Quantum Information Theory

Joseph M. Renes
With elements from the script of Renato Renner & Matthias Christandl

December 13, 2013

Lecture Notes ETH Zurich HS2013

Revision: fa366cb508c2 Branch: default Date: Fri, 13 Dec 2013 17:02:27 +0100

Contents

1 Introduction

"Information is physical" claimed the late physicist Rolf Landauer,[1] by which he meant that

> Computation is inevitably done with real physical degrees of freedom, obeying the laws of physics, and using parts available in our actual physical universe. How does that restrict the process? The interface of physics and computation, viewed from a very fundamental level, has given rise not only to this question but also to a number of other subjects...[1]

The field of quantum information theory is among these "other subjects". It is the result of asking what sorts of information processing tasks can and cannot be performed if the underlying information carriers are governed by the laws of *quantum* mechanics as opposed to *classical* mechanics. Using the spin of a single electron to store information, for instance, rather than the magnetization of a small region of magnetic material.

Famously, it is possible for two separated parties to communicate securely using only *insecure* classical and quantum transmission channels (plus a short key for authentication), using a protocol for quantum key distribution (QKD). Importantly, the security of the protocol rests on the correctness of quantum mechanics, rather than any assumptions on the difficulty of particular computational tasks—such as factoring large integers—as is usual in today's cryptosystems. This is also fortunate from a practical point of view because, just as famously, a quantum computer can find prime factors very efficiently. On the other hand, as opposed to classical information, quantum information cannot even be copied, nor can it be deleted!

The goal of this course is to provide a solid understanding of the foundations of quantum information theory with which we can examine some of the counterintuitive phenomena in more detail. In the next few lectures we will study the foundations more formally and completely, but right now let's just dive in and get a feel for the subject.

1.1 Bits versus qubits

Classical information, as you already know, usually comes in *bits*, random variables which can take on one of two possible values. We could also consider "dits", random variables taking on one of d values, but this can always be thought of as some collection of bits. The point is that the random variable takes a definite value.

In contrast, quantum information comes in *qubits*, which are normalized vectors in $\mathbb{C}^2$. Given some basis $|0\rangle$ and $|1\rangle$, the qubit state, call it ψ, can be written $|\psi\rangle = a|0\rangle + b|1\rangle$, with $a, b \in \mathbb{C}$ such that $|a|^2 + |b|^2 = 1$. The qubit is generally not definitely in either state $|0\rangle$ or $|1\rangle$; if we make a measurement whose two outcomes correspond to the system being in $|0\rangle$ and $|1\rangle$, then the probabilities are

$$\text{prob}(0) = |\langle 0|\psi\rangle|^2 = |a|^2 \qquad \text{prob}(1) = |\langle 1|\psi\rangle|^2 = |b|^2 \tag{1.1}$$

The state of n qubits is a vector in $\mathbb{C}^{2^n}$, a basis for which is given by states of the ofrm $|0,\ldots,0\rangle = |0\rangle \otimes \cdots \otimes |0\rangle$, $|0,\ldots,1\rangle$, $|0,\ldots,1,0\rangle$, etc. Then we write the quantum state of the entire collection as

$$|\psi\rangle = \sum_{s\in\{0,1\}^n} \psi_s |s\rangle, \tag{1.2}$$

[1] Rolf Wilhelm Landauer, 1927-1999, German-American physicist.

where s are binary strings of length n and once again $\psi_s \in \mathbb{C}$ with $\langle\psi|\psi\rangle = 1 = \sum_s |\psi_s|^2$.

Allowed transformations of a set of qubits come in the form of *unitary* operators, which just transform one basis of $\mathbb{C}^{2^n}$ into another. Knowing this, we can already prove the no-cloning theorem!

1.2 No cloning

Suppose we have a cloning machine, which should perform the following transformation

$$|\psi\rangle|0\rangle \longrightarrow |\psi\rangle|\psi\rangle, \tag{1.3}$$

for any qubit state $|\psi\rangle$. According to the laws of quantum mechanics, the transformation should be described by a unitary U. In particular, U should clone the standard basis states:

$$U|00\rangle = |00\rangle \qquad \text{and} \qquad U|10\rangle = |11\rangle. \tag{1.4}$$

But the action on a basis fixes the action on an arbitrary qubit state, due to the linearity of U. Thus, for $|\psi\rangle = a|0\rangle + b|1\rangle$ we find

$$U|\psi\rangle|0\rangle = a\,U|00\rangle + b\,U|10\rangle = a|00\rangle + b|11\rangle. \tag{1.5}$$

But what we wanted was

$$\begin{align}|\psi\rangle|\psi\rangle &= (a|0\rangle + b|1\rangle)(a|0\rangle + b|1\rangle) \tag{1.6}\\ &= a^2|00\rangle + ab|01\rangle + ba|10\rangle + b^2|11\rangle, \tag{1.7}\end{align}$$

which is not the same. Thus, $U|\psi\rangle|0\rangle \neq |\psi\rangle|\psi\rangle$ for arbitrary qubit states. Note that U *does* clone the basis properly, but by the linearity of quantum mechanics, it can therefore *not* clone arbitrary states.

1.3 Measurement and disturbance

As mentioned before, a generic qubit is not definitely in one of the states $|0\rangle$ or $|1\rangle$. But what happens after a measurement? Surely if we repeat the measurement, we should get the same result (provided nothing much has happened in the meantime). Indeed this is the case in quantum mechanics. Starting from $|\psi\rangle = a|0\rangle + b|1\rangle$ and making the $|0\rangle/|1\rangle$ measurement leaves the system in state $|0\rangle$ with probability $|a|^2$ or the state $|1\rangle$ with probability $|b|^2$, so that a subsequent identical measurement yields the same result as the first.

We can measure in other bases as well. For instance, consider the basis $|\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)$. Now the probabilities for the two outcomes are

$$\text{prob}(+) = |\langle +|\psi\rangle|^2 = \tfrac{1}{2}|a+b|^2 \qquad \text{prob}(-) = |\langle -|\psi\rangle|^2 = \tfrac{1}{2}|a-b|^2. \tag{1.8}$$

Thus, if $|\psi\rangle = |0\rangle$, then $p_\pm = \frac{1}{2}$. That is, the measurement outcome is completely random. And after the measurement the state is either $|+\rangle$ or $|-\rangle$. In this way, measurement disturbs the system by changing its state.

This phenomenon makes QKD possible. Very roughly, a potential eavesdropper attempting to listen in on a quantum transmission by measuring the signals will unavoidably disturb the signals, and this disturbance can be detected by the sender and receiver.

1.4 Quantum key distribution

We can get a flavor of how this works by taking a quick look at the original BB84 protocol, formulated by Charles Bennett and Gilles Brassard in 1984. The goal, as in any QKD protocol, is to create a secret key between the two parties, which may be then be used to encrypt sensitive information using classical encryption methods. A secret key is simply a random sequence of bits which are unknown to anyone but the two parties.

Here's how it works. One party (invariably named Alice) transmits quantum states to the other (invariably named Bob), where the states are randomly chosen from the set $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$. Physically these could correspond to various polarization states of a single photon (horizontal, vertical, +45°, −45°), or anything else whose quantum description is given by the states above. When Bob receives each signal, he immediately measures it, randomly choosing either the "standard" $|k\rangle$ basis ($k = 0, 1$) or the "conjugate" $|\pm\rangle$ basis.

If the quantum states arrive at Bob's end unchanged, then when he measures in the same basis Alice used to prepare the state, he will certainly get the corresponding outcome. That is, if Alice prepares a standard basis state and Bob makes a measurement in the standard basis, they will have the same classical bit describing which basis element was transmitted/received. When Alice prepares $|0\rangle$, Bob is certain to see $|0\rangle$, so they can create one bit of secret key (with value 0). On the other hand, if Bob's basis does not match Alice's then Bob's "which-basis-element" bit is totally uncorrelated with Alice's, and hence useless. When Alice sends $|0\rangle$ but Bob measures in the conjugate basis, his outcome is completely random. Alice and Bob can separate the good cases from the bad ones by simply announcing publicly which basis they used in each instance.

Due to the fragility of quantum states, any attempt by a would-be eavesdropper (invariably named Eve) to spy on the quantum signals can be noticed by Alice and Bob. Suppose Eve intercepts the signals, measures them in one basis or the other, and then resends the state corresponding to the outcome she observed. This will cause errors in the bits created by Alice and Bob, which they can observe by sacrificing a portion of the key and directly comparing it publicly.

Specifically, Eve's action causes an error with probability $1/4$. For concreteness, suppose Alice sends $|0\rangle$. Half the time Eve measures in the standard basis and passes $|0\rangle$ to Bob without error. The other half of the time she measures in the conjugate basis, which produces a random outcome. Each of the two possible states $|\pm\rangle$ has a probability of $1/2$ of generating the correct outcome $|0\rangle$ when measured by Bob, so the overall error probability is $1/4$.

Thus, if Alice and Bob compare a portion of the key and observe *no* errors, then they can be relatively certain that the remainder of the key is secure against this "intercept-resend" attack: Eve could not have gained any information about the key.

Although we haven't proven that QKD can be secure against *arbitrary* attacks, this example illustrates the basis mechanism of security. The crucial point is that the fragility of quantum information implies that the information gained by Eve is linked to the errors observed in the key.

I should hasten to add that although in this example Alice and Bob abort the protocol for any nonzero error rate, it is possible to construct QKD protocols which can tolerate a finite amount of error.

1.5 Quantum computation is not like classical computation

From a computer science perspective, we might now wonder why quantum computers could be more powerful than classical computers, given the rough sketch of quantum information theory we have seen so far. After all, quantum states are vectors, operations on them are unitary operators, and measurements correspond to taking an inner product, all of which can be simulated on a classical computer. Right! A quantum computer cannot compute anything that a classical computer cannot, since we can always simulate the former with the latter. But what is really important are the necessary resources, in particular how much space (memory) we are going to need and how much time it is going to take.

A quantum computation, like a classical computation, is the calculation of a given function of the (classical) input. In a quantum computer we feed in $|x\rangle$ for input x. For instance, the factoring algorithm is a means to compute $f(x) = (p_1, p_2, \ldots)$, the prime factors of input x. The goal is to do this quickly, in an amount of time t which scales algebraically with the length of the input, i.e. $t \approx \text{poly}(|x|)$, where $|x|$ is the number of bits of x. Algorithms scaling exponentially in $|x|$, on the other hand, quickly become too slow.

Algorithms are sequences of simple operations which yield the action of the desired function. For instance, we can build up any function we want (assuming it takes binary strings to binary strings) out of AND, OR, and NOT operations on just two bits at a time (or one for NOT). Indeed, NAND or NOR gates alone suffice to compute any function. The runtime of the computation is then how many steps we need to execute all of the required gates.

Quantum algorithms are largely the same, sequences of unitary operations acting on just one and two qubits at a time. A quantum computer is therefore any device with which we can perform suitable unitary gates to the initial state and then read out (measure) the final state to get the answer, as in

$$|x\rangle \xrightarrow{f} U_f|x\rangle = |f(x)\rangle \qquad U_f = V_n V_{n-1} \cdots V_1, \tag{1.9}$$

where the V_j are single- and two-qubit operations. Actually, we only need something like

$$p_{f(x)} = |\langle f(x)|U_f|x\rangle|^2 \geq 2/3, \tag{1.10}$$

so that the probability of getting the right answer is large. By repeating the computation a modest number of times we can achieve whatever probability of error we like.

Where does the power of a quantum computer come from? I don't think anyone has a very precise answer to that question, but we can get an idea by thinking about how we might simulate it classically and where that approach goes wrong. Since the algorithm is just equivalent to multiplication of unitary matrices, the simplest thing is just to do that ourselves. But wait! The matrices are $2^n \times 2^n$ dimensional for n qubits! $2^{36} \approx 70$ Gb, so we can only simulate around 36 qubits with today's hardware. Still thinking in terms of matrices, after each step we have a vector giving the amplitude to be in each of the various computational states. The trouble is, these amplitudes are complex numbers, and therefore the states interfere with each other when going from one step to the next. Thus, we have to keep track of *all* of them (or, it is not clear how to get by without doing this).

To see this more concretely, suppose we want to calculate $|\langle y|U_f|x\rangle|^2$ for some value of y. Since $U_f = V_n V_{n-1} \cdots V_1$, we can express this in terms of the matrix elements of the V_k:

$$|\langle y|U_f|x\rangle|^2 = \Big|\sum_{z_1,\ldots,z_{n-1}} \langle y|V_n|z_{n-1}\rangle \underbrace{\langle z_{n-1}|V_{n-1}|z_{n-2}\rangle}_{\text{matrix element}} \cdots \langle z_1|V_1|x\rangle\Big|^2. \tag{1.11}$$

This is the product of matrices that we wanted to calculate earlier. Instead of doing that, we could try to keep track of the amplitude associated with each computational path, i.e. sequence $|x\rangle$, $|z_1\rangle$, $\ldots$,

$|y\rangle$. This is just the path integral of quantum mechanics, adapted to the present scenario of dynamics by discrete jumps represented by unitaries. To each path is associated an amplitude α_k,

$$\alpha_k = \langle y|V_n|z_{n-1}\rangle \underbrace{\langle z_{n-1}|V_{n-1}|z_{n-2}\rangle}_{\text{matrix element}} \cdots \langle z_1|V_1|x\rangle, \tag{1.12}$$

so that

$$|\langle y|U_f|x\rangle|^2 = \Big| \sum_{\text{paths } k} \alpha_k \Big|^2. \tag{1.13}$$

The idea would then be to estimate the expression by randomly sampling a modest number of paths. But this does not work either, again due to interference—the overall magnitude can be quite small even though each α_k might not be. We need to know a sizable fraction of the α_k to be able to predict the transition probability. Alas, there are an exponential number of paths.

Observe that if the algorithm were such that after each step, most of the probability amplitude were concentrated on one or a few of the states $|z\rangle$, then we could simulate the computation efficiently. In the case of weight on just one state, this essentially *is* a classical computation, since we just jump from $x \to z_1 \to z_2 \to \cdots \to y$.

One often hears the claim that quantum computers get their power because n qubits can encode or represent 2^n numbers. That is true, in the sense that it takes 2^n complex numbers to specify a quantum state. But it also takes 2^n numbers, now just reals, to specify the probability distribution of n bits! If the initial distribution and all computational steps are deterministic, then the computation takes just one path. But the bigger point is that even if it were probabilistic, we could still potentially sample from the set of paths to get an idea of the transition probability. The possibility of interference between paths precludes us from doing this in the quantum case.

1.6 Further reading

It is not the 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 [2], as well as the more recent offerings from Rieffel and Polak [3], Barnett [4], and especially Schumacher and Westmoreland [5]. An inspiration for many of these books and early lecture notes is the book by Peres [6]. Wilde [7] presents in detail the main results pertaining to information processing tasks such as compression and communication; in the classical setting, these are treated by Cover and Thomas [8]. Mackay [9] treats information theory and many other interesting topics such as Bayesian inference and neural networks from a physics point of view. Mermin [10] gives a concise introduction to quantum algorithms. There are too many lecture notes for quantum information available online to list here; of particular note are those by Preskill [url] as well as Watrous [url]. The argument about computational paths is adapted from Aaronson [11] (see also [url]).

2 Probability Theory

A nice way to understand the formalism of quantum mechanics (but not the physics) is as a generalization of classical probability theory. Moreover, classical information theory is formulated in the language of probability theory, so quantum information theory will be as well. Therefore, we begin by recalling some key notions of probability theory. This chapter is not meant as an introduction to probability theory, however. Instead, 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?

The notion of probability is actually a rather delicate philosophical question, and it is not the topic of this course to answer it. For the purpose of this course, it might make sense to take a *Bayesian*[1] point of view, meaning that probability distributions are generally interpreted as a *state of knowledge*. To illustrate this approach, consider a game where a quizmaster hides a prize behind one of three doors and 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 door is equally likely to hide the prize. Hence, the probability distribution P_X^{cand} that the candidate would assign to X is uniform,

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

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

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

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

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

When interpreting a probability distribution as a *state of knowledge* and, hence, as *subjective* quantity, we must 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 a *sender* would like 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 assign different probability distributions to M. In the following, when describing such situtations, we will ascribe all distributions as states of knowledge of an *outside observer*.

[1] Thomas Bayes, c. 1701 – 1761, English mathematician and Presbyterian minister.

[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*.

2.2 Probability spaces and random variables

Both the concepts of probability and random variables are important in both physics and information theory. Roughly speaking, one can think of a random variable as describing the value of some physical degree of freedom of a classical system. Hence, in classical information theory, it is nature to think of data as being represented by random variables.

In this section we define probability spaces and random variables. For completeness, we first give the general mathematical formulation based on probability spaces, known as the Kolmogorov[3] axioms. Later, we will restrict to *discrete* spaces and 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 the information-theoretic considerations of this course.

2.2.1 Probability space

The basic notion in the Kolmogorov approach to probability theory is a *probability space*, which models an experiment with random outcomes or, in our Bayesian interpretation, a physical system with properties that are not fully known. It is a collection of three things:

1. a *sample space* Ω, which represents the set of all possible outcomes,
2. a set of *events* $\mathscr{E}$, which are collections of possible outcomes, and
3. a *probability measure* P, which gives the probability of any event.

The set of events is required to be a *σ-algebra*, which means that (i) $\mathscr{E} \neq \emptyset$, i.e. $\mathscr{E}$ is not trivial, (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 countable family of events then $\bigcup_{i \in \mathbb{N}} E_i$ is an event. In particular, from these requirements one can show that Ω and $\emptyset$ are events, called the *certain event* and the *impossible event*. The requirements of a σ-algebra reflect the probabilistic setting. For any given even there ought to be an "opposite" event such that one or the other is certain to occur, hence the requirement that complements exist. And for any two events one should be able to find an event which corresponds to either one occurring, hence the requirement that unions exist.

The *probability measure* P on $(\Omega, \mathscr{E})$ is a function $P : \mathscr{E} \to \mathbb{R}_+$ that assigns to each event $E \in \mathscr{E}$ a nonnegative real number $P[E]$, called the *probability of E*. It must satisfy the Kolmogorov probability axioms

1. $P[\Omega] = 1$ and
2. $P\left[\bigcup_{i \in \mathbb{N}} E_i\right] = \sum_{i \in \mathbb{N}} P[E_i]$ for any countable family $(E_i)_{i \in \mathbb{N}}$ of pairwise disjoint events.

The axioms are precisely what is needed to be compatible with the σ-algebra structure of events. The second axiom directly echoes the union-property of events, and since E and E^c are disjoint, $P[E] + P[E^c] = P[\Omega] = 1$ so that indeed either E or E^c is certain to occur, since the certain event has probability one. Of course, the impossible event has probability zero, since it is the complement of the certain event.

The above applies for quite general sample spaces, including those which are uncountably infinite such as $\mathbb{R}$. To properly deal with such cases one needs to be able to take limits of sequences of events, hence the constant attention to *countable* collections of events and so forth. The pair $(\Omega, \mathscr{E})$ is known

[3] Andrey Nikolaevich Kolmogorov, 1903 – 1987, Russian mathematician.

in this general context as a *measureable space*, and the uncountably infinite case is important in the mathematical study of (Lebesgue[4]) integration. In this course we will be concerned with discrete sample spaces Ω, for which the set of events $\mathscr{E}$ can be taken to be the *power set* $\mathscr{E} = 2^{\Omega}$, the set of all subsets of Ω.

2.2.2 Random variables

It may be slightly surprising, but in this formulation of probability theory, random variables are best thought of as *functions* from Ω to the space of values taken by the random variable. The precise definition is as follows. Suppose that $(\Omega, \mathscr{E}, P)$ is a probability space and let $(\mathscr{X}, \mathscr{F})$ be another measurable space. A *random variable* X is a function from Ω to $\mathscr{X}$,

$$X: \quad \omega \mapsto X(\omega), \tag{2.1}$$

which is *measurable* with respect to the σ-algebras $\mathscr{E}$ and $\mathscr{F}$. Measurable means that the preimage of any $F \in \mathscr{F}$ is an event in $\mathscr{E}$, i.e. $X^{-1}(F) \in \mathscr{E}$. Therefore the events $\mathscr{F}$ inherit a probability measure P_X from the probability space, like so:

$$P_X[F] := P[X^{-1}(F)] \quad \forall F \in \mathscr{F}. \tag{2.2}$$

The space $(\mathscr{X}, \mathscr{F})$ is often called the *range* of the random variable X.

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

$$(X, Y): \quad \omega \mapsto X(\omega) \times Y(\omega). \tag{2.3}$$

Here, $\mathscr{F} \times \mathscr{G}$ denotes the set $\{F \times G : F \in \mathscr{F}, G \in \mathscr{G}\}$, and it is easy to see that $\mathscr{F} \times \mathscr{G}$ is a σ-algebra over $\mathscr{X} \times \mathscr{Y}$.

We will typically write P_{XY} to denote the *joint probability measure* $P_{(X,Y)}$ on $(\mathscr{X} \times \mathscr{Y}, \mathscr{F} \times \mathscr{G})$ 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 $(X_1, \ldots, X_n)$.

2.2.3 Events from random variables

Events are often themselves 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 $\{X > x_0\}$.

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 $(\Omega, \mathscr{E}, P)$ is ultimately irrelevant. In fact, as long as we are only interested in events defined in terms of the random variables $X_1, \ldots, X_n$, we can without loss of generality identify the sample space $(\Omega, \mathscr{E})$ with the range of the tuple $(X_1, \ldots, X_n)$ and define the probability measure P to be equal to $P_{X_1 \cdots X_n}$.

[4] Henri Léon Lebesgue, 1875 – 1941, French mathematician.

2.2.4 Conditional probability

Any event $E' \in \mathscr{E}$ with $P(E') > 0$ gives rise to a new probability measure $P[\cdot|E']$ on $(\Omega, \mathscr{E})$, the conditional probability, defined by

$$P[E|E'] := \frac{P[E \cap E']}{P[E']} \quad \forall E \in \mathscr{E}. \tag{2.4}$$

The probability $P[E|E']$ of E conditioned on E' can be interpreted as the probability that the event E occurs if we already know that the event E' has occurred. The logic of the definition is that restricting Ω to the elements in the event E' effectively gives a new sample space, whose events are all of the form $E \cap E'$. The probability of any of the new events is its original probability, rescaled by the probability of the new sample space. Analogously to (2.2), the conditional probability measure also gives rise to a conditional probability measure of any random variable X, $P[\cdot|E']$, i.e.,

$$P_{X|E'}[F] := P[X^{-1}(F)|E'] \quad \forall F \in \mathscr{F}. \tag{2.5}$$

Two events E and E' are said to be *mutually independent* when $P[E \cap E'] = P[E] \cdot P[E']$, which implies $P[E|E'] = P[E]$.

In the Bayesian framework, the conditional probability describes the change in our state of knowledge when we acquire additional information about a system that we describe with the probability space $(\Omega, \mathscr{E}, P)$, in particular when we learn that the event E is certain. With a view toward our later formulation of quantum mechanics, we can think of the process of acquiring information as a *measurement* of the system. If, prior to the measurement, our probability were $P[\cdot]$, then after learning that E' is certain our probability becomes $P[\cdot|E']$.

2.3 A vector representation of finite discrete spaces

In the remainder of these lecture notes, we specialize to the case of finite discrete probability spaces $(\Omega, \mathscr{E}, P)$. Now Ω is a discrete set, which we will assume to contain finitely many elements $N = |\Omega|$. Further, we take the σ-algebra of events to be the power set 2^Ω, i.e. $\mathscr{E} := \{E \subseteq \Omega\}$, which one can easily verify to indeed be a valid σ-algebra. Such spaces have a simple representation in terms of real-valued vectors in a finite-dimensional space; this will prove useful later in understanding the similarities and differences between classical probability theory and the formalism of quantum mechanics.

2.3.1 Representing the probability space

Since Ω is finite, we may take the elements $\omega \in \Omega$ to be the integers $1, \dots, N$ and associate to the ωth element the vector $\vec{s}_\omega \in \mathbb{Z}_2^N$ which has a single 1 in the ωth component and all other components zero. Any event is a collection of elements from the sample space, which corresponds to the sum of the associated sample space vectors. The vector $\vec{e}(E) \in \mathbb{Z}_2^N$ associated with the event E is defined by

$$\vec{e}(E) = \sum_{\omega \in E} \vec{s}_\omega, \tag{2.6}$$

i.e. $\vec{e}(E)$ has a 1 in any component corresponding to an ω contained in the event E. Thus, the possible $\vec{e}$ are *all* the vectors in $\mathbb{Z}_2^N$, while the sample space corresponds to the usual basis of $\mathbb{Z}_2^N$. Notice that the inner product between $\vec{e}(E)$ and $\vec{s}_\omega$ indicates whether the ωth element of Ω is contained in E: $\vec{e}(E)\cdot\vec{s}_\omega = 1$ if $\omega\in E$ and 0 otherwise.

Since the probability is additive for families of pairwise disjoint events, and the sample space elements are pairwise disjoint as events, by the second axiom we have

$$P(E) = \sum_{\omega\in E} P[\{\omega\}] = \sum_{\omega\in\Omega} \vec{e}(E)\cdot\vec{s}_\omega P[\{\omega\}] = \vec{e}(E)\cdot\left(\sum_{\omega\in\Omega}\vec{s}_\omega P[\{\omega\}]\right). \tag{2.7}$$

This suggests we define a vector $\vec{p} = \sum_{\omega\in\Omega}\vec{s}_\omega\, P[\{\omega\}]\in\mathbb{R}^N_+$, which is just the list of probabilities of the sample space elements. Taking $E=\Omega$ in the above, we see that the first axiom implies $||\vec{p}||_1 = 1$. The nice feature of this representation is that from $\vec{p}$ the probability of any event can be found via the inner product:

$$P[E] = \vec{e}(E)\cdot\vec{p}. \tag{2.8}$$

2.3.2 Random variables and conditional probabilities

Real-valued random variables can also be represented as vectors in $\mathbb{R}^N_+$, in order to represent the expectation value. Let $X(\omega) = x_\omega$ and define $\vec{x} = \sum_{\omega\in\Omega} x_\omega\,\vec{s}_\omega$. Then the *expected value* of X is just the average value under the probability distribution,

$$\begin{align} \langle X\rangle &:= \sum_{\omega\in\Omega} P[\{\omega\}]X(\omega) \tag{2.9}\\ &= \vec{x}\cdot\vec{p}. \tag{2.10}\end{align}$$

We can also succinctly represent the rule for conditional probability, (2.4), in this framework. For some event E', let us call the vector representation of the conditional probability $\vec{p}\,'$. What is $\vec{p}\,'$ in terms of $\vec{p}$? The denominator of (2.4) is simple enough: $P[E'] = \vec{e}(E')\cdot\vec{p}$. For the numerator, we need only consider the probabilities of the singleton events $\{\omega\}$, since all other events are just unions of these. Then, the event $\{\omega\}\cap E'$ is just $\{\omega\}$ when $\omega\in E'$ and $\emptyset$ otherwise. Therefore we have

$$\begin{align} \vec{p}\,' &= \frac{1}{\vec{e}(E')\cdot\vec{p}}\sum_{\omega\in E'}\left(\vec{s}_\omega\cdot\vec{p}\right)\vec{s}_\omega \tag{2.11}\\ &= \frac{1}{\vec{e}(E')\cdot\vec{p}}\sum_{\omega\in\Omega}\left(\vec{s}_\omega\cdot\vec{p}\right)\left(\vec{e}(E')\cdot\vec{s}_\omega\right)\vec{s}_\omega. \tag{2.12}\end{align}$$

The conditional probability vector is formed by discarding or projecting out the components of $\vec{p}$ which are inconsistent with E', and then normalizing the result.

2.3.3 Measurement

Earlier we motivated the notion of conditional probability in the Bayesian framework as relevant when making a measurement on a physical system. If the measurement reports that the event E' is certain, we accordingly update the probability distribution to the conditional probability distribution. But this does not describe the whole measurement procedure, for we only considered one measurement outcome E' and surely at least the event E'^c was also, in principle, possible.

We can think of a measurement as a *partition* of Ω into a collection of disjoint events $E_1, E_2, \ldots, E_M$, where M is the number of outcomes of the measurement. The most intuitive measurement in this sense is just the collection of all singletons $\{\omega\}$, bur really any partition will do. The measurement then reports the kth outcome with probability $P[E_k] = \vec{e}(E_k)\cdot\vec{p}$ and updates $\vec{p}$ to $\vec{p}_k$ according to (2.11). Notice that if we average the new probability vector over the measurement outcomes themselves, we end up with the original $\vec{p}$:

$$\sum_{k=1}^{M} P(E_K)\vec{p}_k = \sum_{k=1}^{M} \vec{e}(E_k)\cdot\vec{p}\,\frac{1}{\vec{e}(E_k)\cdot\vec{p}} \sum_{\omega\in E_k} (\vec{s}_\omega\cdot\vec{p})\,\vec{s}_\omega \tag{2.13}$$

$$= \sum_{\omega\in\Omega} (\vec{s}_\omega\cdot\vec{p})\,\vec{s}_\omega = \vec{p}. \tag{2.14}$$

Averaging the updated distribution over the measurement outcome is like making a measurement and then forgetting the outcome. So here we see that doing this has no effect: Forgetting the outcome is like undoing the measurement.

2.3.4 Transformations of probabilities

Lastly, since we have represented the probability space in a (linear) vector space, we may ask what sorts of linear transformations preserve the probability structure. That is, what matrices T take probability vectors to probability vectors?

If we are to have $\vec{p}' = T\vec{p}$ be an element of $\mathbb{R}_+^N$ for arbitrary input $\vec{p}$, then T must have positive entries. Moreover, $||\vec{p}'||_1 = 1$ should also hold, meaning $\sum_{\omega,\omega'\in\Omega} T_{\omega',\omega}p_\omega = 1$. Choosing $\vec{p}$ to be the various $\vec{s}_\omega$ themselves, we find that T must satisfy

$$\sum_{\omega'\in\Omega} T_{\omega',\omega} = 1 \quad \forall\,\omega\in\Omega. \tag{2.15}$$

Thus, T must be a positive matrix whose column-sums are all one; such matrices are called *stochastic matrices*. If the row-sums are also all one, the matrix is *doubly stochastic*, and one can show that $T\vec{p} = \vec{p}$ for $\vec{p}$ having all entries equal to $1/N$. Generally, we can think of $T_{\omega',\omega}$ as a transition probability, the probability to end up in ω' when starting from ω; we will return to this later in §2.5.

Among the doubly-stochastic matrices are the *permutation matrices*, which have a single 1 in each row and column. Since the action is just to rearrange the elements of the sample space, they can be undone (using the matrix representing the inverse permutation). That is, permutation matrices describe *reversible transformations*.

2.4 Discrete random variables

In information theory we will need to work with many random variables over a probability space. It will prove more convenient to use the random variables to refer to events, as in §2.2.3, rather than constantly referring to the probability space itself. There is hardly any difference between these two conventions if we consider only one random variable, but when working with many it is easier to just refer to events by the values of the random variables rather than clutter up the notation. Thus, given a collection of discrete random variables X_1, X_2, and so on, each taking on a finite number of values, the sample space is just $\mathcal{X}_1\times\mathcal{X}_2\times\cdots$ and the σ-algebra of events is again the power set.

We will often refer to $\mathcal{X}$ as the *alphabet* of the random variable X. We also make use of the *probability mass function*, $P_X(x)$, which gives the probability of the event $X = x$ and satisfies the normalization condition $\sum_{x\in\mathcal{X}} P_X(x)=1$. For a single random variable, the values of this function are just the entries of $\vec{p}$.

Certain probability distributions or probability mass functions are important enough to be given their own names. We call P_X *flat* if all non-zero probabilities are equal. By the normalization condition, $P_X(x)=\frac{1}{|\text{supp}P_X|}$ for all $x\in\mathcal{X}$, where $\text{supp}P_X:=\{x\in\mathcal{X}:P_X(x)>0\}$ is the *support* of the function P_X. Furthermore, P_X is *uniform* if it is flat and has no zero probabilities, whence $P_X(x)=\frac{1}{|\mathcal{X}|}$ for all $x\in\mathcal{X}$.

2.4.1 Joint, marginal, and conditional distributions

When working with more than one random variable the concepts of joint, marginal, and conditional distributions become important. The following definitions and statements apply to arbitrary n-tuples of random variables, but we 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}, where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabets of X and Y, respectively. The extension to arbitrary n-tuples is straightforward.

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

$$P_Y(y)=\sum_{x\in\mathcal{X}} P_{XY}(x,y) \quad \forall y\in\mathcal{Y}, \tag{2.16}$$

and likewise for P_X. 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

$$P_{X|Y=y}(x)=\frac{P_{XY}(x,y)}{P_Y(y)} \quad \forall x\in\mathcal{X}. \tag{2.17}$$

2.4.2 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 distribution then satisfies $P_{XY}(x,y)=P_X(x)P_Y(y)$.

Related is the notion of *Markov chains*.[5] A sequence of random variables $X_1,X_2,\ldots$ is said to form a *Markov chain*, denoted $X_1\leftrightarrow X_2\leftrightarrow\cdots\leftrightarrow X_n$, if for all $i\in\{1,\ldots,n-1\}$

$$P_{X_{i+1}|X_1=x_1,\ldots,X_i=x_i}=P_{X_{i+1}|X_i=x_i} \quad \forall x_1,\ldots,x_i. \tag{2.18}$$

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 $X_1,\ldots,X_{i-1}$. Note that the arrows in the notation for the Markov property go both ways; the reader is invited to verify that under (2.18) it also holds that X_{i-1} is independent of $X_{i+1},\ldots,X_n$ given a fixed value of X_i.

2.4.3 Functions of random variables 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 $Y=f(X)$ the random variable defined by the concatenation $f\circ X$. Obviously, $f(X)$ has alphabet $\mathcal{Y}$

[5] Andrey Andreyevich Markov, 1856 – 1922, Russian mathematician.

and, in the discrete case we consider here, the corresponding probability mass function P_Y is given by

$$P_Y(y) = \sum_{x \in f^{-1}(\{y\})} P_X(x). \tag{2.19}$$

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*[6]

$$\langle f(X) \rangle \geq f(\langle X \rangle). \tag{2.20}$$

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

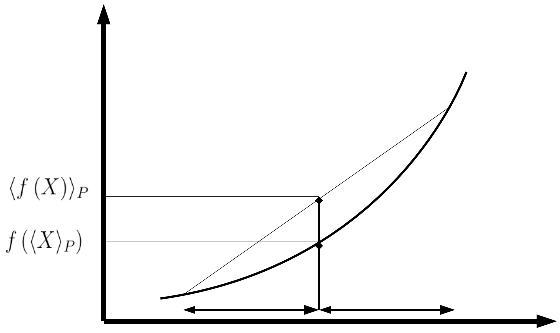


Figure 2.1: Jensen's inequality for a convex function

2.4.4 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

$$P_{X_1 \cdots X_n} = P_X^{\times n} := P_X \times \cdots \times P_X. \tag{2.21}$$

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 §3.4).

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

$$Z_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.22}$$

[6] Johan Ludwig William Valdemar Jensen, 1859 – 1925, Danish mathematician and engineer.

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.,

$$\lim_{n\to\infty} P\big[|Z_n - \mu| < \varepsilon\big] = 1 \quad \forall \varepsilon > 0. \tag{2.23}$$

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

$$P\big[\lim_{n\to\infty} Z_n = \mu\big] = 1. \tag{2.24}$$

2.5 Channels

A *channel* W is a probabilistic mapping that assigns to each value of an *input alphabet* $\mathcal{X}$ a value of the *output alphabet* $\mathcal{Y}$. In doing so, it transforms the random variable X to the random variable $Y = W(X)$. Formally, a channel is any linear transformation of the type encountered in §2.3.4, i.e. a map which preserves the probability structure, only now from the space $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$ to some other space $(\mathcal{Y}, 2^{\mathcal{Y}}, P_Y)$, not back to the original space. It is specified by assigning a number $W(y|x)$ to each input-ouput pair (x,y) such that such that $W(\cdot|x)$ is a probability mass function for any $x \in \mathcal{X}$. The distribution P_Y is then the marginal of the joint distribution

$$P_{XY}(x,y) = P_X(x)W(y|x). \tag{2.25}$$

Since $W(y|x)$ is a properly-normalized distribution for each x, P_{XY} is indeed a valid distribution.

Looking at (2.17), we can see that $W(y|x)$ is just the conditional probability $P_{Y|X=x}(y)$. This justifies the earlier statement in §2.3.4 that $T_{\omega',\omega}$ is the transition probability from ω to ω'; this was not well-defined previously because we lacked the joint input-output space in which to properly formulate the conditional probability.

Moreover, 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 W is given by

$$W(y|x) = \delta_{y,f(x)}. \tag{2.26}$$

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

Not only do channels carry information over space, they also carry information 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.

Example 2.5.1. 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 the reliable transmission of one classical bit of information.

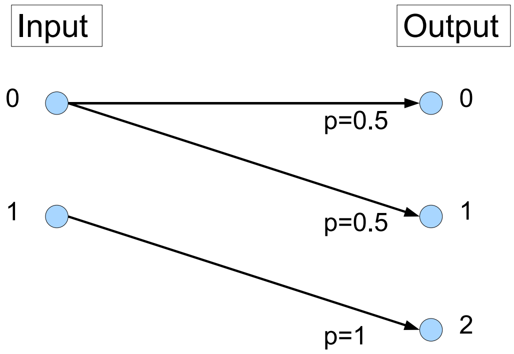


Figure 2.2: Example 1. A reliable channel

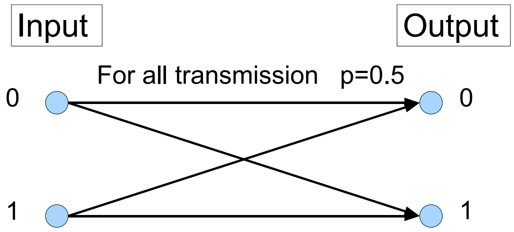


Figure 2.3: Example 2. An unreliable channel

Example 2.5.2. 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 for transmitting information.

2.6 Measurement as a channel

The process of measurement, described in §2.3.3 can also be thought of as a channel, where the input X is the system to be measured and the output Y is the output of the measurement. Consider again a partition of the sample space $\mathcal{X}$ into a set of disjoint events, i.e. a collection of sets E_y, $y = 1, \ldots, |\mathcal{Y}|$ of values that X can take on, with all such sets pairwise disjoint and every possible value $X = x$ an element of some set in the collection. Then define the channel W by

$$W(y|x) = \begin{cases} 1 & x \in E_y \\ 0 & \text{else} \end{cases}, \tag{2.27}$$

which may be succinctly written $W(y|x) = \vec{e}(E_y) \cdot \vec{s}_x$ using the vector representation.

Now consider the joint distribution P_{XY}, given by (2.25). If $\vec{p}$ denotes the vector representative of P_X, the marginal P_Y is simply

$$P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) W(y|x) \tag{2.28}$$

$$= \sum_{x \in \mathcal{X}} (\vec{e}(E_y) \cdot \vec{s}_x)(\vec{s}_x \cdot \vec{p}) \tag{2.29}$$

$$= \vec{e}(E_y) \cdot \vec{p}, \tag{2.30}$$

the probability distribution of the measurement outcomes!Moreover, the conditional distribution of X given Y is also easy to compute:

$$P_{X|Y=y(x)} = \frac{1}{P_Y(y)} P_X(x) W(y|x) \tag{2.31}$$

$$= \frac{1}{\vec{e}(E_y) \cdot \vec{p}} (\vec{s}_x \cdot \vec{p})(\vec{e}(E_y) \cdot \vec{s}_x). \tag{2.32}$$

Comparison with (2.12) reveals that $P_{X|Y=y}(x)$ is nothing more than the xth component of the new probability vector $\vec{p}_y'$.

Thus, the joint distribution induced by the channel (2.27) incorporates both the probabilities of the outcomes of the measurement, as well as the distributions of the original system X conditional on the measurement outcome. The fact that forgetting the outcome undoes the measurement is reflected in the fact that the unconditional distribution of X, i.e. the marginal distribution not conditioned on the value of Y, is simply the original P_X.

The importance of this analysis is that *if channels represent physical operations, then measurement itself is a physical operation*. We obviously expect this to be true, but the previous discussion of measurements and channels did not rule out the possibility that measurement is somehow distinct from the action of a channel.

2.7 Comparing probability distributions

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

$$\delta(P,Q) := \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|. \tag{2.33}$$

In the literature, the trace distance is also called *statistical distance*, *variational distance*, or *Kolmogorov distance*.[7] 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

$$\delta(P,Q) \leq \delta(P,R) + \delta(R,Q). \tag{2.34}$$

Furthermore, $\delta(P,Q) \leq 1$ with equality if and only if P and Q have disjoint support. Because P and Q are normalized, the trace distance can equivalently be written as

$$\delta(P,Q) = 1 - \sum_{x \in \mathcal{X}} \min[P(x), Q(x)]. \tag{2.35}$$

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

Lemma 2.7.1. *For any two probability mass functions P_{XY} and Q_{XY},*

$$\delta(P_{XY}, Q_{XY}) \geq \delta(P_X, Q_X). \tag{2.36}$$

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

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

$$\frac{1}{2}\sum_{x,y}|P_{XY}(x,y)-Q_{XY}(x,y)| \geq \frac{1}{2}\sum_{x}|\sum_{y}P_{XY}(x,y)-Q_{XY}(x,y)| \tag{2.37}$$

$$= \frac{1}{2}\sum_{x}|P_X(x)-Q_X(x)|, \tag{2.38}$$

where the second equality is (2.16). The assertion then follows from the definition of the trace distance. □

Finally, and perhaps most importantly, the trace distance gives a direct bound on how well two distributions can be distinguished. The following lemma shows that the trace distance is itself the biggest difference in probabilities that are possible for some event E under the two distributions:

Lemma 2.7.2. *For P_X and Q_X probability distributions on $\mathcal{X}$,*

$$\delta(P_X, Q_X) = \max_{E\subseteq\mathcal{X}} |P_X[E]-Q_X[E]|, \tag{2.39}$$

where the maximum is taken over all events E.

Proof. The proof is given in the exercises. □

2.8 Further reading

For a nice introduction to the philosophy of probability theory, see [12].

3 Information Theory

3.1 Background: The problem of reliable communication & storage

3.1.1 What is information?

The field of information theory was established by Shannon[1] with his publication "A Mathematical Theory of Communication". It opens by stating

> The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point.[13]

Communication in this sense encompasses the usual meaning of sending a message from one party to another, but also storing a message to be able to read it later. The trouble is, of course, that the means of communication are not inherently reliable or noiseless. Compact discs can be scratched, radio signals can be distorted by the atmosphere on the way from sender to receiver, and so on.

Prior to Shannon's paper, the main approach to improving the quality of communication was to improve the quality of the channel itself. In other words, to engineer channels that more and more closely approximate an ideal noiseless channel. Information theory, however, takes a "software" approach, focusing on changing the way messages are transmitted over noisy channels so that they can nevertheless be faithfully understood by the receiver.

An important step in this direction was the realization that, for the purposes of reliable communication, the "information" being transmitted has nothing to do with the *meaning* of the message. Instead, as Hartley[2] wrote in 1928,

> Hence in estimating the capacity of the physical system to transmit information we should ignore the question of interpretation...and base our result on the possibility of the receiver's distinguishing the result of selecting any one symbol from that of selecting any other.[14]

The task of communication thus divorced from somehow reproducing the meaning of the message, one can then consider manipulating messages in different ways to ensure that its identity can be correctly inferred by the receiver. The method for doing so is to add redundancy to the transmission. The simplest form is just repetition: by repeating a transmitted message three times, the receiver has three attempts to determine the input. If two agree, then the receiver can be more confident in having correctly determined the message than if the message were transmitted only once.

3.1.2 Quantifying information and uncertainty

Shannon's breakthrough was to quantify the minimal amount of redundancy needed to ensure reliable communication. This can be seen as a measure of the information-carrying capacity of the channel. Closely related is a measure of the information contained in the message, or conversely, the uncertainty in the message. Shannon termed this uncertainty *entropy* on the advice of von Neumann,[3] who told him,

[1] Claude Elwood Shannon, 1916 – 2001, American mathematician and electrical engineer.

[2] Ralph Vinton Lyon Hartley, 1888 – 1970, American electrical engineer.

[3] John von Neumann, 1903 – 1957, Hungarian-American mathematician and polymath.

> You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.[15]

Shannon's concept of entropy of a random variable—his "uncertainty function"—is based on the probability distribution of the random variable, and it is a measure of how uncertain we are of the actual value. The flatter the probability distribution, the higher the entropy, and the more peaked, the lower.

Shannon's approach should be contrasted with another approach to quantifying entropy, known as the *algorithmic entropy* or *Kolmogorov complexity*, which quantifies the uncertainty or randomness of a *particular* value (of a random variable) as the length of the shortest computer program needed to recreate the precise value. For example, a bitstring $X = \mathtt{00}\cdots\mathtt{0}$ 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 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.

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[4] render the algorithmic complexity unsuitable as a measure of entropy for most practical applications.

3.2 Entropy

3.2.1 Entropy of events

We will first take an axiomatic approach to motivate the definition of the Shannon entropy and related quantities. First 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 $H(E)$,

$$\begin{aligned} H: \ \mathcal{E} &\rightarrow \mathbb{R}\cup\{\infty\} \\ E &\mapsto H(E)\,. \end{aligned} \tag{3.1}$$

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 $P[E] = \frac{1}{2}$.

[4] 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 are 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 3.2.1. *The function H satisfies the above axioms if and only if it has the form*

$$H: \quad E \longmapsto -\log_2 P[E]. \tag{3.2}$$

Proof. All the axioms are straightforwardly satisfied by H; all that remains to show is uniqueness. To do so, make the ansatz

$$H(E) = f(-\log_2 P[E]), \tag{3.3}$$

where f is an arbitrary function from $\mathbb{R}^+ \cup \{\infty\}$ to $\mathbb{R} \cup \{\infty\}$. Observe that, apart from taking the first axiom into account, this choice can be made without loss of generality, as any possible function of $P[E]$ can be written in this form.

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

$$f(-\log_2 p) + f(-\log_2 p') = f(-\log_2 p p'). \tag{3.4}$$

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

$$f(a) + f(a') = f(a + a'). \tag{3.5}$$

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 $\gamma = 1$. □

3.2.2 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,

$$h(x) := H(E_x) = -\log_2 P_X(x). \tag{3.6}$$

This quantity is sometimes referred to as the *surprisal*. It is clearly positive for any x, since $P_X(x) \leq 1$. The *Shannon entropy* is then defined as the expected surprisal,

$$H(X) := \langle h(X) \rangle = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x), \tag{3.7}$$

with the proviso that $0 \log_2 0$ is taken to be 0. If the probability measure P is unclear from the context, we will include it in the notation as a subscript and write $H(X)_P$.

Similarly, the *min-entropy*, denoted $H_{\min}$, is defined as the *minimum* surprisal,

$$H_{\min}(X) := \min_{x \in \mathcal{X}} h(x) = -\log_2 \max_{x \in \mathcal{X}} P_X(x). \tag{3.8}$$

Another important entropy measure is the *max-entropy*, denoted $H_{\max}$, and defined by Hartley. 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 $\operatorname{supp} P_X := \{x \in \mathcal{X} : P_X(x) > 0\}$ of P_X,

$$H_{\max}(X) := \log_2 \left| \operatorname{supp} P_X \right|. \tag{3.9}$$

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

$$H_{\min}(X) \leq H(X) \leq H_{\max}(X), \tag{3.10}$$

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. 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 $|\operatorname{supp} P_X| = 1$.

The first two clearly hold for all three entropy measures. We shall return to the latter two in §3.2.5.

3.2.3 Conditional entropy

Even more useful than the entropy is the *conditional entropy*, since in information theory we are often interested in the uncertainty about X given the information represented by Y. Thus, we need to generalize the entropy measures introduced above.

Analogously to (3.6) we define the *conditional surprisal*

$$h(x|y) := -\log_2 P_{X|Y=y}(x), \tag{3.11}$$

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,

$$H(X|Y) := \langle h(X|Y) \rangle = -\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} P_{XY}(x,y) \log_2 P_{X|Y=y}(x). \tag{3.12}$$

Note that it could equally-well be defined as the average of the conditional entropies: $H(X|Y) = \langle H(X|Y=y) \rangle_{P_Y}$. For the definition of the *min-entropy of X given Y*, the expectation over Y is replaced by a minimum,

$$H_{\min}(X|Y) := \min_{y \in \mathcal{Y}} H_{\min}(X|Y=y) = \min_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} h(x|y) = -\log_2 \max_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} P_{X|Y=y}(x). \tag{3.13}$$

While in the *max-entropy of X given Y* the expectation is replaced by a maximum,

$$H_{\max}(X|Y) := \max_{y\in\mathcal{Y}} H_{\max}(X|Y=y) = \max_{y\in\mathcal{Y}} \log_2 |\mathrm{supp} P_{X|Y=y}|. \tag{3.14}$$

The conditional entropies H, $H_{\min}$, and $H_{\max}$ satisfy the rules listed in Section 3.2.2. Moreover, it is straightforward to verify that the Shannon entropy H satisfies the *chain rule*

$$H(X|YZ) = H(XY|Z) - H(Y|Z)\,. \tag{3.15}$$

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

$$H(X|Y) = H(XY) - H(Y), \tag{3.16}$$

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.

Furthermore, the conditional entropies can only decrease when conditioning on an additional random variable Z, which is known as *strong subadditivity* or *conditioning reduces entropy*:

$$H(X|Y) \geq H(X|YZ). \tag{3.17}$$

The proof is given in §3.2.5.

3.2.4 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,

$$I(X:Y) := H(X) - H(X|Y). \tag{3.18}$$

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

$$I(X:Y|Z) := H(X|Z) - H(X|YZ). \tag{3.19}$$

Notice that we could equally-well define the conditional mutual information by averaging over Z: $I(X:Y|Z) = \langle I(X:Y|Z=z)\rangle$.

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

$$I(X:Y|Z) = I(Y:X|Z). \tag{3.20}$$

Furthermore, as we shall see in the following section, 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.

3.2.5 Relative entropy and further properties of entropy

The entropies defined in the previous sections will turn out to have direct operational interpretations. That is, they quantify the ultimate limits of information processing tasks such as data compression or noisy channel coding. Here we introduce the relative entropy, mainly for its use in proving the properties we have asserted above, though it also has operational interpretations as well.

Unlike the entropies above, the relative entropy is defined in terms of probability distributions. For two distributions P and Q over the same alphabet $\mathcal{X}$, the *relative entropy* is given by

$$D(P,Q) := \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}. \tag{3.21}$$

The conditional and unconditional Shannon entropies as well as the mutual information can be expressed in terms of the relative entropy, as follows. First let U_X be the uniform distribution on $\mathcal{X}$. Then we have

1. $H(X)_P = \log|\mathcal{X}| - D(P_X, U_X)$,
2. $H(X|Y)_P = \log|\mathcal{X}| - D(P_{XY}, U_X \times P_Y)$,
3. $I(X:Y)_P = D(P_{XY}, P_X \times P_Y)$.

Essentially all of the nontrivial entropic properties stated in previous sections follow from the Gibbs[5] inequality:

Lemma 3.2.2 (Gibbs inequality). *$D(P,Q) \geq 0$, with equality if and only if $P = Q$.*

Proof. Note that $\log_2 x \leq (x-1)/\ln 2$ for all $x > 0$, with equality if and only if $x = 1$. Applied to $Q(x)/P(x)$ we have

$$-\log_2 \frac{P(x)}{Q(x)} = \log_2 \frac{Q(x)}{P(x)} \leq \frac{1}{\ln 2}\left(\frac{Q(x)}{P(x)} - 1\right). \tag{3.22}$$

Averaging over x using $P(x)$ gives the result. □

The Gibbs inequality immediately gives the statement that $I(X:Y) \geq 0$ with equality if and only if $P_{XY} = P_X \times P_y$, i.e. when X and Y are independent. Moreover, $I(X:Y|Z) \geq 0$ follows immediately by the alternative definition $I(X:Y|Z) = \langle I(X:Y|Z=z)\rangle$. In turn, the original definition of conditional mutual information gives strong subadditivity of the conditional entropy.

Since conditioning reduces entropy, the entropy is concave in the probability distribution by the alternative definition $\langle H(X|Y=y)\rangle$. That is, for P, Q, and R each distributions of a random variable X, such that $R(x) = \lambda P(x) + (1-\lambda)Q(x)$ for some $0 \leq \lambda \leq 1$, then $H(X)_R \geq \lambda H(X)_P + (1-\lambda)H(X)_Q$. (Here we see the limits of the notation in which the entropy is a "function" of the random variable and not the distribution, even though it actually depends on the latter, not the former.) To see that the statement holds, take P to be the distribution $P_{X|Y=0}$ and Q to be $P_{X|Y=1}$ for a binary valued random variable Y with distribution $P_Y(0) = \lambda$ and $P_Y(1) = 1-\lambda$.

Finally, we establish latter two properties listed in §3.2.2. The third clearly holds for the max-entropy; it holds for the min-entropy since the largest probability over an alphabet of size $|\mathcal{X}|$ cannot be smaller than $1/|\mathcal{X}|$ (otherwise the distribution would not be normalized). The Shannon

[5] Josiah Willard Gibbs, 1839 – 1903, American physicist.

entropy is also bounded by the alphabet size by $H(X)_P = \log|\mathcal{X}| - D(P_X, U_X)$ and the Gibbs inequality. Moreover, the only distribution for which $H(X) = \log|\mathcal{X}|$ is the uniform distribution U_X. Finally, the fourth property clearly holds for the min- and max-entropies. For the Shannon entropy, consider any distribution as a convex combination of deterministic distributions, namely $P_X(x) = \sum_{x' \in \mathcal{X}} P_X(x')\delta_{x,x'}$ and use the concavity of entropy.

3.2.6 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 $\{1, \ldots, 2^\ell\}$ and probability mass function P_X^ε given by

$$P_X^\varepsilon(x) = \begin{cases} 1-\varepsilon & x=1 \\ \frac{\varepsilon}{2^\ell - 1} & \text{else} \end{cases}, \tag{3.23}$$

where $\varepsilon \in [0,1]$. It is easy to see that $H_{\max}(X)_{P_X^0} = 0$ while $H_{\max}(X)_{P_X^\varepsilon} = \ell$ for any $\varepsilon > 0$. However, observe 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: The max-entropy is not continuous. 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

$$H_{\min}^\varepsilon(X|Y) := \max_{Q_{XY} \in \mathcal{B}^\varepsilon(P_{XY})} H_{\min}(X|Y)_{Q_{XY}}, \tag{3.24}$$

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

$$H_{\max}^\varepsilon(X|Y) := \min_{Q_{XY} \in \mathcal{B}^\varepsilon(P_{XY})} H_{\max}(X|Y)_{Q_{XY}}. \tag{3.25}$$

Note that the original definitions of $H_{\min}$ and $H_{\max}$ are recovered for $\varepsilon = 0$.

3.2.7 Asymptotic equipartition

We have already seen that the Shannon entropy always lies between the min- and the max-entropy (see (3.10)). 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 3.2.3. *For any $n \in \mathbb{N}$, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sequence of i.i.d. pairs of random variables, i.e. $P_{X_1 Y_1 \cdots X_n Y_n} = P_{XY}^{\times n}$ and define $\varepsilon_n = \frac{\sigma^2}{n\delta_n^2}$ for some $\delta_n = o(\frac{1}{\sqrt{n}})$ and σ^2 the variance of the conditional surprisal $h(x|y)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} H_{\min}^{\varepsilon_n}(X_1 \cdots X_n | Y_1 \cdots Y_n) = H(X|Y)_{P_{XY}}, \tag{3.26}$$

$$\lim_{n\to\infty} \frac{1}{n} H_{\max}^{\varepsilon_n}(X_1 \cdots X_n | Y_1 \cdots Y_n) = H(X|Y)_{P_{XY}}. \tag{3.27}$$

Proof. The lemma is a consequence of the law of large numbers, §2.4.4, applied to the random variables $Z_i := h(X_i|Y_i)$, for $h(x|y)$ defined by (3.11). More details are given in the exercises. □

This phenomenon is termed *asymptotic equipartition*; the convergence of the smooth entropies implies that, in some sense, the i.i.d probability distribution is flat, except on a small portion of the alphabet of size $O(\varepsilon)$.

Note that, since the probability of independent events is additive, the Shannon entropy obeys $H(X^{\times n}|Y^{\times n}) = nH(X|Y)$.

3.3 Data compression

3.3.1 Definition of the problem

Now let us put our entropy measures to use. Consider the task of *data compression*, mapping a random variable X over alphabet $\mathscr{X}$ to a new random variable Y over a smaller alphabet $\mathscr{Y}$, such that the mapping can be undone. The task of data compression is also known as *source coding*, as we may imagine X is produced by a *source* of information. Formally, we imagine a compression map $\mathsf{comp} : X \rightarrow Y$ and a decompression map $\mathsf{dec} : Y \rightarrow \hat{X}$, such that both maps are channels. The overall mapping of random variables is

$$X \underset{\mathsf{comp}}{\longrightarrow} Y \underset{\mathsf{dec}}{\longrightarrow} \hat{X}. \tag{3.28}$$

We will further assume that $\mathscr{Y} = \{0,1\}^\ell$ for some $\ell \in \mathbb{N}$ so that we may count the number of bits which are needed to store X reliably.

Here we must make an important distinction in how well the output $\hat{X}$ matches the input X. how the mapping is undone. Schemes in which decompression *exactly* reproduces the input are called *lossless*. Examples include *Huffman coding*[6] or *arithmetic coding*, which are both *variable-length* schemes because while some inputs are made shorter, others become longer. Indeed, this must be the case for lossless schemes, since a mapping which shortens all inputs cannot be perfectly reversible. However, we will be interested in *approximately lossless* schemes in which the probability that the output does not equal the input is allowed to be a (presumably small) parameter $\varepsilon > 0$.

We will judge the compression scheme to be successful if the average error probability is less than ε; the average error probability is defined by

$$p_{\text{err}} := 1 - \sum_{x \in \mathscr{X}} P_{X\hat{X}}(x,x). \tag{3.29}$$

Then we are interested in the smallest ℓ such that the average error probability is less than ε:

$$\ell^\varepsilon(X) := \min\{\ell \in \mathbb{N} : \exists\, \mathsf{comp}, \mathsf{dec} : \, p_{\text{err}} \leq \varepsilon\}. \tag{3.30}$$

3.3.2 Direct source coding theorem

Shannon's genius in determining the ultimate achievable limits to data compression (and channel coding) was to study the properties of *randomly-chosen* compression maps. Often this is referred to as "random coding" but that is slightly misleading, as there is nothing necessarily random about the compression or decompression functions themselves. The basic idea is to consider averaging the

[6] David Albert Huffman, 1925 – 1999, American computer scientist.

average error over the choice of compression map. If this doubly-averaged quantity is small, then there must exist at least one compression map for which the average error probability (3.29) is at least as small.

Moreover, one can show that in fact most compression maps have bounded error probability, by appealing to the *Markov inequality*. This was proven in the exercises and is stated here for later convenience:

Lemma 3.3.1 (Markov inequality). *Let X be a random variable taking* positive *values. Then*

$$P[X \geq \varepsilon] \leq \frac{\langle X \rangle}{\varepsilon}. \tag{3.31}$$

We shall also need the *union bound*, which states the intuitive fact that the probability of the union of some events cannot be larger than the sum of the probabilities of the events:

Lemma 3.3.2 (Union bound). *For any countable set of events $E_1, E_2, \ldots$, we have*

$$P[\cup_i E_i] \leq \sum_i P[E_i]. \tag{3.32}$$

It can be easily proven by induction, but we omit the formal proof here.

Using Shannon's random coding technique, we can prove the *direct source coding theorem*,

Theorem 3.3.3. *For any random variable X and $\varepsilon \geq 0$,*

$$\ell^{\varepsilon}(X) \leq H_{\max}(X) + \log\frac{1}{\varepsilon} + 1. \tag{3.33}$$

Proof. Following Shannon, consider a randomly-chosen, deterministic compression map which maps $\mathcal{X}$ to $\{0,1\}^{\ell}$, for some ℓ to be specified later. The decompressor maps $y \in \mathcal{Y}$ to the value $x \in \mathcal{X}$ which has the largest *posterior* probability, i.e. $\text{argmax}_{x:\mathsf{comp}(x)=y} P_{X|Y=y}(x)$. An error potentially occurs if there is more than one compatible x for a given y. Even in this case there is a chance that the decompressor will make a lucky guess, but we can bound the average error probability by assuming that all such events lead to an error:

$$p_{\text{err}} \leq \sum_{x\in\mathcal{X}} P_X(x) P[\exists x' \neq x : \mathsf{comp}(x') = \mathsf{comp}(x), x' \in \text{supp} P_X]. \tag{3.34}$$

Notice that we need only consider x' which have nonzero probability, as these are the only candidates for inclusion in the decompressor's output (based on $P_{X|Y=y}$), no matter the value of Y. By the union bound, Lemma 3.3.2, the latter factor is bounded as follows

$$P[\exists x' \neq x : \mathsf{comp}(x') = \mathsf{comp}(x), x' \in \text{supp} P_X] \leq \sum_{x'\neq x, x'\in \text{supp} P_X} P[\mathsf{comp}(x') = \mathsf{comp}(x)]. \tag{3.35}$$

Now we average over the random choice of compression map, and denote the resulting averaged average error probability by $\langle p_{\text{err}} \rangle_{\mathsf{comp}}$. Since the probability of two inputs being mapped to the same output (the *collision probability*) under a randomly-chosen function is just 1 divided by the size of the output alphabet, we find

$$\langle p_{\text{err}} \rangle_{\mathsf{comp}} \leq \sum_{x\in\mathcal{X}} P_X(x) \sum_{x'\neq x, x'\in \text{supp} P_X} \frac{1}{2^{\ell}} \tag{3.36}$$

$$\leq 2^{H_{\max}(X)_P - l}. \tag{3.37}$$

Therefore, choosing $\ell = \lceil H_{\max}(X)_P + \log \frac{1}{\varepsilon} \rceil$ ensures that $\langle p_{\text{err}} \rangle_{\text{comp}} \leq \varepsilon$. Since the average over all maps leads to a small probability of error, there must be at least one mapping for which $p_{\text{err}} \leq \varepsilon$. Finally, using $\lceil x \rceil \leq x + 1$ completes the proof. □

Next, we make use of the smoothed max-entropy to derive a tighter bound. For any distribution P_X, suppose we apply an ε-error source coding scheme designed for the $P'_X \in \mathscr{B}^{\varepsilon'}(P_X)$ such that $H_{\max}(X)_{P'} = H^{\varepsilon'}_{\max}(X)_P$. By the properties of the trace distance, the average error probability cannot be larger than $\varepsilon + \varepsilon'$. Thus, we have the following corollary.

Corollary 3.3.4. $\ell^{\varepsilon+\varepsilon'}(X) \leq H^{\varepsilon'}_{\max}(X)_P + \log \frac{1}{\varepsilon} + 1$.

3.3.3 Comments on the source coding protocol

We have shown that there must exist one compression map with the desired properties. But actually, by the Markov inequality, Lemma 3.3.1, no more than a fraction $\sqrt{\varepsilon}$ of compression maps have an error larger than $\sqrt{\varepsilon}$, so even a random choice will most likely work well.

Moreover, while the proof makes use of randomly-chosen compression mappings, a closer look at the proof itself reveals that we only needed to randomly choose among a set of mappings such that the collision probability takes the value needed for (3.36). We need not pick randomly from the set of all mappings for this, indeed randomly choosing a *linear* map is actually enough. The technique of picking randomly from a set of functions so as to have a desired collision probability is called *universal hashing*.

The decompression technique we have used is called MAP decompression, which stands for *maximum a posteriori*, as we use the maximum posterior probability (as opposed to the maximum *prior* probability, $\text{argmax}_{x\in\mathscr{X}} P_X(x)$). This is similar, but not identical to *maximum likelihood* decompression, where the decompressor chooses $\text{argmax}_{x:\text{comp}(x)=y} P_{Y=y|X=x}$.

Note that in the source coding scheme we have constructed, the compressor does not need to know P_X, just the entropy $H^{\varepsilon}_{\max}(X)$. However, the decompressor does need to know P_X, since it uses $P_{X|Y=y}$. Protocols in which the distribution need not be explicitly known are called *universal* or *blind*. Thus, our scheme may be termed *universal at the compressor*. This is important for the task of source coding with side information at the decoder; that is, compression of a source X when the decoder has access to some additional random variable Z which may be correlated with X.

3.3.4 Direct source coding theorem for i.i.d. sources

Now suppose that we wish to compress n instances of a random variable X, that is, the random variable $X^{\times n}$, with distribution $P_{X^{\times n}} = P_{X_1} \times \cdots \times P_{X_n}$. In this case we are interested in the smallest possible *rate* at which $X^{\times n}$ can be compressed, as n tends to infinity, provided the error ε also tends to zero. Let us call this optimal rate the *compressibility of* X. It is formally defined as

$$C(X) = \inf\left\{ \lim_{n\to\infty} \frac{1}{n} \ell^{\varepsilon_n}(X^{\times n}) : \lim_{n\to\infty} \varepsilon_n = 0 \right\}. \tag{3.38}$$

Together with the smooth entropy corollary to the direct source coding theorem, Corollary 3.3.4, a simple application of the asymptotic equipartition property, Lemma 3.2.3, immediately implies the *direct i.i.d. source coding theorem*,

Corollary 3.3.5. *For any random variable* X, $C(X) \leq H(X)$.

Proof. Let ε and ε' in Corollary 3.3.4 both be equal to the error in the AEP, i.e.

$$\varepsilon = \varepsilon' = \frac{\sigma^2}{n\delta_n^2} \quad \text{with} \quad \delta_n = o(\tfrac{1}{\sqrt{n}}) \tag{3.39}$$

Clearly, $\lim_{n\to\infty} \varepsilon = 0$. Then we have

$$\lim_{n\to\infty} \frac{1}{n}\ell^{2\varepsilon}(X^{\times n}) \leq \lim_{n\to\infty} \left[\frac{1}{n}H_{\max}^{\varepsilon}(X^{\times n}) - \frac{1}{n}\log\frac{1}{\varepsilon} + \frac{1}{n}\right] \tag{3.40}$$

$$= H(X) - \lim_{n\to\infty} \frac{1}{n}\log\frac{n\delta_n^2}{\sigma^2} = H(X). \tag{3.41}$$

□

The protocol constructed by this method is an instance of *block compression*, in that both the compressor and decompressor act on blocks of n bits at a time. More practical, however, are means of *stream compression* where the sequence of random variables is compressed one at a time. The aforementioned schemes of Huffman coding and arithmetic coding are examples of stream compression which can operate at (essentially) the rate given above, but we will not go into them here.

3.3.5 Converse source coding theorem for i.i.d. sources

Now let us demonstrate that the rate $H(X)$ is in fact optimal for the task of i.i.d. compression. For this we make use of the *Fano inequality*[7] which links the conditional entropy with the probability of error.

Lemma 3.3.6 (Fano inequality). *For any two random variables X and Y, let $\hat{X}$ be a guess of X generated from Y (by means of a channel). Then*

$$P[\hat{X} \neq X] \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}. \tag{3.42}$$

Proof. The proof will be given in the exercises. □

Now we can state the *converse i.i.d. source coding theorem*,

Theorem 3.3.7. *For any random variable X, $C(X) \geq H(X)$.*

Proof. Suppose we have source coding schemes for each blocklength n with errors ε_n such that $\varepsilon_n \to 0$ as $n \to \infty$. By Fano's inequality, the chain rule for the Shannon entropy, and the fact that the compressor is a deterministic mapping, we have

$$\varepsilon_n n \log|\mathcal{X}| + 1 \geq H(X^{\times n}|Y) \tag{3.43}$$

$$= H(X^{\times n}Y) - H(Y) \tag{3.44}$$

$$= H(X^{\times n}) - H(Y) \tag{3.45}$$

$$= nH(X) - H(Y). \tag{3.46}$$

[7] Robert Mario Fano, born 1917, Italian-American computer scientist.

Since $H(Y) \leq \log|\mathscr{Y}|$, we therefore obtain

$$\frac{1}{n}\log|\mathscr{Y}| \geq H(X) - \varepsilon_n \log|\mathscr{X}| - \frac{1}{n}. \tag{3.47}$$

Taking the limit $n \to \infty$ on both sides completes the proof. □

This result is sometimes called the *weak converse*, since we have shown that compression at rates below the compressibility must result in error rates ε_n which do not converge to zero. In fact, it is possible to prove a *strong converse*, which asserts that compression at rates below the compressibility imply $\varepsilon_n \to 1$. Thus, the compressibility is a sharp transition and no tradeoff between error rate and compression rate is possible in the asymptotic limit.

3.4 Noisy channel coding

3.4.1 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 W that takes inputs X from Alice and outputs Y on Bob's side (see Section 2.5). 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 unlike the case of data compression, it is generally impossible to reach $\varepsilon = 0$ exactly.

To describe the encoding and decoding process, we assume without loss of generality[8] that the message M is represented as an ℓ-bit string, i.e., M takes values from the set $\mathscr{M} = \{0,1\}^\ell$. Alice then applies an *encoding function* $\mathsf{enc}_\ell : \{0,1\}^\ell \to \mathscr{X}$ that maps M to a channel input X. On the other end of the line, Bob applies a *decoding function* $\mathsf{dec}_\ell : \mathscr{Y} \to \{0,1\}^\ell$ to the channel output Y in order to retrieve $\hat{M}$. Then the entire transmission process looks like

$$M \xrightarrow[\mathsf{enc}_\ell]{} X \xrightarrow[W]{} Y \xrightarrow[\mathsf{dec}_\ell]{} \hat{M}. \tag{3.48}$$

The transmission is successful if $M = \hat{M}$. 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

$$p_{\text{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m) := P[\mathsf{dec}_\ell \circ W \circ \mathsf{enc}_\ell(M) \neq M | M = m] \tag{3.49}$$

as the probability that the decoded message $\hat{M} := \mathsf{dec}_\ell \circ W \circ \mathsf{enc}_\ell(M)$ generated by the process (3.48) does not coincide with M.

In the following, we analyze the maximum number of message bits ℓ that can be transmitted in one use of the channel W if we tolerate a *maximum* error probability ε,

$$\ell^\varepsilon(W) := \max\{\ell \in \mathbb{N} : \exists\, \mathsf{enc}_\ell, \mathsf{dec}_\ell : \max_{m \in \mathscr{M}} p_{\text{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m) \leq \varepsilon\}\,. \tag{3.50}$$

[8] 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.

3.4.2 Direct channel coding theorem

The *direct channel coding theorem* provides a lower bound on the quantity $\ell^\varepsilon(W)$. 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 $\log_2 \frac{1}{\varepsilon}$).

Theorem 3.4.1. *For any channel W and any $\varepsilon \geq 0$,*

$$\ell^\varepsilon(W) \geq \max_{P_X}\big(H_{\min}(X) - H_{\max}(X|Y)\big) - \log_2 \frac{1}{\varepsilon} - 3, \tag{3.51}$$

where the entropies on the right hand side are evaluated for the random variables X and Y jointly distributed according to $P_{XY}(x,y) = P_X(x)W(y|x)$, as in (2.25).

The proof idea is illustrated in Fig. 3.1.

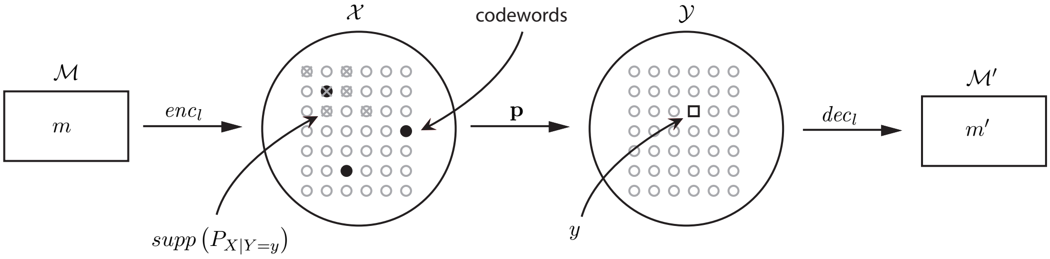


Figure 3.1: The figure illustrates the proof idea of the channel coding theorem. The range of the encoding function enc_ℓ is called the *code* and their elements are the *codewords*.

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

$$\ell = \lfloor H_{\min}(X) - H_{\max}(X|Y) - \log_2 \frac{2}{\varepsilon} \rfloor. \tag{3.52}$$

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

$$\big\langle p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(M)\big\rangle = P[\mathsf{dec}_\ell \circ W \circ \mathsf{enc}_\ell(M) \neq M] \leq \frac{\varepsilon}{2}. \tag{3.53}$$

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

$$p_{\mathrm{err}}^{\mathsf{enc}'_{\ell-1},\mathsf{dec}'_{\ell-1}}(m) \leq \varepsilon \quad \forall m \in \{0,1\}^{\ell-1}. \tag{3.54}$$

From (3.53) and (3.54) we then have

$$
\begin{aligned}
\ell^{\varepsilon}(W) &\geq \ell - 1 && (3.55)\\
&= \lfloor H_{\min}(X) - H_{\max}(X|Y) - \log_2(2/\varepsilon) \rfloor - 1 && (3.56)\\
&\geq H_{\min}(X) - H_{\max}(X|Y) - \log_2(1/\varepsilon) - 3. && (3.57)
\end{aligned}
$$

To prove (3.53), let enc_ℓ and M be chosen at random as described, let $Y := W \circ \mathsf{enc}_\ell(M)$ be the channel output, and let $M' := \mathsf{dec}_\ell(Y)$ 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 $\mathsf{enc}_\ell(m') \in \operatorname{supp} P_{X|Y=y}$. Hence, the probability that the decoding fails is bounded by

$$P[M \neq M' | M = m, Y = y] \leq P[\exists m' \neq m : \mathsf{enc}_\ell(m') \in \operatorname{supp} P_{X|Y=y}]. \tag{3.58}$$

Furthermore, by the union bound in Lemma 3.3.2, we have

$$P[\exists m' \neq m : \mathsf{enc}_\ell(m') \in \operatorname{supp} P_{X|Y=y}] \leq \sum_{m' \neq m} P[\mathsf{enc}_\ell(m') \in \operatorname{supp} P_{X|Y=y}]. \tag{3.59}$$

Because, by construction, $\mathsf{enc}_\ell(m')$ 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

$$
\begin{aligned}
P[\mathsf{enc}_\ell(m') \in \operatorname{supp} P_{X|Y=y}] &= \sum_{x \in \operatorname{supp} P_{X|Y=y}} P_X(x) && (3.60)\\
&\leq |\operatorname{supp} P_{X|Y=y}| \max_x P_X(x) && (3.61)\\
&\leq 2^{-(H_{\min}(X) - H_{\max}(X|Y))}, && (3.62)
\end{aligned}
$$

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

$$P[M \neq M' | M = m, Y = y] \leq 2^{\ell - (H_{\min}(X) - H_{\max}(X|Y))} \leq \frac{\varepsilon}{2}. \tag{3.63}$$

Because this holds for any m and y, we have

$$P[M \neq M'] \leq \max_{m,y} P[M \neq M' | M = m, Y = y] \leq \frac{\varepsilon}{2}. \tag{3.64}$$

This immediately implies that (3.53) 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 (3.53) holds.

It remains to show inequality (3.54). For this, we divide the set of messages $\{0,1\}^\ell$ into two equally large sets $\underline{\mathcal{M}}$ and $\overline{\mathcal{M}}$ at the median, i.e. such that $p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(\underline{m}) \leq p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(\overline{m})$ for any $\underline{m} \in \underline{\mathcal{M}}$ and $\overline{m} \in \overline{\mathcal{M}}$. We then have

$$\max_{m \in \underline{\mathcal{M}}} p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m) \leq \min_{m \in \overline{\mathcal{M}}} p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m) \leq 2^{-(\ell-1)} \sum_{m \in \overline{\mathcal{M}}} p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m). \tag{3.65}$$

Using (3.53), we conclude

$$\max_{m \in \underline{\mathcal{M}}} p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m) \leq 2 \sum_{m \in \{0,1\}^\ell} 2^{-\ell} p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(m) = 2 \langle p_{\mathrm{err}}^{\mathsf{enc}_\ell,\mathsf{dec}_\ell}(M) \rangle \leq \varepsilon. \tag{3.66}$$

Inequality (3.54) then follows by defining $\mathsf{enc}'_{\ell-1}$ as the encoding function enc_ℓ restricted to $\underline{\mathcal{M}}$, and adapting the decoding function accordingly. □

As with data compression, we can tighten the bound by constructing the decoder not using the actual joint distribution P_{XY} but a nearby distribution, from which one can derive bounds on the smooth entropies. This leads to the following corollary, whose proof is included only for completeness and is not examinable in this course.

Corollary 3.4.2. *For any channel W and any $\varepsilon, \varepsilon' \geq 0$,*

$$\ell^{\varepsilon+3\varepsilon'}(W) \geq \max_{P_X}\left(H^{\varepsilon'}_{\min}(X) - H^{\varepsilon'}_{\max}(X|Y)\right) - \log_2 \frac{1}{\varepsilon} - 3, \tag{3.67}$$

where the entropies on the right hand side are evaluated for P_{XY}.

Proof. Consider an arbitrary P_X and the P_{XY} derived from it by the action of the channel. For a smoothing parameter $\varepsilon' > 0$, let P'_{XY} be the distribution for which $H_{\max}(X|Y)_{P'} = H^{\varepsilon'}_{\max}(X)_P$. The smoothed distribution differs from the original in that some events (x,y), which under P_{XY} have small probability, now have zero probability under P'_{XY}, so as to reduce the support of the conditional distribution.

Next, by the properties of the trace distance, $\delta(P_X, P'_X) \leq \varepsilon'$. Now consider smoothing P'_X by $2\varepsilon'$ to P''_X so that $H_{\min}(X)_{P''} = H^{2\varepsilon'}_{\min}(X)_{P'}$, for reasons which will become clear shortly. This will reduce the probability of events with large probability, redistributing the excess to events with small probability. We can extend the smoothing to P'_{XY}, simply by distributing the probability reduction proportionally over all y for a given x. This will maintain the trace distance, so that $\delta(P'_{XY}, P''_{XY}) \leq 2\varepsilon'$. And even more importantly, observe that the min-entropy smoothing step does not adversely affect the previous max-entropy smoothing step. That is, $H_{\max}(X)_{P''} \leq H_{\max}(X)_{P'}$.

From the definition of smooth min-entropy, $H_{\min}(X)_{P''} = H^{2\varepsilon'}_{\min}(X)_{P'} \geq H^{3\varepsilon'}_{\min}(X)_P$, while the above shows $H_{\max}(X)_{P''} \leq H_{\max}(X)_{P'} = H^{\varepsilon'}_{\max}(X)_P$. Therefore, constructing the decoding function using $P''_{X|Y=y}$ will only increase the error by at most $3\varepsilon'$. □

3.4.3 Comments on the channel coding protocol

Many of the same comments regarding the source coding protocol also apply to our channel coding protocol. It is also universal at the encoder, also called universal with an informed decoder. It also need not make use of the full set of random encoding maps. From the proof, it is sufficient that the codewords be merely *pairwise independent* rather than completely independent, and as in the source coding case, this allows us to narrow the set of encoders to linear mappings.

3.4.4 Direct channel coding theorem for i.i.d. channels

Realistic communication channels (e.g., an optical fiber) can usually be used repeatedly. Moreover, such channels are often not inaccurately 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 $W^{\times n} = W \times \cdots \times W$. 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 3.4.1 or Corollary 3.4.2 applied to $W^{\times n}$.

In applications, the number n of channel uses is typically large. It is thus convenient to consider the *rate* of communication, i.e. the number of bits per channel use which can be reliably transmitted, again in the limit $n \to \infty$ such that $\varepsilon \to 0$. The optimal rate is called the *capacity* $C(W)$ of the channel, and is formally defined by

$$C(W) = \sup\left\{ \lim_{n\to\infty} \frac{1}{n}\ell^{\varepsilon_n}(W^{\times n}) : \lim_{n\to\infty} \varepsilon_n = 0 \right\}. \tag{3.68}$$

As with data compression, combination of the asymptotic equipartition property of Lemma 3.2.3 with the smooth entropy corollary to the direct noisy channel coding theorem, Corollary 3.4.2 directly implies the *direct i.i.d. noisy channel coding theorem*,

Theorem 3.4.3. *For any channel W*

$$C(W) \geq \max_{P_X}\big(H(X) - H(X|Y)\big) = \max_{P_X} I(X:Y). \tag{3.69}$$

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

3.4.5 Converse channel coding theorem for i.i.d. channels

We conclude our treatment of channel coding with a proof that the bound given in Theorem 3.4.3 is tight. The proof is similar to the converse for data compression, Theorem 3.3.7, with two additional ingredients, the *data processing processing inequality* and *single-letterization*.

Lemma 3.4.4 (Data processing)**.** *Suppose* $X \leftrightarrow Y \leftrightarrow Z$ *is a Markov chain. Then*

$$I(X:Z) \leq I(X:Y). \tag{3.70}$$

Proof. The proof is given in the exercises. □

Lemma 3.4.5 (Single-letterization)**.** *For any random variable* X^n *on* $\mathcal{X}^{\times n}$*, let* W *be a channel and define* $Y^n = W^{\times n}(X^n)$*. Then*

$$I(X^n : Y^n) \leq nC(W). \tag{3.71}$$

Proof. Using the chain rule we obtain

$$I(X^n : Y^n) = H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\dots,Y_{i-1}X^n). \tag{3.72}$$

Since each channel use is independent of all others, Y_i only depends on X_i, which is to say that it is independent of all other variables when conditioned on the value of X_i. Thus, $H(Y_i|Y_1,\dots,Y_{i-1}X^n) = H(Y_i|X_i)$ and we have $I(X^n : Y^n) = H(Y^n) - \sum_{i=1}^n H(Y_i|X_i)$. Then, by subadditivity of entropy ($H(Y_1Y_2) \leq H(Y_1) + H(Y_2)$), we find

$$I(X^n : Y^n) \leq \sum_i H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) = \sum_{i=1}^n I(X:Y) \leq nC(W). \tag{3.73}$$

□

Now we can state the *converse i.i.d. channel coding theorem*,

Theorem 3.4.6. $C(W) \leq \max_{P_X} I(X:Y)$.

Proof. Suppose we have coding schemes for each blocklength n and maximum error probability ε_n, with $\varepsilon_n \to \infty$. Further, consider the message $M \in \mathcal{M}$ to be a random variable with uniform probability distribution. Clearly, the average error probability will also be less than ε_n. Then by the chain rule and definition of mutual information we have

$$\log|\mathcal{M}| = H(M) = H(M|M') + I(M:M'). \tag{3.74}$$

We can apply the Fano inequality, Lemma 3.3.6, to the first term and data processing to the second; as in (3.48), $M \leftrightarrow X^n \leftrightarrow Y^n \leftrightarrow M'$ form a Markov chain, where X^n is the random variable induced from M by the encoder and $Y^n = W^{\times n}(X^n)$. We then have

$$\log|\mathcal{M}| \leq 1 + P(M \neq M')\log|\mathcal{M}| + I(X^n : Y^n). \tag{3.75}$$

In view of the single-letterization lemma, we can then conclude

$$\log|\mathcal{M}| \leq 1 + \varepsilon_n \log|\mathcal{M}| + nC(W). \tag{3.76}$$

Dividing both sides by n and taking the limit $n \to \infty$ on both sides completes the proof. □

As was the case with source coding, the above represents a *weak converse*, showing that trying to communicate at rates exceeding the capacity will incur some error. Again one can show a *strong converse* stating that the error would in that case actually converge to 1.

3.5 Further reading

Mackay [9] has an excellent discussion of arithmetic coding. For more on universal hashing and pairwise independence, see [16].

4

The Formalism of Quantum Mechanics

4.1 The 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 describe one of the standard approaches to quantum mechanics. It is based on a number of postulates formulated by Dirac and von Neumann regarding the states of physical systems as well as their evolution. (For more details, we refer to Section 2 of [2], 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* $|\phi\rangle \in \mathcal{H}$ which is unique up to a phase factor. In the following, we will call $\mathcal{H}$ the *state space* of the system.

2. *Dynamics:*
   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 $|\phi\rangle \in \mathcal{H}$ at time t_0 to the state
   $$|\phi'\rangle = U|\phi\rangle \tag{4.1}$$
   at time t_1. The unitary U is unique up to a phase factor.

3. *Observables:*
   Any physical property of a system that can be measured is an observable and all observables are represented by self-adjoint linear operators acting on the state space $\mathcal{H}$. Each eigenvalue x of an observable O corresponds to a possible value of the observable. Since O is self-adjoint, it takes the form $O = \sum_x x P_x$, where P_x is the projector onto the subspace with eigenvalue x.

4. *Measurements:*
   The measurement of an observable O yields an eigenvalue x. If the system is in state $|\phi\rangle \in \mathcal{H}$, then the probability of observing outcome x is given by
   $$P_X(x) = \mathrm{tr}(P_x|\phi\rangle\langle\phi|). \tag{4.2}$$
   The state $|\phi'_x\rangle$ of the system after the measurement, conditioned on the event that the outcome is x, is just
   $$|\phi'_x\rangle := \sqrt{\frac{1}{P_X(x)}} P_x|\phi\rangle. \tag{4.3}$$

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

$$|\Psi\rangle = |\phi\rangle \otimes |\phi'\rangle \in \mathscr{H}_A \otimes \mathscr{H}_B. \tag{4.4}$$

4.1.1 Comparison with classical probability theory

We can make an analogy between classical probability theory and the formalism of quantum mechanics, as follows

Quantum				Classical
state vector	$\vert\phi\rangle$	$\approx$	$\vec{p}$	probability distrib.*
observable	O	$\approx$	X	random variable
projector	P_x	$\approx$	E	event
evolution operator	U	$\approx$	T	transformation*
probability rule	$\mathrm{tr}[P_x\vert\phi\rangle\langle\phi\vert]$	$\approx$	$\vec{e}[E]\cdot\vec{p}$	
post-measurement state	$P_x\vert\phi\rangle/\sqrt{P_X(x)}$	$\approx$	$P_{\vec{e}[E]}\vec{p}/\vec{e}[E]\cdot\vec{p}$	

This table highlights the fact that not only are there analogs in the quantum domain of objects in classical probability theory, but that they interact with each other in similar ways. Most notably, the probability rule is a “linear combination” of states and events in each case. The mathematical spaces in which the objects live is quite different, but nonetheless linearity is at the heart of both.

A couple of caveats are in order, corresponding to the starred items. First, state vectors are analogous to “sharp” probability distributions, which are those such that $p_j = \delta_{jk}$ for some k. This is because we can always find a measurement associated with a state $|\phi\rangle$ for which one outcome is certain, namely the measurement associated with the orthogonal projectors $P_\phi = |\phi\rangle\langle\phi|$ and $P_{\bar{\phi}} = \mathrm{id} - |\phi\rangle\langle\phi|$. Second, the unitary operators implementing time evolution are reversible, so they are analogous to reversible transformations (permutations) of the classical sample space.

Despite the elegance of the above analogy, there is one glaring omission: the sample space. What is the analog of the sample space in quantum theory? One could say that the $|\psi\rangle$ are the quantum version of the $\vec{s}_j$, since sharp distributions $\vec{p}$ are just the $\vec{s}_j$. But then comes the famous measurement problem: Why does the state (now a physical thing) evolve unitarily under “normal” dynamics but differently for measurement? Is measurement not a dynamical process? The view of $|\psi\rangle$ as akin to a probability distribution does not have this problem; even in classical probability theory the probability changes upon measurement. After all, the measurement reveals something about they underlying state of the system. But quantum-mechanically this approach leaves us in the awkward position of having no sample space to refer probability to. What is it about a quantum system that a measurement is supposed to reveal? If there is a useful sample space, why don’t we just formulate quantum mechanics directly in these terms? As we’ll see when discussing the Bell[1] inequalities, there are essentially no good options for an underlying sample space in these terms. So what should we do? Should we think of the state vector as a physical quantity, like $\vec{s}_j$, or just as a nice way to encode the probability of events, like $\vec{p}$? As far as I know, there’s no satisfactory answer to this question, though many are convinced by the various approaches to solve the riddle. One could also hope that the very

[1] John Stewart Bell, 1928 – 1990, Northern Irish physicist.

question is the wrong one to be asking, but it is by no means clear what the right question would be. Thus we are forced to live with the strange structure of quantum mechanics as we currently understand it.

4.2 Bipartite states and entanglement

The analogy presented in the previous section also does not deal with the last postulate, dealing with the structure of composite quantum systems. This structure is quite different than in the setting of classical probability theory, in particular due to the existence of *entangled* states. As we shall see, in one form or another entanglement is responsible for weirdness of quantum mechanics.

Consider an arbitrary state of a bipartite quantum system, i.e. a state $|\Psi\rangle$ on the space $\mathscr{H}_A \otimes \mathscr{H}_B$. Given orthonormal bases $\{|b_j\rangle\}$ and $\{|b'_k\rangle\}$ for these two spaces, any bipartite state can be expressed as

$$|\Psi\rangle = \sum_{j=1}^{d_A}\sum_{k=1}^{d_B} \Psi_{jk}|b_j\rangle \otimes |b'_k\rangle. \tag{4.5}$$

Here d_A (d_B) is the dimension of $\mathscr{H}_A$ ($\mathscr{H}_B$). Thinking of the components Ψ_{jk} as forming a $d_A \times d_B$ matrix, we may use the singular-value decomposition to form the *Schmidt decomposition* of the state. Let the singular value decomposition be $\Psi_{j,k} = U_{j,\ell}D_{\ell,\ell}[V^*]_{\ell,k}$. The entries of $D_{\ell,\ell}$ are all positive; let their values be $D_{\ell,\ell} = \lambda_\ell$. At most there are $d_{\min} = \min(d_A, d_B)$ nonzero singular values, so we may express $|\Psi\rangle$ as

$$\begin{aligned}
|\Psi\rangle &= \sum_{j=1}^{d_A}\sum_{k=1}^{d_B}\sum_{\ell}^{d_{\min}} U_{j,\ell}\lambda_\ell [V^*]_{\ell,k}|b_j\rangle \otimes |b'_k\rangle && (4.6)\\
&= \sum_{\ell}^{d_{\min}} \lambda_\ell \left(\sum_{j=1}^{d_A} U_{j,\ell}|b_j\rangle\right) \otimes \left(\sum_{k=1}^{d_B} V^*_{k,\ell}|b'_k\rangle\right) && (4.7)\\
&= \sum_{\ell}^{d_{\min}} \lambda_\ell |\xi_\ell\rangle \otimes |\eta_\ell\rangle, && (4.8)
\end{aligned}$$

where we have implicitly defined the states $|\xi_\ell\rangle$ and $|\eta_\ell\rangle$ in the last step. Since U and V are unitary, these two sets are each orthonormal bases. Thus, for any given bipartite state it is possible to find an orthonormal basis for each subsystem such that the coefficients of the global state are *diagonal* in this basis. Moreover, since the singular values are positive and the state is assumed to be normalized, the set $\{\lambda_\ell^2\}$ forms a probability distribution.

If there is only one nonzero *Schmidt coefficient* λ_ℓ, the state is a *product state* $|\Psi\rangle = |\xi\rangle \otimes |\eta\rangle$. On the other hand, if there is more than one nonzero Schmidt coefficient, the state is said to be *entangled*; if $\lambda_\ell = 1/\sqrt{d_m}$, the state is said to be *maximally entangled.*

The possibility of entanglement is due to the linear structure of the state space, and is responsible for the no-cloning argument we saw in the Introduction. Attempting to clone a general qubit state $|\psi\rangle = a|0\rangle + b|1\rangle$ results in the entangled state $a|0\rangle \otimes |0\rangle + b|1\rangle \otimes |1\rangle$.

4.3 Superdense coding and teleportation

There are two basic quantum information processing protocols involving entangled states of two systems which have no classical analog: superdense coding and teleportation.

4.3.1 Superdense Coding

The canonical maximally entangled state of two qubits is

$$|\Phi\rangle_{AB} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{AB}. \tag{4.9}$$

Consider the action of one of the Pauli operators on system B, say σ_x:

$$|\Phi_x\rangle_{AB} = \big(\mathrm{id}_A \otimes (\sigma_x)_B\big)\,|\Phi\rangle_{AB} = \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)_{AB}. \tag{4.10}$$

Clearly this state is orthogonal to $|\Phi\rangle$. What about σ_z?

$$|\Phi_z\rangle_{AB} = (\mathrm{id}_A \otimes (\sigma_z)_B)|\Phi\rangle_{AB} = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)_{AB}. \tag{4.11}$$

Also orthogonal to $|\Phi\rangle$, and to $|\Phi_x\rangle$. And σ_y:

$$|\Phi_y\rangle_{AB} = \Big(\mathrm{id}_A \otimes (-i\sigma_y)_B\Big)\,|\Phi\rangle_{AB} = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)_{AB}, \tag{4.12}$$

orthogonal to all others. We have constructed a basis for $\mathbb{C}^2 \otimes \mathbb{C}^2$ comprised of maximally entangled states, all related by Pauli operators on system B *alone.* As a side note, these four states turn out to be an interesting basis in terms of angular momentum. $|\Phi_y\rangle$ is the singlet, the state of two spin-$\frac{1}{2}$ systems having total angular momentum zero. The other states span the triplet space, having total angular momentum 1. $|\Phi_x\rangle$ is the eigenstate having $J_z = 0$, while $|\Phi_z\rangle$ is the eigenstate with $J_x = 0$ and $|\Phi\rangle$ $J_y = 0$. The latter two can be identified by direct calculation or by noticing that since σ_y commutes with rotations about the y axis, it cannot change the value of J_y.

Now return to the setup of our two separated parties, Alice and Bob. Alice would like to send a message to Bob, a message composed of two bits (sell stocks? buy gold?), but she has only got enough postage for either one classical bit or one quantum bit. Clearly one classical bit is insufficient. But quantum postage was even cheaper in the past, and Alice predicting that it would go up, sent a qubit to Bob back when the rates were cheap.

How does that help her now? Suppose she originally prepared $|\Phi\rangle_{AB}$ and then sent system A using the cheap postage. Now she can apply one of the 3 Pauli operators, or do nothing, to B and send this qubit to Bob. This creates one of the 4 entangled basis states $|\Phi_j\rangle_{AB}$, and Bob can read out the message using the measurement $P_j = |\Phi_j\rangle\langle\Phi_j|$

Notice that Alice managed to send 2 bits of information using just 1 qubit — when she sent the first one she had not yet made up her mind about selling stocks and buying gold. That is why this scheme is called superdense coding: one qubit is used to transfer 2 classical bits, though of course two qubits are ultimately involved (Bob needs 4 orthogonal projectors to read out the message).

4.3.2 Teleportation

Now imagine Alice and Bob are in the opposite situation: Instead of Alice wanting to send 2 classical bits and having only a quantum channel (plus preshared entanglement), she wants to send a qubit, but only has access to a classical channel. Can she somehow send the state to Bob using only a classical channel?

If that is all the resources they share, the answer is no. Alice could try to measure the qubit in some way, for instance to learn the values of the coefficients a and b in the expression $|\psi\rangle = a|0\rangle + b|1\rangle$ by building up statistics (since $\Pr(0) = |a|^2$ and never mind she also needs the relative phase between a and b), but she only has 1 copy of $|\psi\rangle$.

On the other hand, if Alice and Bob already share an entangled state, then it is possible to transfer $|\psi\rangle$ to Bob, and it only requires 2 bits! The "2 bits" are reminiscent of the 4 entangled states $|\Phi_j\rangle$ (called Bell states) used in superdense coding, and they play the same role as measurement in teleportation.

The protocol is very simple. Alice has a qubit prepared in $|\psi\rangle_{A'}$ as well as half of a maximally entangled state $|\Phi\rangle_{AB}$. She then measures her two systems in the Bell basis, producing a two-bit outcome. What happens when the outcome corresponds to $|\Phi\rangle$?

$$\begin{aligned}
{}_{A'A}\langle\Phi|\psi\rangle_{A'}|\Phi\rangle_{AB} &= {}_{A'A}\langle\Phi|\frac{1}{\sqrt{2}}(a|000\rangle + a|011\rangle + b|100\rangle + b|111\rangle)_{A'AB} && (4.13)\\
&= \frac{1}{2}(\langle 00| + \langle 11|)_{A'A}(a|000\rangle + a|011\rangle + b|100\rangle + b|111\rangle)_{A'AB} && (4.14)\\
&= \frac{1}{2}(a|0\rangle + b|1\rangle)_B && (4.15)\\
&= \frac{1}{2}|\psi\rangle_B. && (4.16)
\end{aligned}$$

The state has been transferred to Bob! The squared norm of the output tells us the probability, so the chance that Alice obtains result $|\psi\rangle$ is $\frac{1}{4}$. And what about the other results?

To figure this out we can use a different method: write $|\psi\rangle_{A'}|\Phi\rangle_{AB}$ in the $|\Phi_j\rangle_{A'A}|k\rangle_B$ basis.

$$\begin{aligned}
|\psi\rangle_{A'}|\Phi\rangle_{AB} &= \frac{1}{\sqrt{2}}(a|000\rangle + a|011\rangle + b|100\rangle + b|111\rangle)_{A'AB} && (4.17)\\
&= \frac{1}{2}\Big[a\left(|\Phi\rangle|0\rangle + |\Phi_z\rangle|0\rangle + |\Phi_x\rangle|1\rangle + |\Phi_y\rangle|1\rangle\right)\\
&\qquad +b\left(|\Phi_x\rangle|0\rangle - |\Phi_y\rangle|0\rangle + |\Phi\rangle|1\rangle - |\Phi_z\rangle|1\rangle\right)\Big] && (4.18)\\
&= \frac{1}{2}\Big[|\Phi\rangle(a|0\rangle + b|1\rangle) + |\Phi_x\rangle(a|1\rangle + b|0\rangle)\\
&\qquad |\Phi_y\rangle(a|1\rangle - b|0\rangle) + |\Phi_z\rangle(a|0\rangle + b|1\rangle)\Big] && (4.19)\\
&= \frac{1}{2}\Big[|\Phi\rangle|\psi\rangle + |\Phi_x\rangle\sigma_x|\psi\rangle + |\Phi_y\rangle(-i\sigma_y)|\psi\rangle + |\Phi_z\rangle\sigma_z|\psi\rangle\Big]. && (4.20)
\end{aligned}$$

Notice how each term is of the form $|\Phi_j\rangle_{A'A}\sigma_j|\psi\rangle_B$, meaning that if Alice measures $A'A$ in the Bell basis and communicates the result to Bob, he can apply the corresponding Pauli operator to obtain the input state $|\psi\rangle$. Alice needs 2 bits to describe the outcome, and since each term has the same weight, the probability of every outcome is $\frac{1}{4}$.

4.4 Further Reading

The Schmidt decomposition was first presented in [17].

5 The Structure of Quantum States

The postulates of quantum mechanics presented in the previous chapter deal only with isolated systems. Moreover, they do not directly allow classical information to be included in the quantum description. But such a description should be possible, according to the theme of the course.

In this chapter and the next e shall see that for parts of a larger system or when including classical information, states are no longer rays, measurements are no longer projectors, and dynamics is no longer given by unitary operators. In this chapter we are specifically concerned with the structure of quantum states.

5.1 Density operators

5.1.1 Mixtures of states

Consider a quantum system $\mathcal{H}_A$ whose state, a pure state, depends on a classical value (random variable) Z and let $|\phi_z\rangle\langle\phi_z|_A \in \mathcal{S}(\mathcal{H}_A)$ be the pure state of the system conditioned on the event $Z = z$. Note that the states $|\phi_z\rangle$ need not be orthogonal. 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. This setup is described by the *ensemble* of states $\{P_Z(z), |\phi\rangle_z\}$.

Assume now that the system $\mathcal{H}_A$ undergoes an evolution U_A followed by a measurement $O_A = \sum_x xP_x$. 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

$$P_{X|Z=z}(x) = \mathrm{tr}(P_x U_A |\phi_z\rangle\langle\phi_z|_A U_A^*). \tag{5.1}$$

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

$$P_X(x) = \sum_z P_Z(z) P_{X|Z=z}(x). \tag{5.2}$$

By linearity, this can be rewritten as

$$P_X(x) = \mathrm{tr}(P_x U_A \rho_A U_A^*). \tag{5.3}$$

where we have implicitly defined the *density operator*

$$\rho_A = \sum_z P_Z(z) |\phi_z\rangle\langle\phi_z|_A. \tag{5.4}$$

Observe that ρ_A has the following properties:

$$\rho_A \geq 0, \tag{5.5}$$

$$\mathrm{tr}\rho_A = 1. \tag{5.6}$$

Operators satisfying these conditions are called *density operators*, and for a state space $\mathcal{H}$, the set of density operators is denoted by $\mathcal{S}(\mathcal{H})$. By the spectral decomposition, we can always express ρ in terms of eigenvalues and eigenvectors as $\rho = \sum_k p_k |b_k\rangle\langle b_k|$; the eigenvalues form a probability distribution since the operator is normalized. States as we defined them originally are equivalent to density operators of the form $\rho = |\phi\rangle\langle\phi|$ and are called *pure states*. Pure states have only one nonzero eigenvalue and therefore satisfy $\mathrm{tr}\rho^2 = 1$. States with more than one nonzero eigenvalue are called *mixed states* since they are mixtures (convex combinations) of their eigenvectors.

Alternatively, expression (5.3) can be obtained by applying the postulates of Section 4.1 directly to the density operator ρ_A defined above. In particular, by replacing $|\phi\rangle\langle\phi|$ with the density operator ρ in (4.2). In other words, from the point of view of an observer with no access to Z, the situation is consistently characterized by ρ_A.

According to the theme of this course, the information contained in the classical random variable Z should be manifested physically. It is an important feature of the framework we are developing that Z can also be described in the density operator formalism. 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 above scenario would read

$$\rho_{AZ} = \sum_z P_Z(z)\rho_A^z \otimes |b_z\rangle\langle b_z|, \tag{5.7}$$

where $\{|b_z\rangle\}_z$ is a family of orthonormal vectors on $\mathcal{H}_Z$. States of this form will be said to be *classical on $\mathcal{H}_B$ with respect to* $\{|b_z\rangle\}_z$ and are called classical-quantum states, or CQ states.

In the previous chapter we defined entanglement of bipartite pure states, but mixed states can be entangled, too. Entanglement in the pure state case was defined by any state which is not the tensor product of states on the constituent systems. In general, product states take the form $\rho_{AB} = \theta_A \otimes \varphi_B$ and can be regarded as classical in the sense that there is a well-defined state for each constituent system. This notion continues to hold for mixtures of product states, since then each system again has a well-defined state conditional on the parameter of the mixture:

$$\sigma = \sum_k p_k \rho_k \otimes \varphi_k. \tag{5.8}$$

Any quantum state of the form (5.8) is called *separable* and any state which is not separable is said to be entangled.

5.1.2 Reduced states

Another motivation for density operators comes from examining a subsystem of a larger composite system in a pure quantum state. One striking feature of entangled states on $\mathcal{H}_A \otimes \mathcal{H}_B$ is that, to an observer with no access to B, the state of A does not correspond to a fixed vector $|\phi\rangle \in \mathcal{H}_A$, but rather a density operator. To see this more concretely, consider the measurement of an observable O_A on one part of a bipartite system in state $|\Psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$. The expectation value of O_A is given by

$$\begin{aligned}\langle O_A\rangle_\Psi &= \mathrm{tr}[O_A \otimes \mathrm{id}_B |\Psi\rangle\langle\Psi|] && (5.9)\\ &= \mathrm{tr}[O_A \mathrm{tr}_B[|\Psi\rangle\langle\Psi|]], && (5.10)\end{aligned}$$

where we have used the partial trace from §A.5. Thus we can define $\rho_A = \mathrm{tr}_B[|\Psi\rangle\langle\Psi|]$, which pertains only to system A which allows us to calculate all expectation values and probabilities. It is often called the *reduced state*. The existence of reduced states is an important *locality* feature of quantum theory. Since any action performed on B will not affect ρ_A, it is impossible to influence system A by local action on B.

Using the Schmidt decomposition, we can write the above calculation out in terms of components, like so:

$$\begin{aligned}\langle O_A\rangle_\Psi &= \langle\Psi|(O_A\otimes \mathrm{id}_B)|\Psi\rangle &(5.11)\\ &=\sum_{jk}\lambda_j\lambda_k\langle\xi_j|\otimes\langle\eta_j|(O_A\otimes\mathrm{id}_B)|\xi_k\rangle\otimes|\eta_k\rangle &(5.12)\\ &=\sum_{jk}\lambda_j\lambda_k\langle\xi_j|O_A|\xi_k\rangle\langle\eta_j|\eta_k\rangle &(5.13)\\ &=\sum_k\lambda_k^2\langle\xi_k|O_A|\xi_k\rangle &(5.14)\\ &=\mathrm{tr}[O_A\sum_k\lambda_k^2|\xi_k\rangle\langle\xi_k|]. &(5.15)\end{aligned}$$

Comparing with the above, we have found that $\rho_A=\mathrm{tr}_B[|\Psi\rangle\langle\Psi|]=\sum_k\lambda_k^2|\xi_k\rangle\langle\xi_k|$. This clearly satisfies (5.5) and (5.6) and is therefore a density operator.

5.1.3 Purification of mixed states

The notion of a density operator was motivated by examining mixtures of pure quantum states. In the previous section we have also seen that all reduced states of a composite system are density operators. Can we connect these two viewpoints and regard any density operator ρ as the reduced state ρ_A of a pure state $|\Psi\rangle_{AB}$ on a larger system? The answer is yes, and such a pure state $|\Psi\rangle_{AB}$ is called a *purification* of ρ. Regarding a mixed state as part of a pure state in this way is done very often in quantum information theory and is called "going to the church of the larger Hilbert space".

Given an ensemble decomposition of a density operator $\rho=\sum_z P_Z(z)|\phi_z\rangle\langle\phi_z|$ as in (5.4), it is easy to construct a purification of ρ. Simply invent an additional system B and define

$$|\Psi\rangle_{AB}=\sum_{z=1}^n\sqrt{P_Z(z)}|\phi_z\rangle_A\otimes|b_z\rangle_B. \tag{5.16}$$

This also works for CQ states, like $\rho_{AZ}=\sum_z P_Z(z)\rho_A^z\otimes|b_z\rangle\langle b_z|_Z$ in (5.7). Now invent two additional systems B and Z' and define

$$|\Psi\rangle_{ABZZ'}=\sum_z\sqrt{P_Z(z)}|\varphi_z\rangle_{AB}\otimes|b_z\rangle_Z\otimes|b_z\rangle_{Z'}, \tag{5.17}$$

where $|\varphi_z\rangle_{AB}$ is a purification of ρ_A^z.

We can also construct a purification in a "component-free" manner, by making use of the *canonical maximally-entangled state* on $\mathcal{H}_A\otimes\mathcal{H}_B$ with $\mathcal{H}_A\simeq\mathcal{H}_B$ of dimension d,

$$|\Phi\rangle_{AB}=\frac{1}{\sqrt{d}}\sum_{k=1}^d|b_k\rangle_A\otimes|b_k\rangle_B. \tag{5.18}$$

Call $|\Omega\rangle$ the unnormalized version of this state, i.e.

$$|\Omega\rangle_{AB}=\sum_{k=1}^d|b_k\rangle_A\otimes|b_k\rangle_B. \tag{5.19}$$

Observe that the partial trace of $|\Omega\rangle$ is the identity: $\mathrm{tr}_B[|\Omega\rangle\langle\Omega|] = \mathrm{id}_A$. Then it is easy to verify that the state

$$|\Psi\rangle_{AB} = (\sqrt{\rho}_A \otimes \mathrm{id}_B)|\Omega\rangle_{AB} \tag{5.20}$$

is a purification of ρ. Here $\sqrt{\rho}$ is the positive operator whose square is ρ; by the spectral decomposition, if $\rho = \sum_k p_k |b_k\rangle\langle b_k|$ for some orthonormal basis $\{|b_k\rangle\}$, then $\sqrt{\rho} = \sum_k \sqrt{p_k}|b_k\rangle\langle b_k|$. Indeed, any state of the form

$$|\Psi\rangle_{AB} = (\sqrt{\rho}_A U_A \otimes V_B)|\Omega\rangle_{AB}, \tag{5.21}$$

for unitary U and V is also a purification of ρ_A.

The Schmidt decomposition of a purification $|\Psi\rangle_{AB}$ of ρ_{AB} is directly related to the eigendecomposition of ρ itself: The Schmidt basis $\{|\xi_k\rangle\}$ of system A is the eigenbasis of ρ. Indeed, we already implicitly encountered this fact in (5.15). The partial trace of a state in Schmidt form immediately gives an eigendecomposition, so the Schmidt basis vectors must be the eigenstates of the original density operator.

Moreover, the Schmidt decomposition immediately implies that any two purifications of a state ρ must be related by a unitary operation on the purifying system B. Suppose $|\Psi\rangle_{AB}$ and $|\Psi'\rangle_{AB}$ are two purifications of ρ_A. In view of the relation to the eigendecomposition, the Schmidt forms of the two states must be

$$|\Psi\rangle_{AB} = \sum_k \sqrt{p_k}|\xi_k\rangle \otimes |\eta_k\rangle \tag{5.22}$$

$$|\Psi'\rangle_{AB} = \sum_k \sqrt{p_k}|\xi_k\rangle \otimes |\eta'_k\rangle. \tag{5.23}$$

But both $\{|\eta_k\rangle\}$ and $\{|\eta'_k\rangle\}$ are orthonormal bases, so they must be related by some unitary transformation U: $U|\eta_k\rangle = |\eta'_k\rangle$. Therefore, we have shown that

Lemma 5.1.1 (Unitary relation of purifications). *For any two purifications $|\Psi\rangle_{AB}$ and $|\Psi'\rangle_{AB}$ of a state ρ_A, there exists a unitary U_B such that $|\Psi'\rangle_{AB} = (\mathrm{id}_A \otimes U_B)|\Psi\rangle_{AB}$.*

This statement might appear to be inconsistent with (5.21), but closer inspection of the state $|\Omega\rangle$ reveals the useful fact that

$$U_A \otimes \mathrm{id}_B|\Omega\rangle_{AB} = \mathrm{id}_A \otimes U_B^T|\Omega\rangle, \tag{5.24}$$

where U^T is the transpose of U relative to the basis $\{|b_j\rangle\}$ which is used to define $|\Omega\rangle$. More specifically, $U^T = \sum_{jk}|b_j\rangle\langle b_k|\langle b_k|U|b_j\rangle$. Thus, (5.21) could equally-well be written as $\sqrt{\rho_A} \otimes (VU^T)_B|\Omega\rangle$ which is consistent with Lemma 5.1.1.

The unitary freedom in choosing a purification of a density operator translates into a freedom in the *decomposition* of the density operator into pure states. Equation (5.4) presents a generic decomposition, but for concreteness consider the state $\rho_A = \frac{1}{2}|b_0\rangle\langle b_0| + \frac{1}{2}|b_1\rangle\langle b_1|$, which we may interpret as describing the fact that A is prepared in one of the two basis states $|b_0\rangle$ with equal probability. However, the decomposition is not unique, as the same state could be written as

$$\rho_A = \frac{1}{2}|\tilde{b}_0\rangle\langle\tilde{b}_0| + \frac{1}{2}|\tilde{b}_1\rangle\langle\tilde{b}_1| \tag{5.25}$$

where $|\tilde{b}_0\rangle := \frac{1}{\sqrt{2}}(|b_0\rangle + |b_1\rangle)$ and $|\tilde{b}_1\rangle := \frac{1}{\sqrt{2}}(|b_0\rangle - |b_1\rangle)$. That is, the system could equally-well be interpreted as being prepared either in state $|\tilde{b}_0\rangle$ or $|\tilde{b}_1\rangle$, each with probability $\frac{1}{2}$.

All possible pure state ensemble decompositions of a density operator are related in a unitary way, via the purification. Suppose that

$$\rho = \sum_k p_k |\phi_k\rangle\langle\phi_k| = \sum_j q_j |\psi_j\rangle\langle\psi_j| \tag{5.26}$$

are two decompositions of ρ. From these, we can construct the purifications

$$|\Psi_1\rangle_{AB} = \sum_k \sqrt{p_k}|\phi_k\rangle_A \otimes |b'_k\rangle_B \quad \text{and} \tag{5.27}$$

$$|\Psi_2\rangle_{AB} = \sum_j \sqrt{q_j}|\psi_j\rangle_A \otimes |b'_j\rangle_B. \tag{5.28}$$

As these are purifications of the same state, there must be a unitary U such that $\mathrm{id}_A \otimes U_B|\Psi_1\rangle_{AB} = |\Psi_2\rangle_{AB}$. But then we have

$$\begin{aligned}
\sqrt{q_k}|\psi_k\rangle &= \sum_j \sqrt{q_j}|\psi_j\rangle\langle b'_k|b_{j'}\rangle && (5.29)\\
&= {}_B\langle b'_k|\Psi_2\rangle_{AB} && (5.30)\\
&= \sum_j \sqrt{p_j}|\phi_j\rangle\langle b'_k|U|b'_j\rangle && (5.31)\\
&= \sum_j U_{kj}\sqrt{p_j}|\phi_j\rangle. && (5.32)
\end{aligned}$$

Thus, we have shown

Lemma 5.1.2 (Unitary relation of ensemble decompositions). *For a density operator ρ with ensemble decompositions $\{p_k, |\phi_k\rangle\}$ and $\{q_k, |\psi_k\rangle\}$, there exists a unitary matrix U such that*

$$\sqrt{q_k}|\psi_k\rangle = \sum_j U_{kj}\sqrt{p_j}|\phi_j\rangle. \tag{5.33}$$

5.1.4 Comparison of probability distributions and quantum states

Looking back at the analogy of quantum theory with classical probability theory, it becomes apparent that density operators are the proper quantum version of probability distributions. This holds for two reasons. First, just as $\vec{p}$ can be regarded as a convex combination of sharp distributions, so too are density operators mixtures of pures states. Pure states are pure ans sharp distributions sharp, because they cannot be expressed as a nontrivial convex combination of other states or distributions. Secondly, neither for unsharp $\vec{p}$ nor for mixed ρ can find an event which is certain to occur.

Purifications do not exist in classical probability theory. That is, given a distribution $\vec{p}_A$, there is no sharp joint distribution $\vec{p}_{AB}$ over two random variables whose marginal is $\vec{p}_A$. Any sharp distribution on $\vec{p}_{AB}$ has components $(\vec{p}_{AB})_{jk} = \delta_{jj'}\delta_{kk'}$ for some j' and k'. The marginal is clearly $(\vec{p}_A)_j = \delta_{jj'}$, which is itself sharp. Only in the formalism of quantum theory can the "distribution" of the compound system be sharp, even though the marginal "distributions" are not.

5.2 Distance measures between states

5.2.1 Trace distance

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 §2.7.

Definition 5.2.1. The *trace distance* between two density operators ρ and σ on a Hilbert space $\mathscr{H}$ is defined by

$$\delta(\rho,\sigma) := \frac{1}{2}\left\|\rho-\sigma\right\|_1. \tag{5.34}$$

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 §2.7. 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

$$\delta(\rho,\sigma) = \delta(P,Q). \tag{5.35}$$

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 5.2.2. *Let $\rho,\sigma \in \mathscr{S}(\mathscr{H})$. Then*

$$\delta(\rho,\sigma) = \max_O \delta(P,Q) \tag{5.36}$$

where the maximum ranges over all observables $O \in \mathrm{Herm}\mathscr{H}$ 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.

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

$$R = \sum_{i:\alpha_i \geq 0} \alpha_i |e_i\rangle\langle e_i| \tag{5.37}$$

$$S = -\sum_{i:\alpha_i < 0} \alpha_i |e_i\rangle\langle e_i|, \tag{5.38}$$

that is,

$$\Delta = R - S \tag{5.39}$$

$$|\Delta| = R + S. \tag{5.40}$$

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

$$\delta(P,Q) = \frac{1}{2}\sum_x \left|P(x) - Q(x)\right| = \frac{1}{2}\sum_x \left|\mathrm{tr}(P_x\rho) - \mathrm{tr}(P_x\sigma)\right| = \frac{1}{2}\sum_x \left|\mathrm{tr}(P_x\Delta)\right|. \tag{5.41}$$

Now, using (5.39) and (5.40),

$$\left|\mathrm{tr}(P_x\Delta)\right| = \left|\mathrm{tr}(P_x R) - \mathrm{tr}(P_x S)\right| \leq \left|\mathrm{tr}(P_x R)\right| + \left|\mathrm{tr}(P_x S)\right| = \mathrm{tr}(P_x|\Delta|), \tag{5.42}$$

where the last equality holds because of (A.29). Inserting this into (5.41) and using $\sum_x P_x = \mathrm{id}$ gives

$$\delta(P,Q) \leq \frac{1}{2}\sum_x \mathrm{tr}\big(P_x|\Delta|\big) = \frac{1}{2}\mathrm{tr}\big(|\Delta|\big) = \frac{1}{2}\|\Delta\|_1 = \delta(\rho,\sigma). \tag{5.43}$$

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 (5.42) 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 5.2.2 is that the trace distance between two states ρ and σ is related to the *maximum distinguishing probability*, the maximum probability by which a difference between ρ and σ can be detected, just as in the case of probability distributions in Lemma 2.7.2. Another consequence of Lemma 5.2.2 is that the trace distance cannot increase under the partial trace, as stated by the following lemma.

Lemma 5.2.3. *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*

$$\delta(\rho_A,\sigma_A) \leq \delta(\rho_{AB},\sigma_{AB}). \tag{5.44}$$

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 5.2.2

$$\delta(\rho_A,\sigma_A) = \delta(P,Q). \tag{5.45}$$

Consider now the observable O_{AB} on the joint system defined by $O_{AB} := O_A \otimes \mathrm{id}_B$. It follows from property (A.31) 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 5.2.2,

$$\delta(\rho_{AB},\sigma_{AB}) \geq \delta(P,Q). \tag{5.46}$$

The assertion follows by combining (5.45) and (5.46). □

5.2.2 Fidelity

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

$$F(\rho,\sigma) := \left\|\sqrt{\rho}\sqrt{\sigma}\right\|_1, \tag{5.47}$$

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 $\delta(|\phi\rangle\langle\phi|, |\psi\rangle\langle\psi|)$.

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

$$F(\rho, |\psi\rangle\langle\psi|) = \|\sqrt{\rho}\sqrt{\sigma}\|_1 = \mathrm{tr}\left(\sqrt{\sqrt{\sigma}\rho\sqrt{\sigma}}\right) = \mathrm{tr}\left(\sqrt{|\psi\rangle\langle\psi|\rho|\psi\rangle\langle\psi|}\right) = \sqrt{\langle\psi|\rho|\psi\rangle}. \tag{5.48}$$

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

$$F(\phi, \psi) = |\langle\phi|\psi\rangle|. \tag{5.49}$$

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

The following statement from Uhlmann[1] generalizes this statement to arbitrary states.

Theorem 5.2.5 (Uhlmann). *Let ρ_A and σ_A be density operators on a Hilbert space $\mathcal{H}_A$. Then*

$$F(\rho_A, \sigma_A) = \max_{\phi_{AB}, \psi_{AB}} F(\phi_{AB}, \psi_{AB}). \tag{5.50}$$

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

Proof. Using Lemma 5.1.1, for some choice of unitaries U and V we can write

$$\begin{aligned} F(\phi_{AB}, \psi_{AB}) &= |\langle\phi|\psi\rangle| & (5.51)\\ &= |\langle\Omega|\sqrt{\rho_A}\sqrt{\sigma_A} \otimes V_B^* U_B |\Omega\rangle| & (5.52)\\ &= |\mathrm{tr}[\sqrt{\rho}\sqrt{\sigma} U^T (V^*)^T]|. & (5.53)\end{aligned}$$

Thus, $\max_{\phi_{AB}, \psi_{AB}} F(\phi_{AB}, \psi_{AB})$ amounts to $\max_U |\mathrm{tr}[\sqrt{\rho}\sqrt{\sigma} U]|$. By Lemma A.7.2 we then have

$$\begin{aligned} \max_{\phi_{AB}, \psi_{AB}} F(\phi_{AB}, \psi_{AB}) &= \max_U |\mathrm{tr}[U\sqrt{\rho_A}\sqrt{\sigma_A}]| & (5.54)\\ &= \|\sqrt{\rho_A}\sqrt{\sigma_A}\|_1 & (5.55)\\ &= F(\rho_A, \sigma_A), & (5.56)\end{aligned}$$

completing the proof. □

By Uhlmann's theorem it is clear that the $0 \leq F(\rho, \sigma) \leq 1$. It also immediately implies the monotonicity of fidelity under partial trace,

Lemma 5.2.6. *Let ρ_{AB} and σ_{AB} be bipartite states. Then*

$$F(\rho_{AB}, \sigma_{AB}) \leq F(\rho_A, \sigma_A). \tag{5.57}$$

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

$$F(\rho_{AB}, \sigma_{AB}) = F(\rho_{ABC}, \sigma_{ABC}). \tag{5.58}$$

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

$$F(\rho_A, \sigma_A) \geq F(\rho_{ABC}, \sigma_{ABC}). \tag{5.59}$$

Combining (5.58) and (5.59) concludes the proof. □

[1] Armin Gotthard Uhlmann, born 1930, German theoretical physicist.

5.2.3 Relation between trace distance and fidelity

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

$$\delta(\phi,\psi)=\sqrt{1-F(\phi,\psi)^2}. \tag{5.60}$$

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 $\{\phi,\phi^\perp\}$,

$$|\phi\rangle\langle\phi|=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} \tag{5.61}$$

$$|\psi\rangle\langle\psi|=\begin{pmatrix}|\alpha|^2 & \alpha\beta^*\\ \alpha^*\beta & |\beta|^2\end{pmatrix} \tag{5.62}$$

In particular, the trace distance takes the form

$$\delta(\phi,\psi)=\frac{1}{2}\big\||\phi\rangle\langle\phi|-|\psi\rangle\langle\psi|\big\|_1=\frac{1}{2}\left\|\begin{pmatrix}1-|\alpha|^2 & -\alpha\beta^*\\ -\alpha^*\beta & -|\beta|^2\end{pmatrix}\right\|_1. \tag{5.63}$$

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

$$\delta(\phi,\psi)=\frac{1}{2}(|\alpha_0|+|\alpha_1|)=\beta. \tag{5.64}$$

Furthermore, by the definition of β, we have

$$\beta=\sqrt{1-|\langle\phi|\psi\rangle|^2}. \tag{5.65}$$

The assertion (5.60) then follows from (5.49).

Equality (5.60) together with Uhlmann's theorem are sufficient to prove one direction of the following lemma.

Lemma 5.2.7. *Let ρ and σ be density operators. Then*

$$1-F(\rho,\sigma)\leq\delta(\rho,\sigma)\leq\sqrt{1-F(\rho,\sigma)^2}. \tag{5.66}$$

Proof. We only prove the second inequality. For a proof of the first, see [2, §9.2.3].

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

$$F(\rho_A,\sigma_A)=F(\rho_{AB},\sigma_{AB}) \tag{5.67}$$

as in Uhlmann's theorem. Combining this with equality (5.60) and Lemma 5.2.3, we find

$$\sqrt{1-F(\rho_A,\sigma_A)^2}=\sqrt{1-F(\rho_{AB},\sigma_{AB})^2}=\delta(\rho_{AB},\sigma_{AB})\geq\delta(\rho_A,\sigma_A). \tag{5.68}$$

□

6 Quantum Measurements and Operations

We have seen in the previous chapter that, as long as we are only interested in the observable quantities of subsystem $\mathscr{H}_A$ of a larger state space $\mathscr{H}_A \otimes \mathscr{H}_B$, 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 and the measurement process is not modelled as a physical operation.

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 $\mathscr{H}_A$ and $\mathscr{H}_B$. The basic mathematical objects to be introduced in this context are *completely positive maps (CPMs)* and *positive operator valued measures (POVMs)*.

6.1 Generalized measurements

6.1.1 The von Neumann picture of measurement

The description of measurement in the axioms is an awkward mixture of quantum and classical. The central problem is that if "measurement" produces a (classical) outcome, should this information not be manifested physically, presumably as a quantum system? So how can there be an "outcome" at all? These are tricky conceptual problems that we will not attempt to answer in this course. However, we should look at the (formal) measurement procedure a little more carefully to see how it fits with the notions both of decompositions and purifications of mixed states. What we will end up with is the von Neumann picture of measurement, introduced in [18].

We have said that measurements are described by a set of projection operators $\{P_x\}$, one P_x for every outcome x. Given a state ρ, we saw in §5.1 that the xth outcome occurs with probability $p_x = \mathrm{tr}[P_x\rho]$. But what about the post-measurement state? Since any density operator has a decomposition into a convex combination of pure states, we can "lift" the structure of post-measurement states from the case of pure to mixed inputs. Suppose $\rho = \sum_z p_z|\phi_z\rangle\langle\phi_z|$ for some pure states $|\phi_z\rangle$. For each z, the measurement produces the state $|\psi_{x,z}\rangle = P_x|\phi_z\rangle/\sqrt{\langle\phi_z|P_x|\phi_z\rangle}$ with probability $P_{X|Z=z} = \langle\phi_z|P_x|\phi_z\rangle$. According to §5.1.1, the density operator describing the post-measurement state must be the mixture of the $|\psi_{x,z}\rangle$ according to the distribution $P_{Z|X=x}$. Therefore, we find

$$\rho_x = \sum_z P_{Z|X=x}(z)|\psi_{x,z}\rangle\langle\psi_{x,z}| \tag{6.1}$$

$$= \sum_z \frac{P_{Z|X=x}(z)}{P_{X|Z=z}(x)} P_x|\phi_z\rangle\langle\phi_z|P_x \tag{6.2}$$

$$= \sum_z \frac{P_Z(z)}{P_X(x)} P_x|\phi_z\rangle\langle\phi_z|P_x \tag{6.3}$$

$$= \frac{1}{P_X(x)} P_x\rho P_x. \tag{6.4}$$

The final expression is independent of the decomposition, as it ought to be if density operators are a complete description of the quantum state, as we argued in the previous chapter.

The above calculation is only for one outcome x, but of course the measurement produces an entire ensemble of states, ρ_x with probability $P_X(x)$. To an observer without access to the measurement result, the description of the state after the measurement is given by

$$\rho' = \sum_x P_X(x)\rho_x = \sum_x P_x \rho P_x. \tag{6.5}$$

Note that $\rho' \neq \rho$ generally. In quantum mechanics, performing a measurement and forgetting the result nonetheless changes the state of the system!

Let us assume for the moment that both the input state $\rho = |\phi\rangle\langle\phi|$ and the ρ_x are pure states and consider the purification of average post-measurement state in (6.5). Does it have any physical meaning? A purification is given by

$$|\Psi\rangle_{AB} = \sum_x P_x |\phi\rangle_A \otimes |b_x\rangle_B. \tag{6.6}$$

The interesting thing is that we can describe the transformation

$$|\phi\rangle_A \otimes |b_0\rangle_B \mapsto |\Psi\rangle_{AB} \tag{6.7}$$

with an operator $U = \sum_x (P_x)_A \otimes (V_x)_B$ which is *unitary*. Here V_k is a unitary operator taking $|b_0\rangle$ to $|b_x\rangle$. For concreteness, we can set $V_k = V^k$ for $V = \sum_j |b_{j\oplus 1}\rangle\langle b_j|$. Unitarity of U is then easy:

$$UU^* = \sum_{xx'} P_x P_x^* \otimes V_x V_x^* = \sum_x P_x \otimes V_x V_x^* \tag{6.8}$$

$$= \sum_x P_x \otimes \mathrm{id} = \mathrm{id} \otimes \mathrm{id}. \tag{6.9}$$

We have arrived, in a somewhat nonstandard fashion, at von Neumann's picture of measurement. The idea is that measurement can be viewed as a fully coherent process (just involving unitary transformations) which establishes a correlation between the system being measured (A) and a system storing the measurement result (B). Actually, this procedure does more than correlate A and B, it *entangles* them.

The measurement process is not quite finished though, since $|\Psi\rangle_{AB}$ describes a coherent superposition of *all* possible outcomes. To realize a particular outcome, we have to assume that B is somehow itself measured in the $\{|b_x\rangle\}$ basis. So how does this really solve the measurement problem? In order to measure B, we need to correlate it with C, and then we will need to measure C, requiring correlation with D, and so on *ad infinitum*! All true, but this is the best we are going to be able to do with a fully coherent description.

The unitary part of the measurement process produces the state in (6.6), and

$$|\xi\rangle_{ABC} = \sum_x P_x |\phi\rangle_A \otimes |b_x\rangle_B \otimes |b_x\rangle_C \tag{6.10}$$

if taken to the next step. In the former case, tracing out system B leaves the density operator $\rho_A = \sum_x P_x |\phi\rangle\langle\phi| P_x$, while in the latter case tracing out C leaves the correlated, but not entangled (classical-quantum) state $\rho_{AB} = \sum_x P_x |\phi\rangle\langle\phi| P_x \otimes |b_x\rangle\langle b_x|$.

6.1.2 Mixtures of measurements & POVMs

Measurements can themselves be "mixed" in the way we saw quantum states can be mixed in §5.1.1. In fact, we already implicitly saw an example of this in the introduction, when discussing quantum key distribution in §1.4. Recall that Bob's task was to measure either $\{P_0 = |0\rangle\langle 0|, P_1 = |1\rangle\langle 1|\}$ or $\{P_\pm = |\pm\rangle\langle\pm|\}$ with equal probability. If we let X be the bit describing which measurement is made and Y its outcome (+ counts as 0, − as 1), and $P_{x,y}$ the corresponding projector, then the probability distribution when the state is ρ is given by

$$P_{XY}(x,y) = \frac{1}{2}\mathrm{tr}[P_{x,y}\rho] = \mathrm{tr}[\Lambda_{x,y}\rho]. \tag{6.11}$$

Here we have implicitly defined the operators $\Lambda_{x,y}$. Observe that these sum to id, just as we insisted for any projective measurement, although they are no longer disjoint.

This example suggests that we should allow arbitrary operators $\{\Lambda_x\}_{x\in\mathscr{X}}$ as long as they satisfy two conditions:

1. $\Lambda_x \geq 0$ for all $x \in \mathscr{X}$,
2. $\sum_{x\in\mathscr{X}} \Lambda_x = \mathrm{id}$.

The two conditions ensure that the probability rule $\mathrm{tr}[\Lambda_x\rho]$ really does yield probabilities.

Such a set is called, somewhat awkwardly, a *positive operator-valued measure* or POVM. The name comes from more a more generic context in which the measurement outcomes are elements of an arbitrary measure space, not a discrete set as we have implicitly chosen here. For instance, the outcome of the measurement might be the position of a particle, which we would associate with elements of $\mathbb{R}$. Then, to each measurable set in the measure space corresponds a positive operator, with the constraint that the operator corresponding to the whole space be the identity.

6.1.3 The Naimark extension

Generalized measurements—POVMs—are consistent with the original axioms in the same way that density operators are: They are equivalent to usual projection measurements on a larger space, like density operators are equivalent to pure states on a larger space. This construction is known as the *Naimark*[1] *extension*.

In fact, we have already met the Naimark extension in §6.1.1. One method of realizing a POVM is the von Neumann approach. For a set of projectors we saw that $U_{AB} = \sum_x (P_x)_A \otimes (V_x)_B$ is a unitary operator taking $|\psi\rangle_A \otimes |0\rangle_B$ to $|\psi'\rangle_{AB}$ such that measuring B with $(P_x)_B$ realizes measurement of $(P_x)_A$ on A. To extend this to an arbitrary POVM with elements Λ_x, define U_{AB} implicitly by the action

$$U_{AB}|\psi\rangle_A \otimes |0\rangle_B = \sum_x \sqrt{\Lambda_x}|\psi\rangle_A \otimes |b_x\rangle_B = |\psi'\rangle_{AB}. \tag{6.12}$$

The probability of outcome x when measuring B with P_x is

$$P_X(x) = \mathrm{tr}[\psi'_{AB}\mathrm{id}_A \otimes (P_x)_B] = \mathrm{tr}[\psi\Lambda_x], \tag{6.13}$$

as intended. But is U_{AB} unitary? Its action is not fully specified, but note that as a map from $\mathscr{H}_A$ to $\mathscr{H}_A \otimes \mathscr{H}_B$ it is an isometry, meaning it preserves the inner product. Letting $|\phi'\rangle_{AB} = U_{AB}|\phi\rangle_A \otimes |0\rangle_B$,

[1] Mark Aronovich Naimark, 1909-1978, Soviet mathematician.

it follows that

$$\langle\phi'|\psi'\rangle = \sum_{x,x'}(\langle\phi|\sqrt{\Lambda_{x'}}\otimes\langle b_{x'}|)(\sqrt{\Lambda_x}|\psi\rangle\otimes|b_x\rangle) \tag{6.14}$$

$$= \sum_x \langle\phi|\Lambda_x|\psi\rangle \tag{6.15}$$

$$= \langle\phi|\psi\rangle. \tag{6.16}$$

Partial isometries from one space $\mathscr{H}$ to another, bigger space $\mathscr{H}'$ can always be extended to be unitaries from $\mathscr{H}'$ to itself. Here $\mathscr{H} = \mathscr{H}_A \otimes |0\rangle_B$ and $\mathscr{H}' = \mathscr{H}_A \otimes \mathscr{H}_B$.

The original formulation of the Naimark extension is the statement that any POVM can be extended to a projection measurement in a larger space, where the projectors may be of arbitrary rank, but the larger space need not come from the tensor product of the original space with an *ancilla* (helper) system. In our presentation the projectors in the larger space all have rank equal to the dimension of A, since they are of the form $\mathrm{id}_A \otimes |b_x\rangle\langle b_x|$. In the finite-dimensional case we are studying it is actually possible to find a Naimark extension of any POVM to a projective measurement consisting of *rank-one* elements, but we will not go into this here. For more details, see [6, §9-6] or [19, §3.1.4].

6.1.4 Post-measurement states

A POVM does not uniquely specify the post-measurement state, as there is some ambiguity in how the POVM is implemented, as follows. Given a POVM $\{\Lambda_j\}$, suppose we find a set of operators $\{M_{jk}\}$ such that

$$\sum_k M_{jk}^* M_{jk} = \Lambda_j. \tag{6.17}$$

The M_{jk} are sometimes called *measurement operators* (not to be confused with *POVM elements* Λ_j). Now suppose that we apply the unitary operator V_{ABC} defined by

$$V_{ABC}|\psi\rangle_A|b_0\rangle_B|b_0\rangle_C = \sum_{jk} M_{jk}|\psi\rangle_A|b_j\rangle_B|b_k\rangle_C, \tag{6.18}$$

and then measure B with projectors $P_j = |b_j\rangle\langle b_j|$. This gives the same probability distribution as the original POVM:

$${}_{BC}\langle b_0, b_0|_A\langle\psi|V_{ABC}^*(P_j)_B V_{ABC}|\psi\rangle_A|b_0, b_0\rangle_{BC} = \sum_k \langle\psi|M_{jk}^* M_{jk}|\psi\rangle = \langle\psi|\Lambda_j|\psi\rangle. \tag{6.19}$$

However, the output of the two implementations is different:

$$|\psi\rangle \xrightarrow{U} \rho_j = \frac{\sqrt{\Lambda_j}|\psi\rangle\langle\psi|\sqrt{\Lambda_j}}{p_j}, \tag{6.20}$$

$$|\psi\rangle \xrightarrow{V} \rho_j' = \frac{\sum_k M_{jk}|\psi\rangle\langle\psi|M_{jk}^*}{p_j}. \tag{6.21}$$

Unlike projection measurements, POVMs are not repeatable; that is, subsequent measurement with the same POVM does not always yield the same answer since the measurement operators M_{jk} are not necessarily mutually orthogonal.

6.2 Quantum operations

6.2.1 Completely positive maps (CPMs)

Let $\mathscr{H}_A$ and $\mathscr{H}_B$ 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 $\mathscr{E}$ from the set of states $\mathscr{S}(\mathscr{H}_A)$ on subsystem $\mathscr{H}_A$ at time t_0 to the set of states $\mathscr{S}(\mathscr{H}_B)$ on subsystem $\mathscr{H}_B$ at time t_1. This and the following sections are devoted to the study of this mapping.

Obviously, not every function $\mathscr{E}$ from $\mathscr{S}(\mathscr{H}_A)$ to $\mathscr{S}(\mathscr{H}_B)$ 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 $\mathscr{E}(\rho)$ is the mixture of $\mathscr{E}(\rho_0)$ and $\mathscr{E}(\rho_1)$. In other words, a physical mapping $\mathscr{E}$ needs to conserve the convex structure of the set of density operators, that is,

$$\mathscr{E}\big(p\rho_0 + (1-p)\rho_1\big) = p\mathscr{E}(\rho_0) + (1-p)\mathscr{E}(\rho_1), \tag{6.22}$$

for any $\rho_0, \rho_1 \in \mathscr{S}(\mathscr{H}_A)$ and any $p \in [0,1]$. If we do not require convexity in this manner, the trouble is that the transformation of a any particular ensemble member depends on the *other* members, even though only one element of the ensemble is actually realized. In other words, the dynamics of the true state would depend on the nonexsistent states!

For our considerations, it will be convenient to embed the mappings from $\mathscr{S}(\mathscr{H}_A)$ to $\mathscr{S}(\mathscr{H}_B)$ into the space of mappings from End($\mathscr{H}_A$) to End($\mathscr{H}_B$). The convexity requirement (6.22) then turns into the requirement that the mapping is linear. Since the mappings $\mathscr{E}$ are linear maps from operators to operators, they are often called *superoperators*.

Two criteria for any mapping $\mathscr{E}$ to map density operators to density operators are immediate:

1. $\rho' = \mathscr{E}(\rho) \geq 0$ for $\rho \geq 0$, and

2. $\mathrm{tr}[\mathscr{E}(\rho)] = 1$ for $\mathrm{tr}[\rho] = 1$.

Superoperators fulfilling the first condition are called *positive* and the second *trace-preserving*. An simple example of a map satisfying both conditions is the *identity map* on End($\mathscr{H}$), in the following denoted $\mathscr{I}$. A more interesting example is the transpose map $\mathscr{T}$, defined by

$$\mathscr{T} : S \mapsto S^T, \tag{6.23}$$

where S^T denotes the transpose with respect to some fixed basis $\{|b_k\rangle\}$. Clearly, $\mathscr{T}$ is trace-preserving, since the transpose does not affect the diagonal elements of a matrix. To see that $\mathscr{T}$ is positive, note that $\langle\phi|S^T|\phi\rangle = \langle\phi|\bar{S}^*\phi\rangle = \langle\bar{S}\phi|\phi\rangle = \overline{\langle\phi|\bar{S}\phi\rangle} = \langle\bar{\phi}|S|\bar{\phi}\rangle \geq 0$, from which we conclude $S^T \geq 0$. Here $|\bar{\phi}\rangle$ denotes the vector formed from $|\phi\rangle$ by taking the complex conjugate of a the components of $|\phi\rangle$ in the basis defining the transpose, $\{|b_k\rangle\}$.

Somewhat surprisingly, positivity by itself is not compatible with the possibility of purifying any mixed state. More concretely, positivity of two maps $\mathscr{E}$ and $\mathscr{F}$ does not necessarily imply positivity of the tensor map $\mathscr{E} \otimes \mathscr{F}$ defined by

$$(\mathscr{E} \otimes \mathscr{F})(S \otimes T) := \mathscr{E}(S) \otimes \mathscr{F}(T). \tag{6.24}$$

A simple example is provided by the superoperator $\mathscr{I}_A \otimes \mathscr{T}_B$ applied to $|\Phi\rangle\langle\Phi|_{AB}$, for $|\Phi\rangle_{AB}$ the canonical maximally-entangled state defined by (5.18). This state is a purification of the maximally-mixed state. The state resulting from the map is simply

$$\rho'_{AB} = \mathscr{I}_A \otimes \mathscr{T}_B(|\Phi\rangle\langle\Phi|_{AB}) = \frac{1}{d}\sum_{jk} |k\rangle\langle j|_A \otimes |j\rangle\langle k|_B. \tag{6.25}$$

Direct calculation reveals that $d\rho'_{AB}$ is the *swap operator*, i.e. $d\rho'_{AB}|\psi\rangle_A|\phi\rangle_B = |\phi\rangle_A|\psi\rangle_B$. But any antisymmetric combination of states, such as $|\psi\rangle_A|\phi\rangle_B - |\phi\rangle_A|\psi\rangle_B$, is an eigenstate of the swap operator with eigenvalue -1; hence $\rho'_{AB} \not\geq 0$.

In order to ensure compatibility with purification, we must demand that quantum operations be *completely positive*: positive on ρ and all its purifications. This translates into the formal requirement given as follows.

Definition 6.2.1. A linear map $\mathscr{E} \in \mathrm{Hom}(\mathrm{End}(\mathscr{H}_A), \mathrm{End}(\mathscr{H}_B))$ is said to be *completely positive* if for any Hilbert space $\mathscr{H}_R$, the map $\mathscr{E} \otimes \mathscr{I}_R$ is positive.

Clearly, $\mathscr{I}_A$ is completely positive, and it is easy to see that the partial trace tr_A is as well. We will use the abbreviation *CPM* to denote completely positive maps. Moreover, we denote by $\mathrm{TPCPM}(\mathscr{H}_A, \mathscr{H}_B)$ the set of trace-preserving completely positive maps from $\mathrm{End}(\mathscr{H}_A)$ to $\mathrm{End}(\mathscr{H}_B)$.

We have already encountered an example of a CPTP map in §6.1. Performing a measurement described by measurement operators $\{M_k\}$ with $\sum_k M_k^* M_k = \mathrm{id}$ results in the ensemble $\{p_k, \rho_k\}$ with $p_k = \mathrm{tr}[M_k \rho M_k^*]$ and $\rho_k = (M_k \rho M_k^*)/p_k$. Averaging over the outputs, i.e. forgetting which outcome occurred, leads to the average state

$$\mathscr{E}(\rho) = \sum_k M_k \rho M_k^*. \tag{6.26}$$

The map $\mathscr{E}$ must be a completely positive superoperator because, as we saw, it can be thought of as a unitary operator U_{AB} followed by tracing out system B, for U_{AB} defined by

$$U_{AB}|\psi\rangle_A|0\rangle_B = \sum_k M_k|\psi\rangle_A|k\rangle_B. \tag{6.27}$$

Both of these operations are CPTP maps, so $\mathscr{E}$ is, too.

In fact, all CPTP maps are of the form (6.26), often called the *operator-sum representation*. This statement is known as the *Kraus*[2] *representation theorem*, and we can easily prove it using the *Choi*[3] *isomorphism*. Since the Kraus form implies the existence of a unitary as in (6.27), this leads to the *Stinespring*[4] *dilation*. Historically, the Stinespring dilation was established first (as a generalization of the Naimark extension, it so happens), but we shall follow the route via the Choi isomorphism as it is simpler for finite-dimensional vector spaces.

6.2.2 The Choi isomorphism

The Choi isomorphism is a mapping that relates superoperators to operators and 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*.

[2] Karl Kraus, 1938 – 1988, German physicist.

[3] Man-Duen Choi, Canadian mathematician.

[4] William Forrest Stinespring, American mathematician.

Actually, we have already encountered the Choi isomorphism in (6.25); there ρ'_{AB} is the *Choi state* of the transpose map $\mathcal{T}_A$ (though it is not a valid state, as we saw). In general, the Choi isomorphism can be defined for any map $\mathcal{E}_{A\to B}$ which takes $\text{End}(\mathcal{H}_A)$ to $\text{End}(\mathcal{H}_B)$. In the following definition we make use of a "copy" of the state space $\mathcal{H}_A$, called $\mathcal{H}_{A'}$, and freely switch between the two in the subsequent discussion. Note that the Choi isomorphism depends on the choice of basis used to define the state $|\Phi\rangle_{A'A}$ of (5.18).

Definition 6.2.2. For $\mathcal{H}_A \simeq \mathcal{H}_{A'}$, the *Choi mapping (relative to the basis $\{|b_i\rangle\}_i$)* is the linear function C from $\text{Hom}(\text{End}(\mathcal{H}_A), \text{End}(\mathcal{H}_B))$ to $\text{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$ defined by

$$\mathsf{C} : \mathcal{E}_{A\to B} \mapsto (\mathcal{I}_A \otimes \mathcal{E}_{A'\to B})(|\Phi\rangle\langle\Phi|_{AA'}). \tag{6.28}$$

Lemma 6.2.3. *The Choi mapping C is an isomorphism. Its inverse C^{-1} takes any O_{AB} to the map $\mathcal{E}_{A\to B} = \mathsf{C}^{-1}(\rho_{AB})$ whose action is specified by*

$$\mathcal{E}_{A\to B} : S_A \mapsto d \cdot \text{tr}_A\Big((\mathcal{T}_A(S_A) \otimes \text{id}_B) O_{AB} \Big), \tag{6.29}$$

where $\mathcal{T}_A$ is the transpose map.

Proof. It suffices to verify that the mapping C^{-1} defined in the lemma is indeed an inverse of C. We first check that $\mathsf{C} \circ \mathsf{C}^{-1}$ is the identity on $\text{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$. For an arbitrary O_{AB}, using the definition of $|\Phi\rangle_{AB}$, we find

$$\begin{aligned}
\mathsf{C} \circ \mathsf{C}^{-1}(O_{AB}) &= \sum_{jk} |b_j\rangle\langle b_k|_A \cdot \text{tr}_{A'}[|b_k\rangle\langle b_j|_{A'} \otimes \text{id}_B O_{A'B}] && (6.30)\\
&= \sum_{jk} (|b_j\rangle_A \langle b_j|_{A'} \otimes \text{id}_B) O_{A'B} (|b_k\rangle_{A'} \langle b_k|_A \otimes \text{id}_B) && (6.31)\\
&= \Big(\sum_j |b_j\rangle_A \langle b_j|_{A'} \otimes \text{id}_B \Big) O_{A'B} \Big(\sum_k (|b_k\rangle_{A'} \langle b_k|_A \otimes \text{id}_B) \Big) && (6.32)\\
&= O_{AB}, && (6.33)
\end{aligned}$$

which establishes the claim.

It remains to show that C is injective. For this, recall that $S_A^T \otimes \text{id}_B |\Phi\rangle_{AA'} = \text{id}_A \otimes S_{A'} |\Phi\rangle_{AA'}$ for arbitrary $S_A \in \text{End}(\mathcal{H}_A)$ and that $\text{tr}_{A'}[|\Phi\rangle\langle\Phi|_{AA'}] = \text{id}_A$. For convenience, let us simply write $\Phi_{AA'}$ for $|\Phi\rangle\langle\Phi|_{AA'}$. Then we have

$$\begin{aligned}
\mathcal{E}_{A\to B}(S_A) &= d \cdot \mathcal{E}_{A\to B}(S_A \text{tr}_{A'}(\Phi_{AA'})) && (6.34)\\
&= d \cdot \text{tr}_{A'}\Big(\mathcal{E}_{A\to B} \otimes \mathcal{I}_{A'} (S_A \otimes \text{id}_{A'} \Phi_{AA'}) \Big) && (6.35)\\
&= d \cdot \text{tr}_{A'}\Big(\mathcal{E}_{A\to B} \otimes \mathcal{I}_{A'} (\text{id}_A \otimes S_{A'}^T \Phi_{AA'}) \Big) && (6.36)\\
&= d \cdot \text{tr}_{A'}\Big((\text{id}_A \otimes S_{A'}^T)(\mathcal{E}_{A\to B} \otimes \mathcal{I}_{A'}(\Phi_{AA'})) \Big). && (6.37)
\end{aligned}$$

Now assume $\mathsf{C}(\mathcal{E}) = 0$. Then, by definition, $(\mathcal{I}_{A'} \otimes \mathcal{E}_{A\to B})(\Phi_{AA'}) = 0$. By virtue of the above equality, this implies $\mathcal{E}(S_A) = 0$ for any S_A and, hence, $\mathcal{E} = 0$. Thus, $\mathsf{C}(\mathcal{E}) = 0$ implies $\mathcal{E} = 0$, meaning C is injective. □

The original interest in the Choi isomorphism was to give a means of determining whether a superoperator is completely positive. Since we have just shown that C is indeed an isomorphism, it follows that $\mathcal{E}_{A\to B}$ is completely positive only if the associated Choi state is positive. We shall return to the 'if' condition later.

In contemporary journal articles on quantum information theory it is common for the above isomorphism to be called the "Choi-Jamiołkowski[5]" isomorphism. However, this conflates two distinct isomorphisms. The *Jamiołkowski isomorphism* J is defined by

$$\mathsf{J}: \mathcal{E}_{A\to B} \mapsto (\mathcal{T}_A \otimes \mathcal{E}_{A'\to B})(|\Phi\rangle\langle\Phi|_{AA'}). \tag{6.38}$$

Despite the appearance of the transpose map, this isomorphism is actually basis independent, owing to the fact that $\mathcal{T}_A \otimes \mathcal{I}_{A'}(\Phi_{AA'})$ is the swap operator (up to normalization) no matter which basis is used to define $|\Phi\rangle_{AA'}$. In turn, this property follows from $U_A \otimes U_{A'}^T|\Phi\rangle_{AA'} = |\Phi\rangle_{AA'}$. The inverse J^{-1} takes any O_{AB} to the map $\mathcal{E} = \mathsf{J}^{-1}(O_{AB})$ whose action is specified by

$$\mathcal{E}: S_A \mapsto d \cdot \mathrm{tr}_A\Big((S_A \otimes \mathrm{id}_B)O_{AB}\Big). \tag{6.39}$$

6.2.3 The Kraus representation

Now we are ready to establish the Kraus representation theorem.

Theorem 6.2.4 (Kraus representation). *For any $\mathcal{E} \in \mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$ there exists a family $\{M_\ell\}_\ell$ of operators $M_\ell \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B)$ such that*

$$\mathcal{E}: S_A \mapsto \sum_\ell M_\ell S_A M_\ell^* \tag{6.40}$$

and $\sum_\ell M_\ell^* M_\ell = \mathrm{id}_A$. Conversely, any mapping $\mathcal{E}$ of the form (6.40) is contained in $\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$.

Proof. The converse follows from the discussion surrounding (6.27). (This will be the content of the Stinespring dilation discussed below.) For the forward direction, let $\rho_{AB} = \mathsf{C}(\mathcal{E}_{A\to B})$. Since $\rho_{AB} \geq 0$, it has eigendecomposition $\rho_{AB} = \sum_\ell \lambda_\ell |\lambda_\ell\rangle\langle\lambda_\ell|_{AB}$. Now define the map

$$M_\ell : |\phi\rangle \mapsto \sqrt{\lambda_\ell}\ {}_A\langle\bar{\phi}|\lambda_\ell\rangle_{AB}. \tag{6.41}$$

The map is linear, since

$$\begin{align} M_\ell\left(\sum_k \phi_k |b_k\rangle\right) &= M_\ell|\phi\rangle \tag{6.42}\\ &= \sqrt{\lambda_\ell}\ {}_A\langle\bar{\phi}|\lambda_\ell\rangle_{AB} \tag{6.43}\\ &= \sqrt{\lambda_\ell}\sum_k \phi_k\ {}_A\langle b_k|\lambda_\ell\rangle_{AB} \tag{6.44}\\ &= \sum_k \phi_k M_\ell|b_k\rangle. \tag{6.45}\end{align}$$

[5] Andrzej Jamiołkowski, Polish physicist.

Using the eigendecomposition of ρ_{AB} in (6.29) gives, for an arbitrary S_A,

$$\begin{aligned}
\mathcal{E}_{A\to B}(S_A) &= \mathrm{tr}_A[S_A^T \otimes \mathrm{id}_B \sum_\ell \lambda_\ell |\lambda_\ell\rangle\langle\lambda_\ell|_{AB}] && (6.46)\\
&= \sum_\ell \lambda_\ell \mathrm{tr}_A[\sum_{jk} \langle b_k|S|b_j\rangle\, |b_j\rangle\langle b_k|_A \otimes \mathrm{id}_B |\lambda_\ell\rangle\langle\lambda_\ell|_{AB}] && (6.47)\\
&= \sum_\ell \lambda_\ell \sum_{jk} \langle b_k|S|b_j\rangle\ {}_{AB}\langle\lambda_\ell|b_j\rangle_A\ {}_A\langle b_k|\lambda_\ell\rangle_{AB} && (6.48)\\
&= \sum_{jk\ell} \langle b_k|S|b_j\rangle\ M_\ell |b_k\rangle\langle b_j| M_\ell^* && (6.49)\\
&= \sum_\ell M_\ell S_A M_\ell^* && (6.50)
\end{aligned}$$

Since $\mathcal{E}_{A\to B}$ is trace preserving,

$$\mathrm{tr}[\mathcal{E}_{A\to B}(\rho)] = \sum_\ell \mathrm{tr}[M_\ell \rho M_\ell^*] = \mathrm{tr}[\sum_\ell M_\ell^* M_\ell \rho] \tag{6.51}$$

holds for arbitrary ρ. This implies that $\sum_\ell M_\ell^* M_\ell = \mathrm{id}$, completing the proof. □

There are two important corollaries to the Kraus representation theorem, both following from the form of the Choi state. First, since $\rho_{AB} = \mathsf{C}(\mathcal{E}_{A\to B}) \in \mathrm{Hom}(\mathcal{H}_A \otimes \mathcal{H}_B)$, it has at most $d_A d_B$ eigenvectors. Therefore, the map $\mathcal{E}_{A\to B}$ always has a Kraus representation with at most $d_A d_B$ *Kraus operators* M_ℓ. Secondly, in the construction of the Kraus operators we are free to use any decomposition of the Choi state into pure states, not only the eigendecomposition. The result would be another set of Kraus operators $\{M'_\ell\}$, generally having more elements. But, by the unitary relation of all possible pure state decompositions from §5.1.3, a similar unitary relation holds among all possible sets of Kraus operators as well. In particular, if $\sqrt{\lambda'_\ell}|\lambda'_\ell\rangle = \sum_m U_{\ell m}\sqrt{\lambda_m}|\lambda_m\rangle$ for $U_{\ell m}$ a unitary matrix, then

$$\sqrt{\lambda'_\ell}\langle\bar{\phi}|\lambda'_\ell\rangle = \sum_m \sqrt{\lambda_m} U_{\ell m}\langle\bar{\phi}|\lambda_\ell\rangle \tag{6.52}$$

and so $M'_\ell = \sum_m U_{\ell m} M_m$.

A careful reading of the proof reveals that we really only used complete positivity to assert that the Choi state is Hermitian, and therefore has a spectral decomposition. The positivity of the eigenvalues is not used in the proof. Since completely positive maps are also Hermiticity-preserving maps, we could have used the Jamiołkowski isomorphism instead of the Choi isomorphism. This is slightly more elegant mathematically, since the former does not depend on the basis choice. The construction proceeds almost exactly as before, only now the Kraus operators are defined by

$$M_\ell|\phi\rangle = \sqrt{\eta_\ell}\ {}_{AB}\langle\eta_\ell|\phi\rangle_A, \tag{6.53}$$

for $|\eta_\ell\rangle$ and η_ℓ the eigenvectors and eigenvalues of $\mathsf{J}(\mathcal{E}_{A\to B})$. Defined this way, the Kraus operators are manifestly linear. The proof using the Choi isomorphism, however, lets us recycle the result on ambiguity in the decomposition of density operators to infer the structure of sets of Kraus operators corresponding to a fixed CPTP map.

6.2.4 Stinespring dilation

The Stinespring dilation now follows immediately from the Kraus representation theorem.

Theorem 6.2.5 (Stinespring dilation)**.** *Let $\mathcal{E}_{A\to B}$ be a CPTP map from* $\mathrm{End}(\mathcal{H}_A)$ *to* $\mathrm{End}(\mathcal{H}_B)$*. Then there exists an isometry $U_{A\to BR} \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B \otimes \mathcal{H}_R)$ for some Hilbert space $\mathcal{H}_R$ such that*

$$\mathcal{E}_{A\to B} : S_A \mapsto \mathrm{tr}_R(U_{A\to BR} S_A U^*_{A\to BR}). \tag{6.54}$$

The dimension of $\mathcal{H}_R$ can be taken to be at most $d_A d_B$.

Proof. One possible isometry $U_{A\to BR}$ is defined by the action

$$U_{A\to BR}|\psi\rangle_A|0\rangle_R = \sum_k M_k|\psi\rangle_A|k\rangle_R, \tag{6.55}$$

just as in (6.27). That this is an isometry was already established in (6.14), but we repeat the calculation here for completeness:

$$\begin{aligned}\langle\phi'|\psi'\rangle &= \sum_{\ell,\ell'}(\langle\phi|M^*_{\ell'} \otimes \langle b_{\ell'}|)(M_\ell|\psi\rangle \otimes |b_\ell\rangle) && (6.56)\\ &= \sum_\ell \langle\phi|M^*_\ell M_\ell|\psi\rangle && (6.57)\\ &= \langle\phi|\psi\rangle. && (6.58)\end{aligned}$$

Since at most $d_A d_B$ Kraus operators are needed, $\dim(\mathcal{H}_R)$ need not be larger than this value. □

The Stinespring dilation shows that general quantum operations (CPTP maps) can be regarded as unitary operations on a larger system: Any CPTP map $\mathcal{E}_{A\to A}$ can be dilated to an isometry $U_{A\to AR}$, which can be extended to a unitary on AR. Thus, we have successfully altered the postulates to describe *open systems*, systems in contact with their surrounding environment, by essentially requiring that the original postulates be satisfied when including the environmental degrees of freedom. We have not been so explicit about this requirement in the preceding discussion, but it is implicit whenever we make use of the purification, as the purification gives the most general quantum description of a system and its environment. Indeed, this is a marked departure from the situation classically, since purification means that in the quantum case the description of the system itself contains the description of the environment.

Using the Stinespring dilation and Kraus representation we can return to the issue of using the Choi state to determine if a superoperator is completely positive, raised in §6.2.2. We have the following

Lemma 6.2.6. *A map $\mathcal{E}_{A\to B}$ is completely positive if and only if* $\mathsf{C}(\mathcal{E}_{A\to B}) \geq 0$.

Proof. The necessity of the condition follows immediately from the definition of C, as already discussed. To establish sufficiency, suppose the Choi state is positive. Then $\mathcal{E}$ has a Kraus representation and hence a Stinespring dilation $U_{A\to BR}$. Therefore, for any $\mathcal{H}_{R'}$ we have

$$\mathcal{E}_{A\to B} \otimes \mathcal{I}_{R'}(\rho_{AR'}) = \mathrm{tr}_R[(U_{A\to BR} \otimes \mathrm{id}_{R'})\rho_{AR'}(U_{A\to BR} \otimes \mathrm{id}_{R'})^*], \tag{6.59}$$

which is completely positive since both unitary action are the partial trace are. □

With the Kraus representation theorem in hand, we can also refine the Choi isomorphism a little bit, to an isomorphism between completely positive superoperators and states of a certain form.

Lemma 6.2.7. *The Choi mapping* C *is an isomorphism between completely positive superoperators* $\mathcal{E}_{A\to B} \in \mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$ *and positive operators* $\rho_{AB} \in \mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$ *with the additional property* $\mathrm{tr}_B[\rho_{AB}] = \frac{1}{d}\mathrm{id}_A$.

Proof. The Choi mapping always outputs a state of the given form. Conversely, given a state of that form, the Kraus representation theorem ensures that the corresponding map $\mathsf{C}^{-1}(\rho_{AB})$ is completely positive. □

6.3 Further Reading

The Choi isomorphism was introduced in [20], Stinespring's dilation in [21], the Kraus form in [22]. The Jamiołkowski isomorphism, which might be more properly attributed to de Pillis, is found in [23].

7

The Decidedly Non-Classical Quantum World

In the preceding chapters we have seen that the formalism of quantum theory differs considerably from classical theory. But how does this translate into physics? How does quantum theory differ physically from classical theory? In this chapter we will examine three main differences: complementarity, uncertainty relations, and Bell inequalities.

7.1 Complementarity

Complementarity of the particle and wave nature of light in the double slit experiment is one of the most well-known examples. Feynman[1] starts off his treatment of quantum mechanics in his famous lectures with a treatment of the double-slit experiment, stating

> In this chapter we shall tackle immediately the basic element of the mysterious behavior in its most strange form. We choose to examine a phenomenon which is impossible, *absolutely* impossible, to explain in any classical way, and which has in it the heart of quantum mechanics. In reality, it contains the *only* mystery. We cannot make the mystery go away by "explaining" how it works. We will just *tell* you how it works. In telling you how it works we will have told you about the basic peculiarities of all quantum mechanics.[24]

7.1.1 Complementarity in the Mach-Zehnder interferometer

In our formalism, we can see that the mystery of the double-slit experiment is intimately related to entanglement. Let's simplify the physics and instead consider a Mach[2]-Zehnder[3] interferometer using polarizing beamsplitters (PBS), depicted in Figure 7.1.

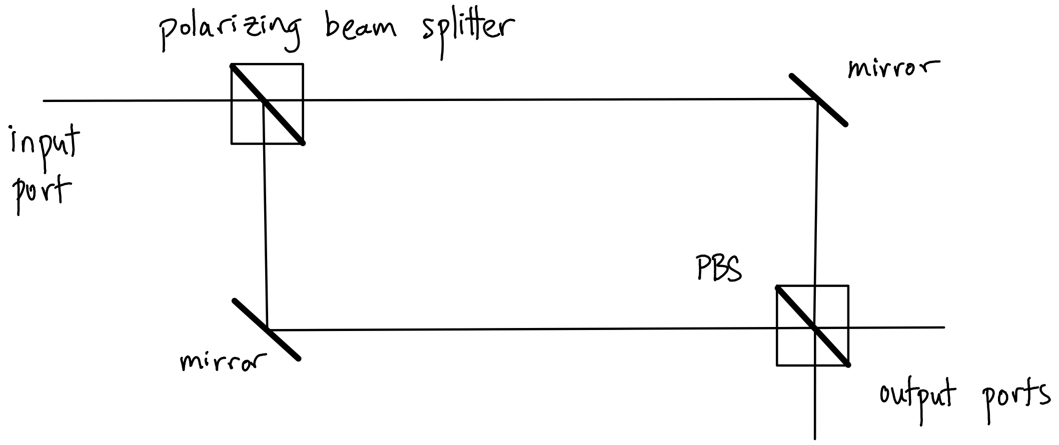


Figure 7.1: A Mach-Zehnder interferometer

[1] Richard Phillips Feynman, 1918 – 1988, American physicist.

[2] Ludwig Mach, 1868 – 1951, Austrian inventor; son of physicist Ernst Mach.

[3] Ludwig Louis Albert Zehnder, 1854 – 1949, Swiss physicist.

Imagine a single photon entering the interferometer. Its polarization could be horizontal, vertical, or any linear combination of these, and its quantum state space is given by $\mathscr{H}_P = \mathbb{C}^2$ with a basis $|0\rangle_P$ for horizontal and $|1\rangle_P$ for vertical polarization. As it travels through the interferometer it can propagate in two spatial modes, call them 'top' and 'bottom' in accord with Figure 7.1. These two modes also form a two-dimensional state space $\mathscr{H}_M$ with basis state $|0\rangle_M$ for the top and $|1\rangle_M$ for the bottom modes.

The beamsplitters separate horizontal from vertical polarization, meaning we can take the action of the polarizing beamsplitter to be

$$U_{\text{PBS}} = \sum_{z=0}^{1} |z\rangle\langle z|_P \otimes |z\rangle_M. \tag{7.1}$$

This equation defines an isometry, not a unitary, since we are ignoring the spatial mode of the input (i.e. we implicitly assume it is in $|0\rangle_M$). Also, we have ignored phases associated with transmission as opposed to reflection from the beamsplitter.

Suppose the photon is initially polarized at +45°, so that its quantum state is

$$|\psi_0\rangle_P = |+\rangle_P = \tfrac{1}{\sqrt{2}}(|0\rangle_P + |1\rangle_P) \tag{7.2}$$

After the first beamsplitter the state becomes

$$|\psi_1\rangle_{PM} = \tfrac{1}{\sqrt{2}}(|0\rangle_P|0\rangle_M + |1\rangle_P|1\rangle_M). \tag{7.3}$$

The path and polarization degrees of freedom have become (maximally) entangled.

The first PBS enables us to measure the polarization of the photon by measuring which arm of the interferometer it is in. In other words, the isometry U_{PBS} gives the von Neumann description of a polarization measurement. However, due to the entanglement between the mode and polarization degrees of freedom, the *coherence* of the polarization state has been lost. The horizontal and vertical states no longer interfere to produce a definite state of polarization. Instead, the polarization state is a mixture of these two states,

$$\rho_P = \text{tr}_M[|\psi_1\rangle\langle\psi_1|_{PM}] = \tfrac{1}{2}(|0\rangle\langle 0|_P + |1\rangle\langle 1|_P). \tag{7.4}$$

Coherence, the possibility of interference between two states of a certain degree of freedom, is only possible if all other degrees of freedom are completely uncorrelated.

But all is not lost in the interferometer, however, as the polarization coherence can be restored simply by passing the photon through the second PBS. Since the output mode is not relevant, the second beamsplitter is described by U^*_{PBS}. That is to say, in the setup the mode and polarization degrees of freedom are completely correlated for every possible input, so we do not need to specify what the second PBS does to vertical polarization propagating the top mode. The two beamsplitters together produce the state

$$|\psi_2\rangle_P = U^*_{\text{PBS}} U_{\text{PBS}} |\psi_0\rangle_P = |\psi_0\rangle_P, \tag{7.5}$$

since $U^*_{\text{PBS}} U_{\text{PBS}} = \text{id}$ as U_{PBS} is an isometry. This effect of "undoing" the measurement is known as the *quantum eraser*, since the which-way information has been erased.

To make the connection with the usual presentation of the double-slit experiment, imagine that we could check which arm the photon is in without destroying it, which is the usual sort of photodetection measurement. For instance, the photon might pass through an optical cavity, altering the state of an atom present in the cavity. The state of the atom can then be measured to determine if a photon passed through the cavity or not. Abstracting away the details, this indirect measurement can be described by the isometry

$$U_{\text{arm}} = \sum_z |z\rangle\langle z|_M \otimes |\varphi_z\rangle_A, \tag{7.6}$$

where the $|\varphi_z\rangle_A$ are the two states of the ancilla produced in the measurement process.

Now the interferometer produces the state

$$|\psi_2'\rangle_{PA} = U^*_{\text{PBS}} U_{\text{arm}} U_{\text{PBS}} |\psi_0\rangle = \tfrac{1}{\sqrt{2}} \sum_z |z\rangle_P |\varphi_z\rangle^A. \tag{7.7}$$

The first PBS is analogous to light being forced through the two slits in the double slit experiment, while the second PBS mimics the result of their interference as the propagate. Measuring the output in the ±45° polarization basis, corresponding to the states $|\pm\rangle$, is analogous to the screen or film.

When the ancilla states are identical, $|\varphi_0\rangle = |\varphi_1\rangle$, then no information about the path has been acquired. The coherence between the two paths (and therefore between the polarizations) is maintained, the paths interfere at the second PBS, and measurement of the output always corresponds to $|+\rangle$. This is the wave nature of the photon. On the other hand, when the ancilla states are orthogonal and we have learned which path the photon took, the paths can no longer interfere. This is the particle nature of the photon. Coherence is lost and measurement after the second PBS just produces a random outcome, since the state is $\rho'_P = \frac{1}{2}\text{id}_P$.

Of course, none of this would occur in a classical world, where light is either a wave or a particle. For a wave, there is no meaning to which path it takes, since it can take both. For a particle, there is no meaning to coherence between paths, since it takes one or the other. Moreover, the particle or wave nature in question is not at the whims of the experimenter, as in the quantum case.

7.1.2 A quantitative statement of complementarity

We can quantify the complementarity of the wave and particle nature of the photon in the above setup. The particle nature corresponds to which path the photon took, and we may quantify this by how well we can predict a hypothetical measurement of the mode (or, equivalently, the polarization) by measuring the states $|\varphi_z\rangle_A$ in the ancilla system. As we saw in the exercises, the probability of correctly guessing the outcome of the hypothetical measurement is quantified by the trace distance. Specifially, the guessing probability is given by $p_{\text{guess}} = \frac{1}{2}(1+\delta(\varphi_0,\varphi_1))$. This motivates the definition of the *distinguishability* of the two paths by $D = \delta(\varphi_0,\varphi_1)$. Its value ranges from zero (complete indistinguishability) to one (complete distinguishability).

On the other hand, interference at the output of the interferometer corresponds to the wave nature. Specifically, if the measurement of the interferometer output in the basis $|\pm\rangle$ is more likely to produce $|+\rangle$ than $|-\rangle$, this can be taken as an indication of the wave nature of the photon. Calling the measurement result X, the above motivates the definition of the *visibility* as $V = |P_X(0) - P_X(1)|$. Again, the value ranges from zero to one. The terminology comes from the visibility of fringe patterns in the double slit experiment. Our definition corresponds to the difference between intensities at the maxima and minima in that case.

With the above definitions, we can then show the following trade-off between the distinguishability and the visibility.

Theorem 7.1.1 (Englert [25]). *For any pure state entering the interferometer above, $D^2 + V^2 \leq 1$.*

Proof. First let us calculate the reduced state of the polarization degree of freedom. Assuming an input state of the form $|\psi\rangle_P = \sum_z \psi_z |z\rangle_P$ for $\psi_z \in \mathbb{C}$ such that $|\psi_0|^2 + |\psi_1|^2 = 1$, the total output state is just

$$|\psi'\rangle_{PM} = \sum_z \psi_z |z\rangle_P |\varphi_z\rangle_A. \tag{7.8}$$

The reduced state of P is then

$$\rho_P = \text{tr}_M[|\psi'\rangle\langle\psi'|_{PM}] = \sum_{zz'} \psi_z \psi^*_{z'} |z\rangle\langle z'|_P \text{tr}[|\varphi_z\rangle\langle\varphi_{z'}|] \tag{7.9}$$

$$= \sum_{zz'} \psi_z \psi^*_{z'} \langle\varphi_{z'}|\varphi_z\rangle |z\rangle\langle z'|_P. \tag{7.10}$$

Next we can compute the visibility as follows.

$$V = |\text{tr}[|+\rangle\langle+|_P \rho_P] - \text{tr}[|-\rangle\langle-|_P \rho_P]| = |\text{tr}[(\sigma_x)_P \rho_P]| \tag{7.11}$$

$$= \Big|\sum_{zz'} \psi_z \psi^*_{z'} \langle\varphi_{z'}|\varphi_z\rangle \text{tr}[\sigma_x |z\rangle\langle z'|]\Big| \tag{7.12}$$

$$= \Big|\sum_{zz'} \psi_z \psi^*_{z'} \langle\varphi_{z'}|\varphi_z\rangle \langle z'|z+1\rangle\Big| \tag{7.13}$$

$$= \Big|\sum_{z} \psi_z \psi^*_{z+1} \langle\varphi_{z+1}|\varphi_z\rangle\Big| \tag{7.14}$$

$$= |\psi_0\psi_1^* \langle\varphi_1|\varphi_0\rangle + \psi_1\psi_0^*\langle\varphi_0|\varphi_1\rangle| \tag{7.15}$$

$$\le |\psi_0\psi_1^* \langle\varphi_1|\varphi_0\rangle| + |\psi_1\psi_0^*\langle\varphi_0|\varphi_1\rangle| \tag{7.16}$$

$$= 2|\psi_0\psi_1^*| \cdot |\langle\varphi_0|\varphi_1\rangle| \tag{7.17}$$

$$\le |\langle\varphi_0|\varphi_1\rangle|. \tag{7.18}$$

The first inequality is the triangle inequality for complex numbers, while the second is the fact that $|\psi_0\psi_1^*| \le \frac{1}{2}$. This holds because we can express the two coefficients as $\psi_0 = \sqrt{p}e^{\theta_0}$ and $\psi_0 = \sqrt{1-p}e^{\theta_1}$ for $0 \le p \le 1$ and two arbitrary angles θ_0 and θ_1. Thus $|\psi_0\psi_1^*| = |\sqrt{p}\sqrt{1-p}| \le \frac{1}{2}$.

Finally, since the two states in the distinguishability are pure, $D = \sqrt{1 - |\langle\varphi_0|\varphi_1\rangle|^2}$, and therefore $|\langle\varphi_0|\varphi_1\rangle|^2 = 1 - D^2$. Using the inequality for V completes the proof. □

7.2 Uncertainty relations

Complementarity is intimately related to uncertainty relations in quantum mechanics. In fact, we can see Theorem 7.1.1 as an uncertainty relation itself, though not of the familiar Heisenberg[4] form $\Delta x \Delta p \ge \frac{\hbar}{2}$. There are actually two interesting interpretations.

In the first, the uncertainty in question pertains to the results of two possible polarization measurements on the photon. One measurement is the actual measurement performed on the output, of polarization at the ±45°, corresponding to basis states $|\pm\rangle$. The second is a hypothetical measurement, *on the input*, of horizontal versus vertical polarization.

A simple calculation shows that $V = 2\delta(P_X, U_X)$, where U_X is the uniform distribution. Let us call the trace distance in this case $q_{\text{flat}}(X) = \delta(P_X, U_X)$, the "flatness" of the distribution. If we call

[4] Werner Karl Heisenberg, 1901 – 1976, German physicist.

Z the outcome of the hypothetical measurement, the distinguishability was defined by starting from the probability $p_{\text{guess}}(Z|A)$ of correctly guessing the value of Z by making use (i.e. measuring) the ancilla system. In particular, $D = 2p_{\text{guess}}(Z|A) - 1$. Inserting this into Theorem 7.1.1, we find

$$(2p_{\text{guess}}(Z|A) - 1)^2 + 4q_{\text{flat}}(X)^2 \leq 1 \tag{7.19}$$

$$\Rightarrow \quad p_{\text{guess}}(Z|A)^2 + q_{\text{flat}}(X)^2 \leq p_{\text{guess}}(Z|A). \tag{7.20}$$

If we weaken the bound a bit by using $p_{\text{guess}}(Z|A) \leq 1$, we end up with the following elegant form:

$$p_{\text{guess}}(Z|A)^2 + q_{\text{flat}}(X)^2 \leq 1. \tag{7.21}$$

Thus, if the probability of guessing the outcome of the hypothetical horizontal/vertical measurement is very high, then the result of the ±45° measurement will be nearly uniform. On the other hand, if the result of the latter measurement is not close to uniform, then the guessing probability must be correspondingly low. This embodies a kind of information gain versus disturbance tradeoff: the more information can be gained about the horizontal/vertical state of the input, the more the ±45° state is altered.

The second interpretation of Theorem 7.1.1 as an uncertainty principle comes from mapping the two actual measurements made after the photon traverses the interferometer to two measurements that could be made on the input state. As we have seen, the whole interferometric setup is described by an isometry which maps the input degree of freedom P to P and A, and subsequently measurements are made on P and A separately. But we can just as well reverse the procedure and use the adjoint of the isometry to transform the measurements on the output spaces into measurements on the input. Then Theorem 7.1.1 may be regarded as an uncertainty principle for these two measurements.

Let us determine what these two measurements are, precisely. First, we need to name the measurements on A and P. Both are projective measurements, and in the latter case the two projectors are $\tilde{P}_x = |\tilde{x}\rangle\langle\tilde{x}|$ for $x = 0, 1$, where $|\tilde{x}\rangle = \frac{1}{\sqrt{2}}(|0\rangle + (-1)^x|1\rangle)$. To simplify the expression of the former measurement, let us first assume that the states $|\varphi_z\rangle$ take the form

$$|\varphi_0\rangle = \cos\theta|0\rangle + \sin\theta|1\rangle \tag{7.22}$$

$$|\varphi_1\rangle = \sin\theta|0\rangle + \cos\theta|1\rangle. \tag{7.23}$$

for some $\theta \in \mathbb{R}$. This choice can be made without loss of generality, by picking the basis states of the ancilla system appropriately. Again from the exercises, we know that the optimal measurement to distinguish these two states is just $P_z = |z\rangle\langle z|$ for $z = 0, 1$.

Now let $W = U^*_{\text{PBS}} U_{\text{arm}} U_{\text{PBS}}$. The measurements on the input space that we are interested in are given by

$$\tilde{\Gamma}_x = W^*(\tilde{P}_x \otimes \text{id})W \tag{7.24}$$

$$\Gamma_z = W^*(\text{id} \otimes P_z)W. \tag{7.25}$$

Note that the original measurements commute with one another, since they act on different degrees of freedom. But this will no longer be true after applying the adjoint of the isometry.

As will be shown in the exercises,

$$\tilde{\Gamma}_x = \tfrac{1}{2}\left(\text{id} + (-1)^x \sin 2\theta\, \sigma_x\right), \tag{7.26}$$

$$\Gamma_z = \tfrac{1}{2}\left(\text{id} + (-1)^z \cos 2\theta\, \sigma_z\right). \tag{7.27}$$

Both Γ and $\tilde{\Gamma}$ are "noisy" versions of measurements in the σ_z or σ_x bases. Indeed, if $\theta = 0$, then Γ is just a measurement in the σ_z basis, while $\tilde{\Gamma}$ is a trivial measurement with elements $\frac{1}{2}\mathrm{id}$, and vice versa for $\theta = \pi/4$. The identity contributions to the POVM elements serve to make the two outcomes more equally-likely, i.e. they reduce the information the POVM collects about the corresponding basis.

Now let X' be the outcome of the $\tilde{\Gamma}$ measurement and Z' the outcome of the Γ measurement. As before, we can define $q_{\text{flat}}(X') = \delta(P_{X'}, U_{X'})$ from the visibility. The interpretation of the distinguishability is less straightforward, since there is no longer any side information. Nonetheless, the guessing probability is still well-defined; now we must simply guess without access to any extra information A. Then $p_{\text{guess}}(Z') = \max\{P_{Z'}(0), P_{Z'}(1)\}$. The above relation continues to hold, so that

$$q_{\text{flat}}(X')^2 + p_{\text{guess}}(Z')^2 \leq 1. \tag{7.28}$$

7.2.1 Joint measurement of noncommuting observables

We could also view the measurements of the interferometer output as one combined POVM and ask what this corresponds to on the input system by applying W^*. The result is in some sense a joint measurement of the noncommuting observables σ_x and σ_z. It has elements $\Lambda_{x,z}$ specified by

$$\Lambda_{x,z} = W^*(|\tilde{x}\rangle\langle\tilde{x}| \otimes |z\rangle\langle z|)W. \tag{7.29}$$

The justification for calling it a joint measurement of σ_x and σ_z comes from the calculations we performed above, which state that

$$\Gamma_z = \sum_x \Lambda_{x,z} \qquad \text{and} \qquad \tilde{\Gamma}_x = \sum_z \Lambda_{x,z}, \tag{7.30}$$

and the fact that the $\tilde{\Gamma}$ and Γ measurements are noisy versions of σ_x and σ_z.

Interestingly, if we ask for general conditions on when there exists a joint measurement of two noncommuting observables in the sense above (i.e. the marginals are noisy versions of the original measurements), then we are lead back to the complementarity tradeoff of Theorem 7.1.1! For more details, see [26].

7.3 The EPR paradox

Complementarity and uncertainty relations assert that physical systems cannot simultaneously display two complementary properties or at least that two such properties cannot both be known to an observer simultaneously. This raises the question: Do systems have these complementary properties and they just refuse to tell us, or do they not have these properties in the first place? Put differently, the question is whether complementarity just results from some kind of inevitable disturbance to a system upon measurement or whether complementary properties somehow do not exist in the first place, and hence cannot be simultaneously known.

Is there any way to tell which of these two options is correct? Before we attempt to answer this question, it is worth specifying more precisely we mean by "real properties" in the first place. A very concise notion is given by Einstein, Podolsky, and Rosen (EPR) in their celebrated 1935 paper in the Physical Review:

> If, without in any way disturbing a system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity.[27]

Now, if disturbance to elements of reality is caused by measurement, then one thing measurement ought *not* do is disturb such elements of reality in systems far from where the measurement takes place. This is the principle of locality, one of the basic principles of modern physics.

Entangled states have a peculiar relation to locality, as noticed by EPR, Einstein in particular. EPR considered what we would regard as two purifications of a given density operator:

$$|\Psi\rangle_{AB} = \sum_k |\psi_k\rangle_A \otimes |u_k\rangle_B = \sum_s |\varphi_s\rangle_A \otimes |v_s\rangle_B, \tag{7.31}$$

where the $|\psi_k\rangle$ and $|\varphi_s\rangle$ are arbitrary pure states, while the $|u_k\rangle$ and $|v_s\rangle$ are unnormalized, but mutually orthogonal states. As we have seen, measurement of system B in the $|u_k\rangle$ basis will result in the post-measurement state $|\psi_k\rangle_A$ in A with probability $\langle u_k|u_k\rangle$. Similarly, measurement of system B in the $|v_s\rangle$ basis will result in the post-measurement state $|\varphi_s\rangle_A$ in A with probability $\langle v_s|v_s\rangle$. Indeed, due to the unitary freedom in the purification, there are many possible post-measurement states that can be prepared in system B by action on system A. Schrödinger termed this sort of phenomenon *steering* and noted the conflict with locality by saying

> It is rather discomforting that the theory should allow a system to be steered or piloted into one or the other type of state at the experimenter's mercy in spite of his having no access to it.[28]

Note that steering does not imply the possibility of superluminal signalling. Although it is true that if Bob measures system B, his description of Alice's system A changes upon obtaining the outcome of the measurement. But Alice's description has not changed since she does not know the measurement outcome. For her the state of A was and is $\rho_A = \text{tr}_B[|\Psi\rangle\langle\Psi|_{AB}]$. Since her state contains no information about Bob's measurement choice or outcome, no communication of any kind is possible, superluminal or otherwise.

Returning to the EPR argument, observe that the various different post-measurement states could correspond to eigenvectors of noncommuting observables on B. But then the values taken by these observables should therefore *all* be elements of reality, at least if the action taken at A does not influence the elements of reality at B. But, as we know from the Robertson-type uncertainty relations, noncommuting observables cannot simultaneously take on well-defined values. This is the EPR paradox.

The conclusion of the EPR paper is that the quantum-mechanical description of systems in terms of state vectors is *incomplete*, that is, there are elements of reality associated with noncommuting observables, the uncertainty principle notwithstanding, but that these are not encapsulated in the state vector $|\psi\rangle$. The state vector should contain all elements of reality, but does not.

Einstein stated a slightly different conclusion in a letter to Schrödinger, eschewing the argument regarding elements of reality and taking aim directly at the state vector as a description of reality:

> Now what is essential is exclusively that $[|\psi_k\rangle_B]$ and $[|\varphi_s\rangle_B]$ are in general different from one another. I assert that this difference is incompatible with the hypothesis that the description is correlated one-to-one with the physical reality (the real state). After the collision [which in the EPR model produces the entangled state], the real state of (AB) consists precisely of the real state of A and the real state of B, which two states have

> nothing to do with one another. *The real state of B thus cannot depend upon the kind of measurement I carry out on A.* ("Separation hypothesis" from above.) But then for the same state of B there are two (in general arbitrarily many) equally justified $[|\psi\rangle_B]$, which contradicts the hypothesis of a one-to-one or complete description of the real states.[5]

Clearly what Einstein has in mind here is that each system has its own elements of reality, or real state, and these should obey locality. We call this sort of description a *locally realistic*. If a locally realistic description of quantum phenomena is possible, it is not found in the use of state vectors.

An important aspect of the EPR argument to note is their reasoning from *counterfactuals*, that is measurements that were not performed. They themselves acknowledge this, noting that

> One could object to this conclusion on the grounds that our criterion of reality is not sufficiently restrictive. Indeed, one would not arrive at our conclusion if one insisted that two or more physical quantities can be regarded as simultaneous elements of reality only when they can be simultaneously measured or predicted. On this point of view, since either one or the other, but not both simultaneously, of the quantities P and Q can be predicted, they are not simultaneously real. This makes the reality of P and Q depend upon the process of measurement carried out on the first system, which does not disturb the second system in any way. No reasonable definition of reality could be expected to permit this.

7.4 Bell inequalities

The EPR argument perhaps raises our hopes that complementarity is due to the inevitable disturbance of measurement, by "revealing" the existence of elements of reality obscured by the uncertainty relation and complementarity. But the elements of reality must be partly in the form of *hidden variables* not contained in the state vector description of a system. Is such a description possible? Is a locally realistic formulation of quantum mechanics possible, one possibly making use of hidden variables? By showing that local realism constrains the possible correlations between measurements made on two separated systems, Bell demonstrated that such a description is *not* possible. Thus, we face two unpalatable alternatives. Either the source of complementarity should be attributed to a lack of existence of local elements of reality, or these independent elements of reality must be nonlocal.

7.4.1 The CHSH inequality

A simplified version of Bell's argument was put forth by Clauser, Horne, Shimony, and Holt, and is known as the CHSH inequality. It involves two systems, upon which the experimenters Alice and Bob can each make one of two possible measurements. Every measurement has two possible outcomes, which we will label ±1. Abstractly, this defines four observables a_0, a_1, b_0 and b_1. According to local realism, deterministic values ±1 can be assigned to all observables, even though it might be

[5] "Wesentlich ist nun ausschliesslich, dass ψ_B und $\psi_{\underline{B}}$ überhaupt voneinander verschieden sind. Ich behaupte, dass diese Verschiedenheit mit der Hypothese, dass die ψ-Beschreibung ein-eindeutig der physikalischen Wirklichkeit (dem wirklichen Zustande) zugeordnet sei, unvereinbar ist. Nach dem Zusammenstoss besteht der wirkliche Zustand von (AB) nämlich aus dem wirklichen Zustand von A und dem wirklichen Zustand von B, welche beiden Zustände nichts miteinander zu schaffen haben. Der wirkliche Zustand von B kann nun nicht davon abhängen, was für eine Messung ich an A vornehme. ('Trennungshypothese' von oben.) Dann aber gibt es zu demselben Zustände von B zwei (überhaupt bel. viele) gleichberechtigte ψ, was der Hypothese einer ein-eindeutigen bezw. vollständigen Beschreibung der wirklichen Zustände widerspricht."[27]

that a_0 and a_1 (and b_0 and b_1) cannot be simultaneously measured (an instance of the reasoning from counterfactuals described above). From this, it immediately follows that

$$C = (a_0 + a_1)b_0 + (a_0 - a_1)b_1 = \pm 2. \tag{7.32}$$

Now imagine that the values of these observables are not directly given in a model of the situation, but require additional hidden variables to pin them down exactly. Calling the hidden variable λ and its distribution $P_{\mathrm{HV}}(\lambda)$, we can express the probability for the observables to take on the definite values a_0, a_1, b_0, and b_1 as

$$P(a_0 = \mathsf{a}_0, a_1 = \mathsf{a}_1, b_0 = \mathsf{b}_0, b_1 = \mathsf{b}_1 | \lambda) P_{\mathrm{HV}}(\lambda). \tag{7.33}$$

But since (7.32) is an equality, averaging over λ like so will only lead to

$$|\langle C \rangle| = |\langle a_0 b_0 \rangle + \langle a_1 b_0 \rangle + \langle a_0 b_1 \rangle - \langle a_1 b_1 \rangle| \leq 2. \tag{7.34}$$

This is the *CHSH inequality*, an instance of a generic *Bell inequality*.

The CHSH inequality can be violated in quantum mechanics, by making use of entangled states. Suppose the bipartite state of two qubit systems A and B is the state $|\Psi\rangle = \frac{1}{\sqrt{2}}(|01\rangle_{AB} - |10\rangle_{AB})$ and let the observables be associated with Bloch vectors $\hat{a}_0$, $\hat{a}_1$, $\hat{b}_0$ and $\hat{b}_1$ so that $a_0 = \vec{\sigma} \cdot \hat{a}_0$ and so forth, where $\vec{\sigma} = \hat{x}\sigma_x + \hat{y}\sigma_y + \hat{z}\sigma_z$ The state $|\Psi\rangle_{AB}$, which is the spin-singlet combination of two spin-$\frac{1}{2}$ particles, is rotationally invariant, meaning that $U_A \otimes U_B |\Psi\rangle_{AB} = |\Psi\rangle_{AB}$ for any unitary U with $\det U = 1$. From rotation invariance it follows that

$$\langle \Psi | (\vec{\sigma}_A \cdot \hat{a})(\vec{\sigma}_B \cdot \hat{b}) |\Psi\rangle_{AB} = -\hat{a} \cdot \hat{b}. \tag{7.35}$$

To see this, compute

$$(\vec{\sigma}_A \cdot \hat{a})(\vec{\sigma}_B \cdot \hat{b})|\Psi\rangle_{AB} = \sum_{jk} a_j b_k (\sigma_j \otimes \sigma_k)|\Psi\rangle_{AB} \tag{7.36}$$

$$= -\sum_{jk} a_j b_k (\mathrm{id} \otimes \sigma_k \sigma_j)|\Psi\rangle_{AB}. \tag{7.37}$$

The second equality holds because $\sigma_j \otimes \sigma_j |\Psi\rangle = -|\Psi\rangle$; $\det(\sigma_j) = -1$, so it is $i\sigma_j$ that has unit determinant Then, in the inner product above only the terms with $j = k$ contribute to the sum, since states of the form $\mathrm{id} \otimes \sigma_k |\Psi\rangle_{AB}$ have nonzero angular momentum.

Now choose $\hat{a}_0 = \hat{x}$, $\hat{a}_1 = \hat{y}$, $\hat{b}_0 = \frac{1}{\sqrt{2}}(\hat{x} + \hat{y})$, and $\hat{b}_1 = \frac{1}{\sqrt{2}}(\hat{x} - \hat{y})$. This gives

$$\langle a_0 b_0 \rangle = \langle a_1 b_0 \rangle = \langle a_0 b_1 \rangle = -\tfrac{1}{\sqrt{2}} \qquad \text{and} \tag{7.38}$$

$$\langle a_1 b_1 \rangle = \tfrac{1}{\sqrt{2}}, \tag{7.39}$$

so that $|\langle C \rangle| = 2\sqrt{2} \nleq 2$. Therefore, Einstein's goal of a locally realistic version of quantum mechanics is impossible.

The use of entangled states is necessary in this argument; no non-entangled states can violate the CHSH inequality. Schrödinger explained the importance of entangled states quite well, though before the advent of Bell inequalities:

> When two systems, of which we know the states by their respective representatives, enter into temporary physical interaction due to known forces between them, and when after a time of mutual influence the systems separate again, then they can no longer be described in the same way as before, viz. by endowing each of them with a representative of its own. I would not call that *one* but rather *the* characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought. By the interaction the two representatives (or ψ-functions) have become entangled.[28]

The violation of the CHSH inequality also highlights the danger of reasoning from counterfactuals in quantum mechanics. It simply is not possible to consider the consequences of hypothetical operations or measurements in quantum mechanics that are not actually performed. Peres put it best in the title of a paper on the subject of Bell inequalities: *Unperformed experiments have no results.*[29]

Interestingly, it was recently discovered that the CHSH inequality is closely related to the question of whether two ±1-valued observables can be jointly measured. Suppose that the two measurements to be jointly measured are the a_0 and a_1 of the CHSH setup. Then, in [30] it is shown that the pair of measurements is not jointly measurable in the sense of §7.2.1 if and only if it is possible to find another pair of observables b_0 and b_1 and a quantum state such that the CHSH inequality is violated.

7.4.2 Tsirel'son's inequality

The value $|\langle C\rangle| = 2\sqrt{2}$ is actually the largest possible in quantum mechanics, a fact known as Tsirel'son's inequality. To prove it, consider the quantity C^2 for a_0, a_1, b_0, and b_1 arbitrary Hermitian operators which square to the identity (so that their eigenvalues are ±1), and for which $[a_x, b_y] = 0$. By direct calculation we find

$$C^2 = 4\mathrm{id} - [a_0, a_1][b_0, b_1]. \tag{7.40}$$

Now compute the infinity norm of C^2, which is defined by

$$||C^2||_\infty := \sup_{|\psi\rangle} \left(\frac{||C^2|\psi\rangle||}{|||\psi\rangle||} \right). \tag{7.41}$$

The infinity norm has the following two properties, (i) $||AB||_\infty \leq ||A||_\infty ||B||_\infty$ and (ii) $||A+B||_\infty \leq ||A||_\infty + ||B||_\infty$. Then, we have

$$\begin{aligned}
||C^2||_\infty &= ||4\mathrm{id} - [a_0, a_1][b_0, b_1]||_\infty & (7.42)\\
&\leq 4 + ||[a_1, a_0]||_\infty + ||[b_0, b_1]||_\infty & (7.43)\\
&\leq 4 + ||a_1||_\infty (||a_0||_\infty + ||-a_0||_\infty) + ||b_0||_\infty (||b_1||_\infty + ||-b_1||_\infty) & (7.44)\\
&= 8. & (7.45)
\end{aligned}$$

In the last step we used the fact that $||\pm c||_\infty = 1$ for c having eigenvalues ±1.

7.4.3 The CHSH game

There is a slightly different way to formulate the CHSH setup which directly reveals the connection to the principle of no superluminal signalling. Abstractly, the CHSH scenario consists of Alice and Bob choosing inputs x and y (their choice of measurements) and then receiving outputs a and b (the measurement results). For later convenience, here we change the convention slightly and regard a and b as also taking on values 0 or 1.

Now consider a game whose goal is to produce outputs a and b given inputs x and y such that $a \oplus b = x \cdot y$. If the outputs a_x and b_y have fixed values, then it is easy to see that there is no way to win the game for all possible inputs x and y. This is because a_x and b_y must satisfy $a_x \oplus b_y = x \cdot y$, but

$$\sum_{xy} a_x \oplus b_y = 0 \qquad \text{while} \qquad \sum_{xy} x \cdot y = 1. \tag{7.46}$$

Examination of the 16 possible settings of a_x and b_y shows that at best Alice and Bob can win with probability 3/4. For instance, $a_0 = 1$, $a_1 = 0$, $b_0 = 0$, and $b_1 = 1$ obeys $a_x \oplus b_y = x \cdot y$ in only three cases, with $x = y = 0$ giving $a_0 + b_0 = 1$. Similarly, $a_0 = 0$, $a_1 = 1$, $b_0 = 0$, and $b_1 = 1$ fails in three cases, only $x = y = 0$ being correct. Mixing these deterministic assignments does not change the bound, so we have found that

$$P(a \oplus b = \cdot y) = \sum_\lambda p(\lambda) P(a|x,\lambda) P(b|y,\lambda) \leq \tfrac{3}{4}, \tag{7.47}$$

where the conditional distributions $P(a|x,\lambda)$ and $P(b|y,\lambda)$ are deterministic.

But the form of the distribution is the most general possible for a deterministic local hidden variable theory, so $P(a \oplus b = x \cdot y) \leq \frac{3}{4}$ is a Bell inequality. Actually it is just a restatement of the CHSH inequality. To see this, let $p_{xy} = P(a \oplus b = x \cdot y|x,y)$. Then each term in C is related to a different p_{xy}. Consider $p_{0,1}$. Denoting by $a'_x = (-1)^{a_x}$ and $b'_y = (-1)^{b_y}$ the original ±1-valued observables, we have

$$\langle a'_0 b'_1 \rangle = \langle (-1)^{a_0 + b_1} \rangle = p_{01} - (1 - p_{01}) = 2p_{01} - 1, \tag{7.48}$$

since $x = 0, y = 1$ means the value of $a'_0 b'_1$ will be $+1$ if they win and -1 if they lose. Similarly, $\langle a'_0 b'_0 \rangle = 2p_{00} - 1$, $\langle a'_1 b'_0 \rangle = 2p_{10} - 1$, while $\langle a'_1 b'_1 \rangle = 1 - 2p_{11}$. In the last case, $a'_1 b'_1$ is -1 if they win and $+1$ if they lose. The CHSH inequality $|\langle C \rangle| \leq 2$ then translates into

$$|\langle C \rangle| = 2 \sum_{xy} p_{xy} - 4 = 2 \cdot 4 p_{\text{win}}^{\text{DLHV}} - 4 \leq 2, \tag{7.49}$$

or $p_{\text{win}}^{\text{DLHV}} \leq \frac{3}{4}$, where $p_{\text{win}}^{\text{DLHV}}$ denotes the probability of winning the game when x and y are chosen randomly, when using a strategy described by a deterministic local hidden variable theory. Using quantum mechanics, we have $|\langle C \rangle| \leq 2\sqrt{2}$, so $p_{\text{win}}^{\text{QM}} \leq \frac{1}{2} + \frac{1}{2\sqrt{2}}$.

The maximum winning probability is 1, of course, and is achieved by the distribution $P(a,b|x,y) = \frac{1}{2}\delta_{a \oplus b, x \cdot y}$. Interestingly, this distribution also does not allow for superluminal signalling, even though the non-local correlations are much stronger than in quantum mechanics. Here, $|\langle C \rangle| = 4$. Nevertheless, the distribution obeys $P(a|x,y) = P(a|x)$, so that the marginal probability of outcome a depends only on the x setting and not the y setting. As much holds in the other direction. This precludes signalling from one party to the other.

8

Quantum Entropy

8.1 The von Neumann entropy and its properties

8.1.1 Entropy of a single system

The von Neumann entropy is just the Shannon entropy of the state's eigenvalues, i.e.,

$$H(\rho) := -\sum_k \lambda_k \log \lambda_k = -\mathrm{tr}[\rho \log \rho]. \tag{8.1}$$

Recall that a function (over $\mathbb{R}$) of a Hermitian operator is to be interpreted as the function applied to the eigenvalues. Some of its properties are

1. $H(\rho) = 0$ iff ρ is a pure state
2. $H(U\rho U^*) = H(\rho)$ for unitary U
3. $H(\rho) \leq \log|\mathrm{supp}\,\rho|$
4. $H\left(\sum_k p_k \rho_k\right) \geq \sum_k p_k H(\rho_k)$
5. $H\left(\sum_k P_k \rho P_k\right) \geq H(\rho)$ for any complete set of projectors P_k.

Properties 1 and 2 are clear by inspection. Just like the classical case, the rest follow from positivity of the relative entropy, a statement known as Klein's inequality[1]

$$D(\rho||\sigma) = \mathrm{tr}[\rho(\log \rho - \log \sigma)] \geq 0, \tag{8.2}$$

with equality iff $\rho = \sigma$. Property 3 follows by defining $\sigma = \frac{1}{d}\mathrm{id}$, where id is the identity operator on $\mathrm{supp}\rho$ and $d = |\mathrm{supp}\rho|$. Then, $D(\rho||\sigma) = \log d - H(\rho) \geq 0$. To prove property 4 let $\rho = \sum_k p_k \rho_k$. Then $\sum_k p_k D(\rho_k||\rho) = H(\rho) - \sum_k p_k H(\rho_k) \geq 0$. Finally to prove property 5, let $\overline{\rho} = \sum_k P_k \rho P_k$. Observe that $[P_k, \overline{\rho}] = 0$. Thus, $D(\rho||\overline{\rho}) = -H(\rho) - \sum_k \mathrm{tr}[P_k(\rho \log \overline{\rho})P_k] = H(\overline{\rho}) - H(\rho) \geq 0$.

8.1.2 von Neumann entropy of several systems

The von Neumann entropy of a single quantum system behaves very similar to the Shannon entropy of a single random variable. But not so for the entropy of several quantum systems! The reason is entanglement, of course. The simplest example is a bipartite pure sate $|\psi\rangle_{AB} = \sum_k \sqrt{p_k}|\varphi_k\rangle_A|\xi_k\rangle_B$, here expressed in Schmidt form. The marginal states of A and B share the same eigenvalues, and we have immediately $H(AB)_\rho = 0$ while $H(A)_\rho = H(B)_\rho$. Thus the entropy (uncertainty) of the whole of an entangled state is *less* than that of the parts, something impossible (and nonsensical) for classical random variables.

The joint entropy does obey the following conditions:

1. Subadditivity: $H(AB) \leq H(A) + H(B)$
2. Triangle inequality: $H(AB) \geq |H(A) - H(B)|$.

[1] Oskar Benjamin Klein, 1894 – 1977, Swedish theoretical physicist.

To see the former, simply compute the relative entropy $D(\rho_{AB}||\rho_A \otimes \rho_B)$, where $\rho_A = \mathrm{tr}_B[\rho_{AB}]$ and similarly for ρ_B. The latter follows from the former by making use of a third purifying reference system R: Let $|\psi\rangle_{RAB}$ be a purification of ρ_{AB}, then

$$H(B) = H(RA) \leq H(A) + H(R) = H(A) + H(AB), \tag{8.3}$$

which implies that $H(AB) \geq H(B) - H(A)$. Swapping A and B in the proof gives the absolute value. It is also easy to see, by direct calculation, that $H(AB) = H(A) + H(B)$ for states of the form $\rho_{AB} = \rho_A \otimes \rho_B$.

8.1.3 Conditional von Neumann entropy and quantum mutual information

We can define the conditional entropy and mutual information using the classical entropy expression:

$$H(A|B) := H(AB) - H(B) \tag{8.4}$$

and

$$I(A:B) := H(A) + H(B) - H(AB). \tag{8.5}$$

Form the discussion above, we see that the conditional entropy can be negative, so perhaps it is best not to think of it as a conditional uncertainty.[2] The quantum mutual information is positive, however, owing to subadditivity. Moreover, by the triangle inequality, $H(AB) \geq H(A) - H(B)$ which implies $2H(B) \geq H(A) + H(B) - H(AB) = I(A:B)$ and thus $I(A:B) \leq 2\min\{H(A), H(B)\}$.

An important property of the conditional von Neumann entropy is *duality*,

$$H(A|B)_\rho + H(A|C)_\rho = 0, \tag{8.6}$$

for all pure states ρ_{ABC}. This follows immediately from the fact that marginals of bipartite pure states have identical eigenvalues and will turn out to have several important implications.

8.1.4 Entropy of CQ states

Classical-quantum states do obey the usual rules of entropy. In particular, $H(X|B) \geq 0$. To show this, let us start by calculating the joint entropy of the CQ state $\rho_{XB} = \sum_x p_x (P_x)_X \otimes (\rho_x)_B$.

$$\begin{aligned}
H(XB) &= -\mathrm{tr}[\rho_{XB} \log \rho_{XB}] && (8.7)\\
&= -\mathrm{tr}[\sum_x p_x P_x \otimes \rho_x \sum_y P_y \otimes \log p_y \rho_y] && (8.8)\\
&= -\mathrm{tr}[\sum_x p_x P_x \otimes \rho_x \log p_x \rho_x] && (8.9)\\
&= -\mathrm{tr}[\sum_x p_x P_x \otimes \rho_x (\log p_x + \log \rho_x)] && (8.10)\\
&= H(X) + \sum_x p_x H(\rho_x) && (8.11)\\
&= H(X) + H(B|X). && (8.12)
\end{aligned}$$

Here we have proven the fact that $H(\sum_x p_x \rho_x) = H(X) + \sum_x p_x H(\rho_x)$, when ρ_x are disjoint.

[2] Nevertheless $H(A|B) \leq H(A)$ by subadditivity.

We have actually already encountered the CQ mutual information $I(X:B)$ in our consideration of the concavity of entropy in the previous section:

$$I(X:B)=H(X)+H(B)-H(XB)=H\left(\sum_x p_x\rho_x\right)-\sum_x p_x H(\rho_x), \tag{8.13}$$

since $\rho_B=\sum_x p_x\rho_x$. This quantity is quite important in quantum information theory and usually goes under the name Holevo information for reasons we shall see shortly.

In contrast to $I(A:B)$, we can prove $I(X:B)\le H(X)$, which is equivalent to $H(X|B)\ge 0$.

Lemma 8.1.1. *$I(X:B)_\rho \le H(X)_\rho$ for any CQ state ρ_{XB}.*

Proof. The standard proof (for instance in Nielsen & Chuang) is to work in the ensemble picture and bootstrap form the pure state case to the mixed state case. But we can avoid having to directly compute too much by considering the following pure state

$$|\psi\rangle_{ABR}=\sum_x \sqrt{p_x}|x\rangle_A|\varphi_x\rangle_{BR}, \quad \text{where } \rho_x=\text{tr}_R[|\varphi_x\rangle\langle\varphi_x|_{BR}]. \tag{8.14}$$

First, since RAB is pure, $H(AR)=H(B)=H\left(\sum_x p_x\rho_x\right)$. Second, measurement increases entropy, so $H\left(\overline{A}R\right)\ge H(AR)$, where $\overline{A}R$ denotes the state after AR measuring A in the $|x\rangle$ basis:

$$H\left(\overline{A}R\right)=H\left(\sum_x p_x P_x\otimes\tilde{\rho}_x\right) \quad \text{for } \tilde{\rho}_x=\text{tr}_B[|\varphi_x\rangle_{BR}]. \tag{8.15}$$

By the results of the joint entropy calculation above

$$H\left(\overline{A}R\right)=H(X)+\sum_x p_x H(\tilde{\rho}_x)=H(X)+\sum_x p_x H(\rho_x), \tag{8.16}$$

where the last step follows because ρ_x and $\tilde{\rho}_x$ are both marginals of the same pure state $|\varphi_x\rangle$. Altogether we have

$$H(X)+\sum_x p_x H(\rho_x)\ge H\left(\sum_x p_x\rho_x\right), \tag{8.17}$$

which implies $H(X)\ge I(X:B)$. □

8.2 Strong Subadditivity

The most important entropy inequality deals with three systems and states that

$$H(A|BC)\le H(A|B), \quad \text{or equivalently} \tag{8.18}$$
$$I(A:BC)\ge I(A:B). \tag{8.19}$$

Classically this inequality is easy to prove; it follows directly form subadditivity:

$$H(X|YZ)=\sum_y p(y)H(X|Y=y,Z)\le\sum_y p(y)H(X|Y=y)=H(X|Y). \tag{8.20}$$

But this will not work quantum mechanically *because there is no notion of a conditional state.* The usual proof for the von Neumann entropy makes use of Lieb's theorem, that the function $f(A,B)=$

$\mathrm{tr}[X^*A^tXB^{1-t}]$ is jointly concave in positive matrices A, B for all X and $0 \leq t \leq 1$. The entire proof is quite long so we will not go into it here.

However, it is worth mentioning that strong subadditivity (SSA) is nearly trivial to prove for the conditional min-entropy. This then implies the result for the von Neumann entropy by going to the smooth min-entropy of the asymptotic i.i.d. case.

Recall that classically the min-entropy of a random variable X is just $-\log$ of the largest probability. Similarly, the min-entropy of a quantum system ρ is $-\log\lambda$ where λ is the smallest number such that $\lambda\,\mathrm{id}-\rho \geq 0$. This extends immediately to the conditional min-entropy: $H_{\min}(\rho_{AB}|\sigma_B) = -\log\lambda$, where λ minimum real such that $\lambda\mathrm{id}_A \otimes \sigma_B - \rho_{AB} \geq 0$. SSA is then immediate. Define λ by $H_{\min}(\rho_{ABC}|\sigma_{BC}) = -\log\lambda$ so that $\lambda\mathrm{id}_A \otimes \sigma_{BC} - \rho_{ABC} \geq 0$. Taking the partial trace preserves this inequality, meaning $\lambda\mathrm{id}_A \otimes \sigma_B - \rho_{AB} \geq 0$ and therefore $H_{\min}(\rho_{AB}|\sigma_B) \geq -\log\lambda$.

The reason SSA is so important is that many useful relations can be derived from it. In particular we shall focus on two: the concavity of the conditional entropy and the fact that local quantum operations cannot increase mutual information.

Lemma 8.2.1 (Concavity of conditional entropy). $H(A|B)_\rho \geq \sum_x p_x H(A|B)_{\rho_x}$ *for* $\rho = \sum_x p_x\rho_x$.

Proof. Consider the state $\rho^{ABX} = \sum_x p(x)\rho_x^{AB} \otimes P_x^X$ and apply $H(A|BX) \leq H(A|B)$. Since X is classical, we can condition on it in the usual way:

$$H(A|BX) = \sum_x p(x)H(A|B)_{\rho_x} \leq H(A|B)_\rho. \tag{8.21}$$

Just to be sure, taking the long way gives

$$H(A|BX) = H(ABX) - H(BX) \tag{8.22}$$

$$= H(X) + \sum_x p_x H(AB)_{\rho_x} - H(X) - \sum_x p_x H(B)_{\rho_x} \tag{8.23}$$

$$= \sum_x p_x H(A|B)_{\rho_x}. \tag{8.24}$$

□

This result is interesting because it tells us that negative $H(A|B)$ is a sign of entanglement (which we might well have suspected already). If ρ_{AB} is pure, then $H(AB) \leq 0$ implies $H(B) > 0$ (indeed $-H(A|B) = H(B)$) and therefore there is more than one Schmidt coefficient. For the case of mixed states, consider $H(A|B)$ for a separable state $\rho^{AB} = \sum_j p_j\sigma_j^A \otimes \xi_j^B$:

$$H(A|B)_\rho \geq \sum_j p_j H(A|B)_j = \sum_j p_j H\left(\sigma_j\right) \geq 0. \tag{8.25}$$

Therefore $H(A|B)_j < 0$ implies ρ is not separable, i.e. ρ is entangled. However, the converse is false: There exist entangled states for which $H(A|B) \geq 0$. Thus, the conditional entropy is not a *faithful* measure of entanglement.

Nonetheless, the duality of conditional entropy translates into the *monogamy* property of entanglement, the fact that a system A cannot be entangled with both B and C at the same time.

The other application of SSA we are interested in is the fact that local quantum operations cannot increase the quantum mutual information, the quantum data processing inequality

Lemma 8.2.2 (Quantum data processing inequality)**.** *For all bipartite states ρ_{AB} and CPTP maps $\mathcal{E}$, $I(A:B)_\rho \geq I(A:B')_{\rho'}$, where $\rho'_{AB'} = \mathcal{I} \otimes \mathcal{E}(\rho_{AB})$.*

Proof. Using the Stinespring representation, we have for some isometry U_{BR}

$$\rho'_{AB} = \mathrm{tr}_R[U_{BR}\rho_{AB}U^*_{BR}] = \mathrm{tr}_R[\psi_{ABR}]. \tag{8.26}$$

As entropy is invariant under isometries, $I(A:B') \leq I(A:BR)_\psi = I(A:B)_\rho$. □

The *Holevo bound* is the data processing inequality for CQ states, and was first established independently of SSA:

Corollary 8.2.3 (Holevo bound)**.** *For any CQ state ρ_{XB} and POVM $\{\Lambda_y\}$ on system B producing the random variable Y, $I(X:B)_\rho \geq I(X:Y)_{\rho'}$.*

Along with the bound $I(X:B) \leq H(X)$, the Holevo bound shows that n qubits cannot be used to carry more than n classical bits about a classical random variable.

8.3 Entropic Uncertainty Relations

From the duality of the conditional von Neumann entropy we can derive two entropic uncertainty relations. The first deals with three parties, and to a certain extent captures the notion that non-commuting observables cannot be simultaneously measured. The second deals with two parties and relates the ability of one system to predict the value of two observables (however, not simultaneously) of the other to their shared entanglement.

The proof relies on several facts about the quantum relative entropy which we will use here without proof: invariance under isometries, monotonicity under CPTP maps (which relies on joint convexity in both arguments), and $D(\rho||\sigma') \leq D(\rho||\sigma)$ for $\sigma \leq \sigma'$.

Theorem 8.3.1 (Entropic Uncertainty)**.** *Given two observables X and Z on a quantum system A, let $|\varphi_x\rangle$ and $|\vartheta_z\rangle$ be the eigenstates of X and Z, respectively and define $c(X,Z) = \max_{xz} |\langle\varphi_x|\vartheta_z\rangle|^2$. Then, for any state ρ_{ABC} and $H(X^A|B)_\rho$ the entropy of the result of measuring X on A conditional on system B, and similarly for $H(Z^A|C)$, we have*

$$H(X^A|B)_\rho + H(Z^A|C)_\rho \geq \log\frac{1}{c(X,Z)}, \quad \textit{and} \tag{8.27}$$

$$H(X^A|B)_\rho + H(Z^A|B)_\rho \geq \log\frac{1}{c(X,Z)} + H(A|B)_\rho. \tag{8.28}$$

Proof. The proof proceeds by showing the first and then deriving the second as a simple consequence. To prove the first statement, observe that by data processing it is sufficient to establish the statement for pure ρ_{ABC}. Then consider the state

$$|\psi\rangle_{XX'BC} = V_{A\to XX'}|\rho\rangle_{ABC} = \sum_x |x\rangle_X|x\rangle_{X'\,A}\langle\varphi_x|\rho\rangle_{ABC}, \tag{8.29}$$

where $V_{A\to XX'}$ is a Stinespring dilation of the measurement process. Applying entropy duality, we have $H(X|B)_\psi + H(X|X'C)_\psi = 0$. By direct calculation it is easy to show that $H(X|X'C) = -D(\psi_{XX'C}||\mathrm{id}_X \otimes \rho_{X'C})$, and thus $H(X|B)_\psi = D(\psi_{XX'C}||\mathrm{id}_X \otimes \rho_{X'C})$. By invariance of the relative entropy under isometries, it follows that $D(\psi_{XX'C}||\mathrm{id}_X\otimes\rho_{X'C})=D(\rho_{AC}||V^*_{A\to XX'}\mathrm{id}_X\otimes\rho_{X'C}\,V_{A\to XX})$. The second argument to the relative entropy is just

$$V^*_{A\to XX'}\mathrm{id}_X\otimes\rho_{X'C}\,V_{A\to XX}=\sum_{xx'}|\varphi_{x'}\rangle_A\langle x'|_X\langle x'|_{X'}\mathrm{id}_X\otimes\rho_{X'C}|x\rangle_X|x\rangle_{X'}\langle\varphi_x|_A \tag{8.30}$$

$$=\sum_x|\varphi_x\rangle\langle\varphi_x|_A\,\mathrm{tr}_{X'}[|x\rangle\langle x|_{X'}\rho_{X'C}]. \tag{8.31}$$

But applying the isometry $U=\sum_z|z\rangle_Z|z\rangle_{Z'}\langle\vartheta_z|_A$ to the relative entropy gives

$$D(\psi_{XX'C}||\mathrm{id}_X\otimes\rho_{X'C})=D(\psi_{ZZ'C}||\sum_x U|\varphi_x\rangle\langle\varphi_x|U^*\,\mathrm{tr}_{X'}[|x\rangle\langle x|_{X'}\rho_{X'C}]) \tag{8.32}$$

$$\geq D(\psi_{ZC}||\sum_x\mathrm{tr}_{Z'}[U|\varphi_x\rangle\langle\varphi_x|U^*]\,\mathrm{tr}_{X'}[|x\rangle\langle x|_{X'}\rho_{X'C}], \tag{8.33}$$

where we have used the monotonicity of the relative entropy under CPTP maps. Again we can simplify the second argument, as follows:

$$\sum_x\mathrm{tr}_{Z'}[U|\varphi_x\rangle\langle\varphi_x|U^*]\,\mathrm{tr}[|x\rangle\langle x|_{X'}\rho_{X'C}]=\sum_{xz}|\langle\vartheta_z|\varphi_x\rangle|^2|z\rangle\langle z|_Z\otimes\mathrm{tr}_{X'}[|x\rangle\langle x|_{X'}\rho_{X'C}] \tag{8.34}$$

$$\leq\sum_{xz}c(X,Z)|z\rangle\langle z|_Z\otimes\mathrm{tr}_{X'}[|x\rangle\langle x|_{X'}\rho_{X'C}] \tag{8.35}$$

$$=c(X,Z)\,\mathrm{id}_Z\otimes\rho_C. \tag{8.36}$$

Since $D(\rho||\sigma')\leq D(\rho||\sigma)$ for $\sigma\leq\sigma'$,

$$D(\psi_{XX'C}||\mathrm{id}_X\otimes\rho_{X'C})\geq D(\psi_{ZC}||c(X,Z)\,\mathrm{id}_Z\otimes\rho_C) \tag{8.37}$$

$$=D(\psi_{ZC}||\mathrm{id}_Z\otimes\rho_C)-\log c(X,Z) \tag{8.38}$$

$$=-H(Z|C)_\rho+\log\frac{1}{c(X,Z)}, \tag{8.39}$$

completing the proof of the first statement.

For the second, it is a simple calculation to verify that $H(Z^AB)_\rho=H(Z^AC)_\rho$ when C is the purification of AB so that ρ_{ABC} is pure. This leads immediately to $H(Z^A|C)_\rho=H(Z^A|B)_\rho-H(A|B)_\rho$. If C is not the purification of AB, then by data processing $H(Z^A|C)_\rho\geq H(Z^A|B)_\rho-H(A|B)_\rho$. Using this expression to replace $H(Z^A|C)$ in the first statement leads to the second. □

9 The Resource Framework

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.

9.1 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:

- $n[q \to q]$: Alice sends n qubits to Bob
- $n[c \to c]$: Alice sends n bits to Bob
- $n[qq]$: Alice and Bob share n maximally entangled states
- $n[cc]$: Alice and Bob share n random classical bits

A resource inequality is a relation $X \geq Y$, which means there exists a protocol to simulate resources Y using only resources X and local operations. For superdense coding the resource inequality reads

$$[qq] + [q \to q] \geq 2[c \to c], \tag{9.1}$$

while teleportation is the inequality

$$2[c \to c] + [qq] \geq [q \to q]. \tag{9.2}$$

Of course, resources are usually not perfect and nor do we require the resource conversion to be perfect. We can then still use resource inequalities to formulate meaningful statements. For instance, Shannon's noiseless channel coding theorem for a channel W of capacity $C(W)$ reads

$$n[W] \geq_{\epsilon} n(C(W) - \epsilon)[c \to c], \tag{9.3}$$

for all $\epsilon > 0$ and n large enough.

In the remainder we will only be concerned with an exact conversion of perfect resources. Our main goal will be to show that the teleportation and superdense coding protocols are optimal.

9.2 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 9.2.1 (Quantum mutual information)**.** For bipartite states, the quantum mutual information is a monotone for the class of local operations, as was shown in Lemma 8.2.2. 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 9.2.2 (Squashed entanglement)**.** The squashed entanglement of a state ρ_{AB} is given by

$$E_{\mathrm{sq}}(A:B):=\tfrac{1}{2}\inf_{E} I(A:B|E), \tag{9.4}$$

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 (often abbreviated as LOCC). 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.

Consider the case where Alice sends a classical system C to Bob (e.g. a bit string). We want to compare $E_{\mathrm{sq}}(AC:B)$ and $E_{\mathrm{sq}}(A:BC)$. 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 $E_{\mathrm{sq}}(AC:B)\geq E_{\mathrm{sq}}(A:BC)$. By symmetry $E_{\mathrm{sq}}(AC:B)=E_{\mathrm{sq}}(A:BC)$ follows.

Theorem 9.2.3. *For any state ρ_{AB}, $E_{\mathrm{sq}}(A:B)=0$ iff ρ_{AB} is separable.*

Proof. We only prove here that a separable state ρ_{AB} implies that $E_{\mathrm{sq}}(A:B)=0$. The converse is beyond the scope of this course and has only been established recently [31]. Consider the following separable classical-quantum state

$$\rho_{ABC}=\sum_i p_i\rho_A^i\otimes\rho_B^i\otimes|i\rangle\langle i|_C, \tag{9.5}$$

for p_i a probability distribution. Using the definition of the mutual information we can write

$$\begin{aligned}
I(A:B|C)_\rho &= H(A|C)_\rho-H(B|AC)_\rho && (9.6)\\
&= \sum_i p_i H(A)_{\rho_A^i}-\sum_i p_i H(A|B)_{\rho_A^i\otimes\rho_B^i} && (9.7)\\
&= 0. && (9.8)
\end{aligned}$$

The first two equalities follow by definition and the final step by the chain rule:

$$H(A|B)_{\rho_A\otimes\rho_B}=H(AB)_{\rho_A\otimes\rho_B}-H(B)_{\rho_B}=H(A)_{\rho_A}. \tag{9.9}$$

□

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}:

$$E_D(\rho_{AB}) = \lim_{\epsilon \mapsto 0} \lim_{n \mapsto \infty} \sup_{\Lambda \text{ LOCC}} \{\frac{m}{n} : \langle \Phi|^{\otimes m} \Lambda(\rho_{AB}^{\otimes n})|\Phi\rangle^{\otimes m} \geq 1 - \epsilon\}$$

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.

9.3 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 the resource notation this reads

$$n[qq] + \infty[c \to c] \geq m[q \to q] \quad \Longrightarrow \quad n \geq m. \tag{9.10}$$

Observe that

$$n[qq] + \infty[c \to c] \geq m[q \to q] \quad \longrightarrow \quad n[qq] + \infty[c \to c] \geq m[qq], \tag{9.11}$$

since we can use the quantum channel to distribute entangled pairs. Thus, we need only show that the number of ebits cannot be increased by classical communication. This certainly sounds intuitive, but proof requires using a monotone, for instance squashed entanglement. Since every possible extension ρ_{ABE} of a pure state ρ_{AB} of n ebits is of the form $\rho_{ABE} = \rho_{AB} \otimes \rho_E$ we find

$$2E_{\text{sq}}(A:B)_\rho = \inf_E I(A:B|E)_\rho = I(A:B)_\rho = 2n. \tag{9.12}$$

According to (9.12), having n ebits can be expressed in term of squashed entanglement. Since local operations and classical communication cannot increase the squashed entanglement as shown in Example 9.2.2, we conclude using again (9.12) that it is impossible to increase the number of ebits by LOCC.

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.

9.4 Superdense coding is optimal

Now we would like to prove that at least one qubit channel is needed to send two classical bits, regardless of how many ebits are available:

$$n[q \to q] + \infty[qq] \geq 2m[c \to c] + \infty[qq] \quad \longrightarrow \quad n \geq m. \tag{9.13}$$

First observe that concatenation of

$$n[q \to q] + \infty[qq] \geq 2m[c \to c] + \infty[qq] \tag{9.14}$$

with teleportation yields

$$n[q \to q] + \infty[qq] \geq m[q \to q] + \infty[qq]. \tag{9.15}$$

Thus, we need to show that shared entanglement does not enable Alice to send additional qubits to Bob. 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 than two times the number of qubits that he receives from Alice, i.e. by $2n$. For this we bound Bob's final quantum mutual information with Charlie by

$$\begin{aligned} I(C:B_f) &\leq I(C:B_iQ) && (9.16)\\ &= I(C:B_i) + I(C:Q|B_i) && (9.17)\\ &\leq I(C:B_i) + 2n && (9.18) \end{aligned}$$

Therefore $m \leq n$. This concludes our proof that the superdense coding protocol is optimal.

For this argument we did not use a monotone such as squashed entanglement. Instead, 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*.

10

Quantum Data Compression and Entanglement Purification

In this chapter we examine two tasks of quantum information processing which, in some sense, only deal with a single quantum system: quantum data compression and entanglement purification.

10.1 Quantum data compression

Quantum data compression is simply the quantum version of classical data compression. The setup is depicted in Figure 10.1. A source produces a bipartite pure state $|\psi\rangle_{SR}$ and delivers S to the compressor. The goal of the data compression scheme is for the compressor to send as few qubits (system C) to the decompressor as possible, such that approximate reconstruction of the state is nonetheless possible.

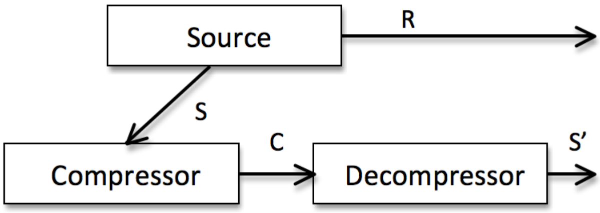


Figure 10.1: Setup of quantum data compression.

There are two distinct figures of merit for the state reconstruction. First, how well does the output state $\rho'_{S'}$ approximate the input $\rho_S = \mathrm{tr}_R[|\psi\rangle\langle\psi|_{RS}]$? Second, how well does does $\psi'_{S'R}$ approximate ψ_{SR}? For classical data compression there is no meaning to the second, since there is no notion of purification. But as we shall see, there is a big difference between the two in the quantum case. Essentially, the point is that for the joint output in $S'R$ to approximate the joint input, there can be no information about the source left at the compressor. In the classical case, whether the compressor retains a copy of its input is of no concern.

10.1.1 A direct approach

Let us now try to construct a quantum compression scheme by directly reusing the classical protocol as much as possible. We can do so by first observing that the quantum task can be mapped to the classical task by working in the eigenbasis of $\rho_S = \sum_x p_x |x\rangle\langle x|_S$. This defines a random variable X with probability distribution $P_X(x) = p_x$.

Now suppose we have a data compression protocol for X, based on some compression function f. We can transform this function into a quantum map by defining its action on the basis vectors, like so

$$\mathscr{C}(|x\rangle\langle x|_S) = |f(x)\rangle\langle f(x)|_C. \tag{10.1}$$

Meanwhile, the classical decompressor takes the compressed input y and outputs the x for which the conditional probability $P_{X|Y=y}$ is the largest. Calling this mapping $r(y)$, we can again promote it to a quantum operation

$$\mathscr{D}(|y\rangle\langle y|_C) = |r(y)\rangle\langle r(y)|_{S'}. \tag{10.2}$$

What is the trace distance of ρ_S and $\rho'_{S'} := \mathscr{D}\circ\mathscr{C}(\rho_S)$? (Since S and S' are isomorphic, we can ignore the distinction between them.) First, we have

$$\rho' := \mathscr{D}\circ\mathscr{C}(\rho) = \sum_x p_x |r(f(x))\rangle\langle r(f(x))| \tag{10.3}$$

Let $\mathscr{X}_g$ be the set of x that are always correctly recovered, i.e. $\mathscr{X}_g = \{x \in \mathscr{X} | r(f(x)) = x\}$. These are the x values with the highest conditional probabilities in each set having fixed value y under f. Therefore,

$$\rho' = \sum_{x\in\mathscr{X}_g} p_x|x\rangle\langle x| + \sum_{x\notin\mathscr{X}_g} p_x|r(f(x))\rangle\langle r(f(x))| = \sum_{x\in\mathscr{X}_g} p'_x|x\rangle\langle x|, \tag{10.4}$$

where p'_x includes the contribution from $\mathscr{X}_g$ and $\mathscr{X}_g^C$. It follows that$p'_x > p_x$ for all $x \in \mathscr{X}_g$. Now,

$$\tfrac{1}{2}||\mathscr{E}(\rho)-\rho||_1 = ||\sum_{x\in\mathscr{X}_g} p'_x|x\rangle\langle x| - \sum_{x\in\mathscr{X}_g} p_x|x\rangle\langle x| - \sum_{x\notin\mathscr{X}_g} p_x|x\rangle\langle x|\,||_1 \tag{10.5}$$

$$= \frac{1}{2}\sum_{x\notin\mathscr{X}_g} p_x + \frac{1}{2}\sum_{x\in\mathscr{X}_g}(p'_x - p_x) \qquad \text{with} \quad \sum_{x\in\mathscr{X}_g} p'_x = 1 \tag{10.6}$$

$$= \frac{1}{2}\sum_{x\notin\mathscr{X}_g} p_x + 1 - \sum_{x\in\mathscr{X}_g} p_x \tag{10.7}$$

$$= \sum_{x\notin\mathscr{X}_g} p_x = p_{\text{err}} \tag{10.8}$$

Therefore, if the classical compression scheme has error probability $p_{\text{err}} \le \varepsilon$, we can find a quantum compression scheme ε-good at accurately reproducing the input to the compressor by directly recycling the classical encoder and decoder.

10.1.2 Maintaining entanglement fidelity

But what about ψ_{SR} approximation? Suppose in the above that $\mathscr{C}$ comes from the following unitary action,

$$U_f|x\rangle_s|0\rangle_C = |x\rangle_S|f(x)\rangle_C, \tag{10.9}$$

and then tracing out S. Even if the compression scheme were perfect, we would have a problem with the fidelity between ψ_{SR} and the output ψ''_{SR}. The states involved are

$$|\psi\rangle_{SR} = \sum_x \sqrt{p_x}|x\rangle_S|\varphi_x\rangle_R \tag{10.10}$$

$$|\psi'\rangle_{SRC} = U|\psi\rangle_{SR}|0\rangle_C = \sum_x \sqrt{p_x}|x\rangle_S|\varphi_x\rangle_R|f(x)\rangle_C \tag{10.11}$$

$$|\psi''\rangle_{SRS'} = \sum_x \sqrt{p_x}|x\rangle_S|\varphi_x\rangle_R|x\rangle_{S'} \qquad \text{(after decompression)} \tag{10.12}$$

$$\psi''_{RS'} = \sum_x p_x|x\rangle\langle x|_{S'} \otimes |\varphi_x\rangle\langle\varphi_x|_R \tag{10.13}$$

Hence the fidelity is $\langle\psi|\psi''|\psi\rangle = \sum_x p_x^2$, which is only close to 1 if $|\psi\rangle$ is essentially unentangled.

As an aside, we can define the *entanglement fidelity* of a mapping $\mathscr{E}$ on an input ρ as follows:

$$F(\mathscr{E},\rho) := \max_{|\Phi_\rho\rangle} F(\Phi_\rho, \mathscr{E}\otimes\mathscr{I}(\Phi_\rho)), \tag{10.14}$$

where $|\Phi_\rho\rangle$ is a purification of ρ. We can find a nice expression for the entanglement fidelity which does not involve maximizations over the possible purifications. Suppose $\mathscr{E}(\rho) = \sum_k A_k\rho A_k^\dagger$ and consider the purification $|\Phi_\rho\rangle_{QR} = \sqrt{\rho_Q}|\Omega\rangle_{QR}$ for $|\Omega\rangle_{QR} = \sum_k |k\rangle_Q|k\rangle_R$, as in (5.20). We then have

$$\begin{aligned} F(\Phi_\rho, \mathscr{E}\otimes\mathscr{I}(\Phi_\rho))^2 &\geq \langle\Omega|\sqrt{\rho}_Q\sum_k A_k\sqrt{\rho}_Q|\Omega\rangle\langle\Omega|\sqrt{\rho}_Q A_k^\dagger\sqrt{\rho}_Q|\Omega\rangle && (10.15)\\ &= \sum_k |\langle\Omega|(\sqrt{\rho}A_k\sqrt{\rho}\otimes\mathbb{1})|\Omega\rangle|^2 && (10.16)\\ &= \sum_k |\mathrm{tr}[\rho A_k]|^2, && (10.17)\end{aligned}$$

since $\langle\Phi|A\otimes\mathbb{1}|\Phi\rangle = \mathrm{tr}[A]$. Moreover, we could have used any other purification, which is necessarily of the form $U_R|\Phi_{\sqrt{\rho}}\rangle_{QR}$ for some unitary U_R, as U_R would cancel out in the first expression.

Now let us be a little more clever about implementing the compressor. The preceding argument shows that we have to be careful to remove all traces of ψ_{SR} from the compressor degrees of freedom.

Instead of the U above, consider the following. First, define $\mathscr{X}_y = \{x : f(x) = y\}$. Then sort $\mathscr{X}_y$ according to $P_{X|Y=y}$. Call the result $\mathscr{X}_y^\downarrow$. Next define $g(x)$ to be the index or position of x in $\mathscr{X}_y^\downarrow$, counting from 0. The point is that $U = \sum_x |f(x)\rangle_C|g(x)\rangle_T\langle x|_S$ is an isometry, because $x \to (f(x), g(x))$ is reversible. For future use, call the inverse map h:

$$x \xleftarrow{h} (f(x), g(x)) \tag{10.18}$$

The compression scheme proceeds as follows. First, the compressor applies U and discards T. Then the decompressor applies $V = \sum_y |h(y,0)\rangle_{S'}\langle y|_C$. That is, the decompressor assumes $g(x) = 0$.

Using (10.17) we can compute the entanglement fidelity. The channel is $\mathscr{N}(\rho) = \mathrm{tr}_\mathrm{T}[\mathrm{VU}\rho\mathrm{U}^*\mathrm{V}^*]$ and so picks up Kraus operators from the partial trace: $A_z = {}_T\langle z|\otimes VU$. Then

$$\begin{aligned} F(\mathscr{N},\rho)^2 &= \sum_z |\mathrm{tr}A_z\rho|^2 && (10.19)\\ &= \sum_z |\mathrm{tr}({}_T\langle z|VU\sum_x p_x|x\rangle\langle x|_S)|^2 && (10.20)\\ &= \sum_z |\sum_x p_x {}_S\langle x|_T\langle z|VU|x\rangle_S|^2 && (10.21)\\ &= \sum_z |\sum_x p_x {}_S\langle x|_T\langle z|\sum_y |h(y,0)\rangle_S\langle y|_C \cdot |f(x)\rangle_C|g(x)\rangle_T|^2 && (10.22)\\ &= \sum_z |\sum_x p_x \sum_y \delta(x,h(y,0))\delta(z,g(x))\delta(y,f(x))|^2. && (10.23)\end{aligned}$$

Note that $\delta(x,h(y,0)) \Leftrightarrow \delta(f(x),y)\delta(g(x),0)$, so only $z=0$ contributes. Therefore,

$$F(\mathscr{N},\rho) = \sum_x p_x\delta(g(x),0) = \sum_{x\in\mathscr{X}_g} p_x, \tag{10.24}$$

since $\mathscr{X}_g = \{x : g(x) = 0\}$ by construction! So we get $F(\mathcal{N}, \rho) = 1 - p_{\text{err}}$.

We can now return to the issue of what the compressor is left with after the protocol and whether it is correlated with the source in any way. The state of RST after the execution of the protocol is

$$|\psi'\rangle_{RST} = \sum_x \sqrt{p_x} |\phi_x\rangle_R |g(x)\rangle_T |h(f(x), 0)\rangle_S. \tag{10.25}$$

Since this state is a purification of the output and we know the output is close to $|\psi\rangle_{RS}$, $|\psi'\rangle$ must be close to $|\psi_{RS}\rangle W|0\rangle_T$ in fidelity, for some isometry W (from Uhlmann's Theorem 5.2.5). This means that not was S compressed, but pure states were created in T in the process. The quantum scheme explicitly *erases* unneeded info in S. As we saw above, it has to!

Note that the two compression maps are different; $\text{tr}_T[U(\cdot)U^*]$ is not the same *equantum* channel as $\mathscr{C}$ used above. Their action is identical on inputs of the form $|x\rangle\langle x|$, but we must also consider "off-diagonal" inputs like $|x\rangle\langle x|'$. For the former we have

$$\text{tr}_T[U|x\rangle\langle x'|U^*] = |f(x)\rangle\langle f(x')|\delta(g(x), g(x')), \tag{10.26}$$

while the latter gives

$$\mathscr{C}(|x\rangle\langle x'|) = |f(x)\rangle\langle f(x')|\delta(x, x'). \tag{10.27}$$

10.2 Entanglement purification

Now let us examine a problem sort of "dual" to quantum data compression, namely entanglement purification. The goal here is to transform a given bipartite pure state $|\Psi\rangle_{AB} = \sum_x \sqrt{p_x} |\varphi_x\rangle_A \otimes |\xi_x\rangle_B$ (expressed here in the Schmidt basis) into an approximate version of $|\Phi_m\rangle = \frac{1}{\sqrt{2^m}} \sum_y |y\rangle_{A'} \otimes |y\rangle_{B'}$, for the largest m possible, using only local operations and classical communication.

As we saw in §10.1.2, data compression can be seen as a means of producing pure states from a given source, that is, an output which has entropy zero. In entanglement purification, on the other hand, the goal is to make the local states as mixed as possible (while still keeping the overall state pure). And, like quantum data compression, it turns out that there is an associated classical task that we can apply more or less directly to achieve the aims of entanglement purification. That task is *randomness extraction*.

10.2.1 Randomness Extraction

Given a random variable X, we would like to find a mapping to a new random variable Y with $\mathscr{Y} = \{0, 1\}^m$ such that $\delta(Y, U_m) \leq eps$, where U_m is the uniform distribution on m bits and δ is the statistical distance. The goal is to make m as large as possible for fixed ε.

Von Neumann gave a simple construction for binary i.i.d. random variables X^n, which goes as follows. Take two consecutive bits X_k and X_{k+1}. If the values of the two match, discard them, but if they differ, then output the first. Since the probabilities of 01 and 10 are identical, the output will be perfectly random. One may extend this construction to make use of discarded bits, but we will not go into that here.

Because the input random variable is not necessarily precisely specified, it is useful to look for universal randomness extraction schemes. That is, means of extracting the randomness of X without having to know too much about its distribution P_X. In computer science, especially cryptography, this has motivated the definition of an *extractor*. Formally, a (k, m, ϵ) seeded extractor E is a function from $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ such that for all random variables X of n bits having min entropy $H_{\min}(X) \geq k$ and U_t the uniformly distributed random variable on t bits,

$$\delta(E(X,U_d),U_m) \leq \epsilon. \tag{10.28}$$

The goal in coming up with new extractors is to make the seed as small as possible.

One might ask if the seed is really necessary. In order for the extractor to be universal, it is. Consider the case that X is a random variable of n bits with $H_{\min}(X) \geq n-1$ and we would like to extract just one bit. For any function $E: \{0,1\}^n \to \{0,1\}$, at least half of the inputs are mapped to one output value or the other. Thus, for X uniformly distributed on this set of inputs, $H_{\min}(X) \geq n-1$ but $E(X)$ is constant. Clearly what we need to do is choose a random E such that the probability of E being improperly matched to X in this way is small. That is where the seed comes in.

The reason the min entropy is involved is easy to understand: the best guess as to the output of the extractor is $f(x)$ for x having the maximum probability, i.e. $2^{-H_{\min}(X)}$. If this best guess probability is to be $\frac{1}{2^m}$, then m should equal $H_{\min}(X)$.

A *strong extractor* is a function $E_s : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ such that $(E_s(X),U_d)$ is ϵ-close to (U_m,U_d). That is, strong extractors produce uniform bits even when the seed is known.

Universal hashing can be used for randomness extraction, a statement which is formalized in the *Leftover Hash Lemma*.

10.2.2 The leftover hash lemma

The leftover hash lemma, due to Impagliazzo, Levin, and Luby, states that for a random variable X one can use universal hashing to extract $\ell = H_2(X) - 2\log(1/\varepsilon) + 2$ ε-good uniformly distributed bits, where H_2 is the *Renyi entropy* of order 2. Specifically, for a universal family $\mathcal{F}$ with $|\mathcal{F}| = 2^d$ of hash functions f, the function $E_f(X,U) = f_u(x)$ (use random variable U to choose the hash function f_u and output $f_u(x)$) is a strong extractor.

Proof. The proof proceeds by observing that the *collision probability* bounds the statistical distance to the uniform distribution and then calculating the collision probability of the output of the extractor.

The collision probability of a distribution is the probability of drawing the same result twice: $p_{\text{coll}}(X) = \sum_x p(x)^2$. Note that it is directly related to the $\alpha = 2$ Renyi entropy: $p_{\text{coll}}(X) = 2^{-H_2(X)}$. The collision probability also shows up in the 2-norm distance of P_X to the uniform distribution Q_X:

$$||P-Q||_2^2 = \sum_x (p(x) - 1/|X|)^2 = p_{\text{coll}}(X) - 1/|X|. \tag{10.29}$$

And the 2-norm distance in $\mathbb{R}^{|X|}$ bounds the statistical distance, $|X|||P-Q||_2^2 \geq ||P-Q||^2$ as shown in Section ??. Thus $||P-Q||^2 \leq |X|\, p_{\text{coll}}(X) - 1$.

What is the collision probability of the output random variable $(E_U(X),U)$?

$$\begin{aligned} p_{\text{coll}}(E_U(X),U) &= \sum_{x,x',f,f'} P[(f(x),f) = (f'(x),f')] && (10.30)\\ &= p_{\text{coll}}(F) P_{X,X',f}[f(x) = f(x')] && (10.31)\\ &= 2^{-d}(p_{\text{coll}}(X) + P[f(x) = f(x')|x \neq x']) && (10.32)\\ &= 2^{-d}(2^{-H_2(X)} + 2^{-m}). && (10.33)\end{aligned}$$

Therefore $||(E_U(X),U_d) - (U_m,U_d)||^2 \leq 2^{(m+d)}(2^{-H_2(X)} + 1/2^m)/2^d - 1 = 2^{m-H_2(X)}$. Choosing $m = H_2(X) - 2\log(1/\epsilon)$ then implies $||(E_U(X),U_d) - (U_m,U_d)|| \leq \epsilon$. □

10.2.3 Extractable randomness

Another closely related question of interest is the following. Given a probability distribution P_X, how many random bits can we extract from it by applying some (possibly seeded) extractor function E?

By the Leftover Hash Lemma (cf. Section 10.2.2), we can get at least roughly $H_2(X)$ bits. A bit of smoothing improves this amount. Suppose that P'_X is the distribution with largest min entropy in the ϵ neighborhood of P_X. Then $||P_X - P'_X|| \leq \epsilon$ and by the leftover hash lemma $||P_{E(X')} - U|| \leq \epsilon'$, where U is the uniform distribution on the same alphabet as $E(X)$, and E is the extractor constructed for P'_X. By the triangle inequality $||P_{E(X)} - U|| \leq \epsilon + \epsilon'$. Therefore, $H_2(X') - 2\log\epsilon'$ sets the number of bits we can get from X. But $H_2(X') \geq H_{\min}(X')$ and $H_{\min}(X') = H^{\epsilon}_{\min}(X)$, so we conclude that at least $H^{\epsilon}_{\min}(X) - 2\log\epsilon'$ $(\epsilon + \epsilon')$-random bits can be extracted from X.

On the other hand, the smooth min entropy also sets an upper bound. If m perfectly random bits can be obtained from X by E, then the most probable $E(X)$ must be $\frac{1}{2^m}$. The most probable $E(X)$ has at least the probability of the most probable X, $2^{-H_{\min}(X)}$. So $m \leq H_{\min}(X)$. By allowing a deviation of ϵ from the perfect case, so as to get ϵ-random bits, this upper bound increases to $m \leq H^{\epsilon}_{\min}(X)$ since with probability at least $1-\epsilon$ the output $E(X)$ is indistinguishable from $E(X')$.

10.2.4 From randomness extraction to entanglement purification

The basic idea of entanglement purification is to extract from a large amount of noisy (i.e., low fidelity) EPR pairs a smaller amount of EPR pairs having a sufficiently high fidelity. In order to do so we consider the following bipartite state written in the Schmidt basis

$$|\psi\rangle_{AB} = \sum_x \sqrt{p_x}|\varphi_x\rangle_A|\xi_x\rangle_B, \tag{10.34}$$

which, using LOCC, can be transform into m-ebits of the form

$$|\phi\rangle_{A'B'} = \frac{1}{\sqrt{2^m}}\sum_y |\eta_y\rangle_{A'}|\eta_y\rangle_{B'}. \tag{10.35}$$

For this transformation we use randomness extraction adapted to a particular X form $|\psi\rangle$. Returning to (10.28), note that it can be interpreted as the average over the choice of seed:

$$\delta(E(X,U_d),U_m) = \frac{1}{2^d}\sum_u \delta(E(X,u),U_m). \tag{10.36}$$

Therefore, for any extractor applied to a specific X, there is an optimal seed choice for which the statistical distance is lowest (and necessarily lower than the average ε). Let f denote this function. Now consider the invertible map $X \to (f(x),x)$ such that

$$V|\varphi_x\rangle_A = |f(x)\rangle_{A'}|\varphi_x\rangle_A. \tag{10.37}$$

Applied to the joint state, we thus obtain

$$|\psi'\rangle_{A'AB} = V_A|\psi\rangle_{AB} = \sum_x \sqrt{p_x}|f(x)\rangle_{A'} \otimes |\varphi_x\rangle_A \otimes |\xi_x\rangle_B. \tag{10.38}$$

Now consider the A' marginal state. Since the $|\xi_x\rangle$ are an orthonormal basis, we find immediately

$$\psi'_{A'} = \sum_x p(x)|f(x)\rangle\langle f(x)| = \sum_y |y\rangle\langle y| \sum_{x\in\mathscr{X}_y} p_x. \tag{10.39}$$

The probability of any given basis state $|y\rangle$ is just $\sum_{x\in\mathscr{X}_y} p_x$, which is precisely that of the output of the function f, so

$$\delta(\psi'_{A'}, \tfrac{1}{2^m}\mathrm{id}_{A'}) \le \varepsilon. \tag{10.40}$$

By Lemma ??, this implies $F(\psi'_{A'}, \frac{1}{2^m}\mathrm{id}_{A'}) \ge 1-\varepsilon$. We can then determine the fidelity for purifications of these two states by Uhlmann's theorem 5.2.5. Clearly, one possible purification of $\psi'_{A'}$ is simply $|\psi'\rangle_{A'AB}$ itself. A possible purification of $\frac{1}{2^m}\mathrm{id}_{A'}$ is $|\Phi_m\rangle_{A'B'}$. To this we are also free to append any pure state we like, for instance $|\psi\rangle_{\bar{B}B}$, which is just the original state, but the A system replaced by $\bar{B}$. Then, by Uhlmann's theorem, there must exist an isometry $W_{AB\to B'\bar{B}B}$ such that

$${}_{A'B'}\langle\Phi_m|\langle\psi|_{\bar{B}B}\, W_{AB\to B'\bar{B}B}|\psi'\rangle_{A'AB} \ge 1-\varepsilon. \tag{10.41}$$

Thus, knowing that the A' system is maximally mixed allows us to infer the existence of a unitary which completes our entanglement purification protocol. This trick of using Uhlmann's theorem is quite widespread in quantum information theory, and goes under the name *decoupling* (since A' is decoupled from everything else).

However, this does not yet yield an LOCC protocol, because W might require joint operations on A and B. To show that there is an LOCC version of W, go back to (10.38) and observe that if the A system were not present, the above argument would nonetheless proceed exactly as before, except that W would only involve system B. We can then get rid of A by the following method. First, define a conjugate basis to $|\varphi_x\rangle$ by

$$|\vartheta_z\rangle = \frac{1}{\sqrt{d}}\sum_{x=0}^{d-1}\omega^{xz}|\varphi_x\rangle, \tag{10.42}$$

where $\omega = e^{2\pi I/d}$ and d is the dimension of the system A. Measuring A in this basis gives the state

$${}_A\langle\vartheta_z|\psi'\rangle_{A'AB} = \frac{1}{\sqrt{d}}\sum_x \sqrt{p_x}\,\omega^{-xz}|f(x)\rangle_{A'}|\xi_x\rangle_B, \tag{10.43}$$

which is unnormalized. The normalization is the probability of the outcome z; note that these probabilities are all equal. Now, if Alice sends the outcome z to Bob, he can simply apply the operation

$$R^z = \sum_x \omega^{xz}|\xi_x\rangle\langle\xi_x| \tag{10.44}$$

and this will produce the state (now normalized)

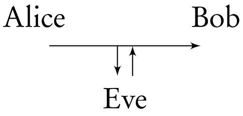
$$|\psi''\rangle_{A'B} = \sum_x \sqrt{p_x}|f(x)\rangle_{A'}|\xi_x\rangle_B. \tag{10.45}$$

Applying the above decoupling argument to this state give an LOCC entanglement purification protocol. The number of ebits produced is just the number of random bits that can be extracted from X distributed according to the Schmidt coefficients of the state. Note also that Bob has ended up with the original state at his side, even though Alice only sent classical information. This is the protocol of *state merging*.

11 Quantum Key Distribution

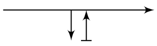
11.1 Introduction

In this chapter, we introduce the concept of quantum key distribution. To start, we introduce the concept of cryptographic resources. A classical *insecure* communication channel is denoted by



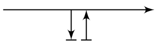
The arrows to the adversary, Eve, indicate that she can receive all messages sent by Alice. Furthermore Eve is able to modify the message which Bob finally receives. This channel does not provide any guarantees. It can be used to model for example email traffic.

A classical *authentic* channel is denoted by

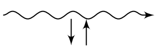


and guarantees that messages received by Bob are sent by Alice. It can be used to describe e.g. a telephone conversation with voice authentification.

The most restrictive classical channel model we consider is the so-called *secure* channel which has the same guarantees as the authentic channel and ensures in addition that no information leaks. It is denoted by

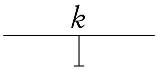


In the quantum setup, an *insecure quantum* channel that has no guarantees is represented by



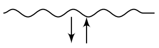
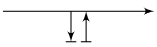
Note that an authentic quantum channel is automatically also a secure quantum channel since reading out a message always changes the message.

The following symbol

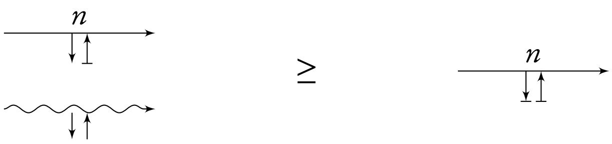


denotes k classical secret bits, i.e. k bits that are uniformly distributed and maximally correlated between Alice and Bob.

A desirable goal of quantum cryptography would be to have a protocol that simulates a secure classical channel using an insecure quantum channel, i.e.,

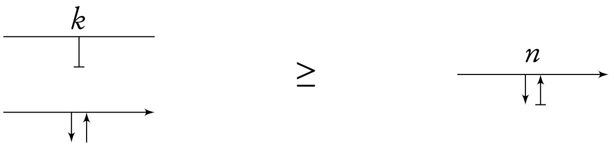
 $\overset{?}{\geq}$ 

However, such a protocol cannot exist since this scenario has complete symmetry between Alice and Eve, which makes it impossible for Bob to distinguish between them. If we add a classical authentic channel in addition to the insecure quantum channel, it is possible as we shall see to simulate a classical secret channel, i.e.,



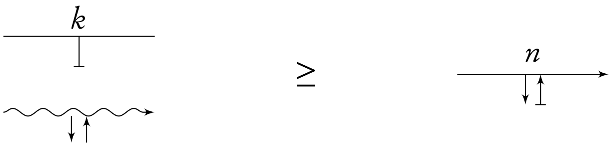
is possible.

In classical cryptography there exists a protocol [32], called *authentication*, that achieves the following



$n \gg k$.

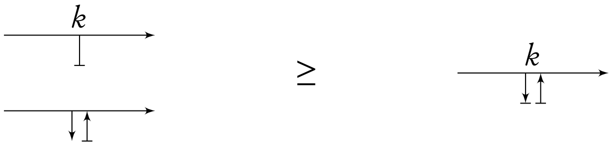
Thus, if Alice and Bob have a (short) password, they can use an insecure channel to simulate an authentic channel. This implies



$n \gg k$.

11.2 Classical message encryption

The *one-time pad* protocol achieves the following using purely classical technology.



Let M be a message bit and S a secret key bit. The operation $\oplus$ denotes an addition modulo 2. Alice first computes $C = M \oplus S$ and sends C over a classical authentic channel to Bob. Bob then computes $M' = C \oplus S$. The protocol is correct as

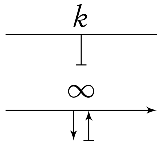
$$M' = C \oplus S = (M \oplus S) \oplus S = M \oplus (S \oplus S) = M.$$

To prove secrecy of the protocol, we have to show that $P_M = P_{M|C}$ which is equivalent to $P_{MC} = P_M \times P_C$. In information theoretic terms this condition can be expressed as $I(M:C) = 0$ which means that the bit C which is sent to Bob and may be accessible to Eve does not have any information about the message bit M. This follows from the observation that $P_{C|M=m}$ is uniform for all $m \in \{0,1\}$. Therefore, $P_{C|M} = P_C$ which is equivalent to $P_{CM} = P_C \times P_M$ and thus proves that the protocol is secret.

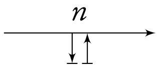
As the name *one-time pad* suggests, a secret bit can only be used once. For example consider the scenario where someone uses a single secret bit to encrypt 7 message bits such that we have e.g. $C = 0010011$. Eve then knows that $M = 0010011$ or $M = 1101100$.

Shannon proved in 1949 that in a classical scenario, to have a secure protocol the key must be as long as the message [33], i.e.

Theorem 11.2.1.



$\geq$



implies $k \geq n$.

Proof. Let $M \in \{0,1\}^n$ be the message which should be sent secretly from Alice to Bob. Alice and Bob share a secret key $S \in \{0,1\}^k$. Alice first encrypts the message M and sends a string C over a public channel to Bob. Bob decrypts the message, i.e. he computes a string M' out of C and his key S. We assume that the protocol fulfills the following two requirements.

1. *Reliability*: Bob should be able to reproduce M (i.e. $M' = M$).

2. *Secrecy*: Eve does not gain information about M.

We consider a message that is uniformly distributed on $\{0,1\}^n$. The secrecy requirement can be written as $I(M:C) = 0$ which implies that $H(M|C) = H(M) = n$. We thus obtain

$$I(M:S|C) = H(M|C) - H(M|CS) = n, \tag{11.1}$$

where we also used the reliability requirement $H(M|CS) = 0$ in the last equality. Using the data processing inequality and the non negativity of the Shannon entropy, we can write

$$I(M:S|C) = H(S|C) - H(S|CM) \leq H(S). \tag{11.2}$$

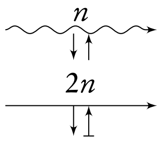
Combining (11.1) and (11.2) gives $n \leq H(S)$ which implies that $k \geq n$. □

Shannon's result shows that information theoretic secrecy (i.e. $I(M:C) \approx 0$) cannot be achieved unless one uses very long keys (as long as the message).

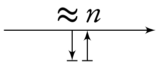
In *computational cryptography*, one relaxes the security criterion. More precisely, the mutual information $I(M:C)$ is no longer small, but it is still computationally hard (i.e. it takes a lot of time) to compute M from C. In other words, we no longer have the requirement that $H(M|C)$ is large. In fact, for public key cryptosystems (such as RSA and DH), we have $H(M|C) = 0$. This implies that there exists a function f such that $M = f(C)$, which means that it is in principle possible to compute M from C. Security is obtained because f is believedto be hard to compute. Note, however, that for the protocol to be practical, one requires that there exists an efficiently computable function g, such that $M = g(C,S)$.

11.3 Quantum cryptography

In this section, we explain why Theorem 11.2.1 does not hold in the quantum setup. As we will prove later, having a quantum channel we can achieve



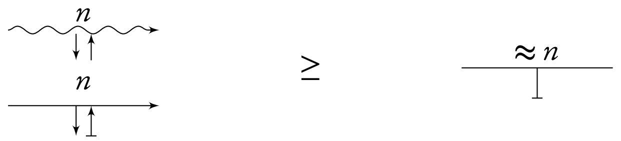
$\geq$



(□)

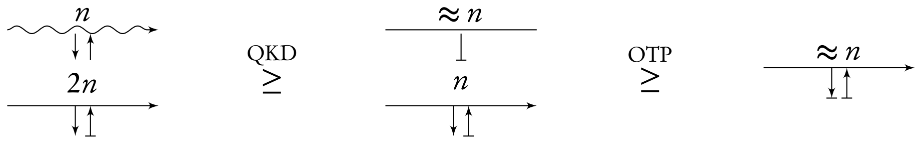
Note that this does not contradict Shannon's proof of Theorem 11.2.1, since in the quantum regime the no-cloning theorem (cf. Section ??) forbids that Bob and Eve receive the same state, i.e., the ciphertext C is not generally available to both of them. Therefore, Shannon's proof is not valid in the quantum setup, which allows quantum cryptography to go beyond classical cryptography.

As we will see in the following, it is sufficient to consider *quantum key distribution* (QKD), which does the following.



(△)

The protocol (△) implies (□) as we can concatenate it with the one-time pad encryption. More precisely,



which is the justification that we can focus on the task of QKD in the following.

We next define more precisely what we mean by a secret key, as denoted by S_A and S_B. In quantum cryptography, we generally consider the following three requirements where $\epsilon \geq 0$

1. Correctness: $\Pr[S_A \neq S_B] \leq \epsilon$
2. Robustness: if the adversary is passive, then[1] $\Pr[S_A = \perp] \leq \epsilon$
3. Secrecy: $\|\rho_{S_A E} - (p\rho_\perp \otimes \rho_{E_\perp} + (1-p)\rho_k \otimes \rho_{E_k})\|_1 \leq \epsilon$, where $\rho_{E_\perp}, \rho_{E_k}$ are arbitrary density operators, $\rho_\perp = |\perp\rangle\langle\perp|$ and ρ_k is a completely mixed state on $\{0,1\}^n$, i.e. $\rho_k = 2^{-n}\sum_{s\in\{0,1\}^n}|s\rangle\langle s|$. The cq-state $\rho_{S_A E}$ describes the key S_A together with the system E held by the adversary after the protocol execution. The parameter p can be viewed as the failure probability of the protocol.

The secrecy condition implies that there is either a uniform and uncorrelated (to E) key or there is no key at all.

11.4 QKD protocols

11.4.1 BB84 protocol

In the seventies, Wiesner had the idea to construct unforgeable money based on the fact that quantum states cannot be cloned [34]. However, the technology at that time was not ready to start up on his idea. In 1984, Bennett and Brassard presented the *BB84 protocol* for QKD [35] which is based on Wiesner's ideas and will be explained next.

In the BB84 protocol, Alice and Bob want to generate a secret key which is achieved in four steps. In the following, we choose a standard basis $\{|0\rangle, |1\rangle\}$ and $\{|\bar{0}\rangle, |\bar{1}\rangle\}$ where $|\bar{0}\rangle := \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|\bar{1}\rangle := \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$.

[1] The symbol $\perp$ indicates that no key has been produced.

BB84 Protocol:

Distribution step Alice and Bob perform the following task N times and let $i = 1,\dots,N$. Alice first chooses $B_i, X_i \in \{0,1\}$ at random and prepares a state of a qubit Q_i (with basis $\{|0\rangle,|1\rangle\}$) according to

B	X	Q
0	0	$\lvert 0\rangle$
0	1	$\lvert 1\rangle$
1	0	$\lvert \bar{0}\rangle$
1	1	$\lvert \bar{1}\rangle$.

Alice then sends Q_i to Bob.

Bob next chooses $B_i' \in \{0,1\}$ and measures Q_i either in basis $\{|0\rangle,|1\rangle\}$ (if $B_i' = 0$) or in basis $\{|\bar{0}\rangle,|\bar{1}\rangle\}$ (if $B_i' = 1$) and stores the result in X_i. Recall that all the steps so far are repeated N-times.

Sifting step Alice sends $B_1,\dots,B_n$ to Bob and vice versa, using the classical authentic channel. Bob discards all outcomes for which $B_i \neq B_i'$ and Alice does so as well. For better understanding we consider the following example situation.

Q	$\lvert 1\rangle$	$\lvert 1\rangle$	$\lvert \bar{1}\rangle$	$\lvert \bar{0}\rangle$	$\lvert 0\rangle$	$\lvert \bar{1}\rangle$	$\lvert \bar{0}\rangle$	$\lvert 1\rangle$	$\lvert \bar{1}\rangle$
B	0	0	1	1	0	1	1	0	1
X	1	1	1	0	0	1	0	1	1
B'	0	0	0	1	1	0	1	1	0
X'	1	1	1	0	1	1	0	0	1
no.	①	②	③	④	⑤	⑥	⑦	⑧	⑨

Hence, Alice and Bob discard columns ③,⑤,⑥,⑧ and ⑨.

Checking step Alice and Bob compare (via communication over the classical *authentic* channel) X_i and X_i' for some randomly chosen sample i of size $\sqrt{n}$. If there is disagreement, the protocol aborts, i.e. $S_A = S_B = \perp$.

Extraction step We consider here the simplest case where we assume to have no errors (due to noise). The key S_A is equal to the remaining bits of $X_1,\dots,X_n$ and the key S_B are the remaining bits of $X_1',\dots,X_n'$. Note that the protocol can be generalized such that it also works in the presence of noise.

11.4.2 Security proof of BB84

It took almost 20 years until the security of BB84 could be proven [36, 37, 38, 39]. We present in the following a proof sketch. The idea is to first consider an entanglement-based protocol (called Ekert91 [40]) and prove that this protocol is equivalent to the BB84 protocol in terms of secrecy. Therefore, it is sufficient to prove security of the Ekert91 protocol which turns out to be easier to achieve. For this, we will use a *generalized uncertainty relation* which states that

$$H(Z|E) + H(X|B) \geq 1, \tag{11.3}$$

where Z denotes a measurement in the basis $\{|0\rangle,|1\rangle\}$, X denotes a measurement in the basis $\{|\bar{0}\rangle,|\bar{1}\rangle\}$ and where B and E are arbitrary quantum systems.

Ekert91 protocol: Similarly to the BB84 protocol this scheme also consists of four different steps.

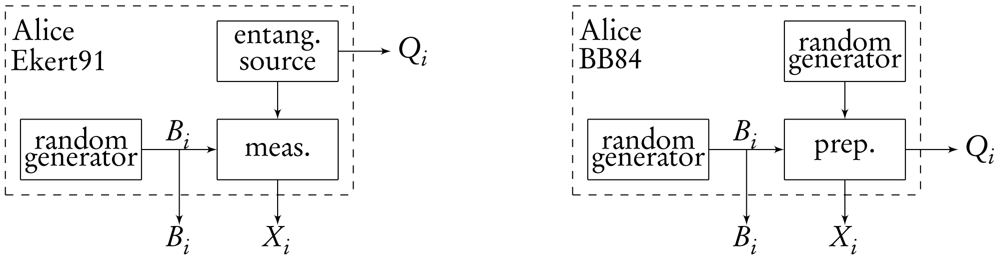
Distribution step (repeated N times) Alice prepares entangled qubit pairs and sends one half of each pair to Bob (over the insecure quantum channel). Alice and Bob then measure their qubit in a random basis B_i (for Alice)[2] and B_i' (for Bob). They report the outcomes X_i (for Alice) and X_i' (for Bop).

Sifting step Alice and Bob discard all (X_i, X_i') for which $B_i \neq B_i'$.

Checking step For a random sample of positions i Alice and Bob check whether $X_i = X_i'$. If the test fails they abort the protocol by outputting $\perp$.

Extracting step Alice's key S_A consists of the remaining bits of $X_1, X_2, \ldots$. Bob's key S_B consists of the remaining bits $X_1', X_2', \ldots$.

We next show that Ekert91 is equivalent to BB84. On Bob's side it is easy to verify that the two protocols are equivalent since Bob has to perform exactly the same tasks for both. The following schematic figure summarizes Alice's task in the Ekert91 and the BB84 protocol.



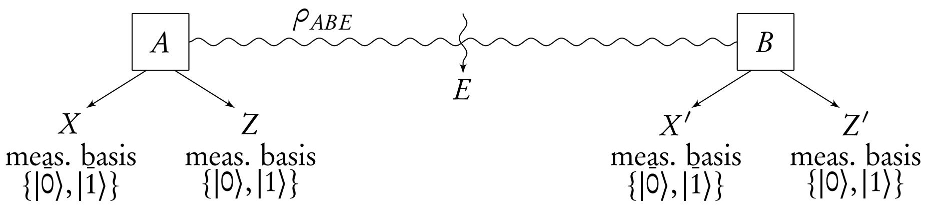
In the Ekert91 protocol Alice's task is described by a state

$$\rho^{\text{Ekert91}}_{B_i X_i Q_i} = \frac{1}{2} \sum_{b \in \{0,1\}} |b\rangle\langle b|_{B_i} \otimes \frac{1}{2} \sum_{x \in \{0,1\}} |x\rangle\langle x|_{X_i} \otimes |\varphi^{b,x}\rangle\langle \varphi^{b,x}|_{Q_i}, \tag{11.4}$$

where $|\varphi^{0,0}\rangle = |0\rangle$, $|\varphi^{0,1}\rangle = |1\rangle$, $|\varphi^{1,0}\rangle = |\bar{0}\rangle$, and $|\varphi^{1,1}\rangle = |\bar{1}\rangle$. The BB84 protocol leads to the same state

$$\rho^{\text{BB84}}_{B_i X_i Q_i} = \frac{1}{2} \sum_{b \in \{0,1\}} |b\rangle\langle b|_{B_i} \otimes \frac{1}{2} \sum_{x \in \{0,1\}} |x\rangle\langle x|_{X_i} \otimes |\varphi^{b,x}\rangle\langle \varphi^{b,x}|_{Q_i}. \tag{11.5}$$

We thus conclude that viewed from outside the dashed box the two protocols are equivalent in terms of security and hence to prove security for BB84 it is sufficient to prove security for Ekert91. Note that both protocols have some advantages and drawbacks. While for Ekert91 it is easier to prove security, the BB84 protocol is technologically simpler to implement.


A
ρ_{ABE}
B
E
X
meas. basis
$\{|0\rangle,|1\rangle\}$
Z
meas. basis
$\{|0\rangle,|1\rangle\}$
X'
meas. basis
$\{|0\rangle,|1\rangle\}$
Z'
meas. basis
$\{|0\rangle,|1\rangle\}$

[2] Recall that $B_i = 0$ means that we measure in the $\{|0\rangle, |1\rangle\}$ basis and if $B_i = 1$ we measure in the $\{|\bar{0}\rangle, |\bar{1}\rangle\}$ basis.

It remains to prove that the Ekert91 protocol is secure. The idea is to consider the state of the entire system (i.e. Alice, Bob and Eve) after the sending the distribution of the entangled qubit pairs over the insecure channel (which may be arbitrarily modified by Eve) but before Alice and Bob have measured. The state ρ_{ABE} is arbitrary except that the subsystem A is a fully mixed state (i.e. ρ_A is maximally mixed). At this point the completeness of quantum theory (cf. Chapter **??**) shows up again. Since quantum theory is complete, we know that anything Eve could possibly do is described within our framework.

We now consider two alternative measurements for Alice ($B=0, B=1$). Call the outcome of the measurement Z if $B=0$ and X if $B=1$. The uncertainty relation (11.3) now implies that

$$H(Z|E)+H(X|B) \geq 1, \tag{11.6}$$

which holds for arbitrary states ρ_{ABE} where the first term is evaluated for ρ_{ZBE} and the second term is evaluated for ρ_{XBE}. The state ρ_{XBE} is defined as the post-measurement state when measuring ρ_{ABE} in the basis $\{|\bar{0}\rangle, |\bar{1}\rangle\}$ and the sate ρ_{ZBE} is defined as the post-measurement state when measuring ρ_{ABE} in the basis $\{|0\rangle, |1\rangle\}$. Using (11.6), we can bound Eve's information as $H(Z|E) \geq 1 - H(X|B)$. We next show that $H(X|B)=0$ which implies that $H(Z|E)=1$, i.e. Eve has no information about Alice's state. The data processing inequality implies $H(Z|E) \geq 1 - H(X|X')$.

In the protocol, there is a step called the testing phase where two alternative things can happen

- if $\Pr[X \neq X'] > 0$, then Alice and Bob detect a deviation in their sample and abort the protocol.
- if $\Pr[X = X'] \approx 1$, Alice and Bob output a key.

Let us therefore assume that $\Pr[X \neq X'] = \delta$ for $\delta \approx 0$. In this case, we have $H(Z|E) \geq 1 - h(\delta) \approx 1 - \sqrt{\delta}$ for small δ, where $h(\delta) := -\delta \log_2 \delta - (1-\delta)\log_2(1-\delta)$ denotes the *binary entropy function*. Note that also $H(Z) = 1$, which implies that $I(Z:E) = H(Z) - H(Z|E) \leq h(\delta)$. Recall that $I(Z:E) = D(\rho_{ZE} \| \rho_Z \otimes \rho_E)$. Thus, if $I(Z:E)=0$, we have $D(\rho_{ZE} \| \rho_Z \otimes \rho_E) = 0$ for $\delta \to 0$. This implies that $\rho_{ZE} = \rho_Z \otimes \rho_E$ which completes the security proof.[3]

Important remarks to the security proof The proof given above establishes security under the assumption that there are no correlations between the rounds of the protocol. Note that if the state involved in the i-th round is described by $\rho_{A_iB_iE_i}$ we have in general

$$\rho_{A_1A_2\ldots A_nB_1B_2\ldots B_nE_1E_2\ldots B_n} \neq \rho_{A_1B_1E_1} \otimes \rho_{A_2B_2E_2} \otimes \ldots \otimes \rho_{A_nB_nE_n}. \tag{11.7}$$

Therefore, it is not sufficient to analyze the rounds individually and hence we so far only proved security against i.i.d. attacks, but not against general attacks. Fortunately, there is a very solution to this problem: The post-selection technique of [41] shows that the proof for individual attacks also implies security for general attacks.

[3] In principle, we have to repeat the whole argument in the complementary basis, i.e. using the uncertainty relation $H(X|E)+H(Z|B) \geq 1$.

A

Mathematical background

A.1 Hilbert spaces and operators on them

Consider a vector space $\mathcal{H}$, for concreteness over the the field of complex numbers $\mathbb{C}$. An *inner product* on $\mathcal{H}$ is a bilinear function $(\cdot,\cdot) : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ with the properties that (i) $(v, v') = (v', v)^*$ where $*$ denotes the complex conjugate, (ii) $(v, \alpha v') = \alpha(v, v')$ for $\alpha \in \mathbb{C}$ and $(v, v' + v'') = (v, v') + (v, v'')$, and (iii) $(v, v) \geq 0$. (Note that the inner product is usually taken to be linear in the first argument in mathematics literature, not the second as here.) The inner product induces a *norm* on the vector space, a function $\|\cdot\| : \mathcal{H} \to \mathbb{C}$ defined by $\|v\| := \sqrt{(v, v)}$. A vector space with an inner product is called an *inner product space*. If it is *complete* in the metric defined by the norm, meaning all Cauchy[1] sequences converge, it is called a *Hilbert*[2] *space* We will restrict attention to finite-dimensional spaces, where the completeness condition always holds and inner product spaces are equivalent to Hilbert spaces.

We denote the set of *homomorphisms* (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 *endomorphisms* (the homomorphisms from a space to itself) on $\mathcal{H}$: $\mathrm{End}(\mathcal{H}) = \mathrm{Hom}(\mathcal{H}, \mathcal{H})$. The identity operator $v \mapsto v$ that maps any vector $v \in \mathcal{H}$ 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

$$(v', Sv) = (S^* v', v), \tag{A.1}$$

for any $v \in \mathcal{H}$ and $v' \in \mathcal{H}'$. 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.

Here we list some properties of endomorphisms $S \in \mathrm{End}(\mathcal{H})$:

- S is *normal* if $SS^* = S^*S$, *unitary* if $SS^* = S^*S = \mathrm{id}$, and *self-adjoint* (or *Hermitian*[3]) if $S^* = S$.

- S is *positive* if $(v, Sv) \geq 0$ for all $v \in \mathcal{H}$. Positive operators are always self-adjoint. 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 $\{b_i\}_i$ of $\mathcal{H}$, we also say that S is *diagonal with respect to* $\{b_i\}_i$ if the matrix $(S_{i,j})$ defined by the elements $S_{i,j} = (b_i, Sb_j)$ is diagonal.

A map $U \in \mathrm{Hom}(\mathcal{H}, \mathcal{H}')$ with $\dim(\mathcal{H}') \geq \dim(\mathcal{H})$ will be called an *isometry* if $U^*U = \mathrm{id}_{\mathcal{H}}$. It can be understood as an embedding of $\mathcal{H}$ into $\mathcal{H}'$, since all inner products between vectors are preserved: $(\phi', \psi') = (U\phi, U\psi) = (\phi, U^*U\psi) = (\phi, \psi)$.

A.2 The bra-ket notation

In this script we will make extensive use of a variant of Dirac's[4] *bra-ket notation*, where vectors are interpreted as operators. More precisely, we can associate any vector $v \in \mathcal{H}$ with an endomorphism

[1] Augustin-Louis Cauchy, 1789 – 1857, French mathematician.

[2] David Hilbert, 1862 – 1943, German mathematician.

[3] Charles Hermite, 1822 – 1901, French mathematician.

[4] Paul Adrien Maurice Dirac, 1902 – 1984, English physicist.

$|v\rangle \in \mathrm{Hom}(\mathbb{C},\mathscr{H})$, called *ket* and defined as

$$|v\rangle : \quad \gamma \mapsto \gamma v, \tag{A.2}$$

for any $\gamma \in \mathbb{C}$. We will often regard $|v\rangle$ as the vector itself, a misuse of notation which enables a lot of simplification. The adjoint $|v\rangle^*$ of this mapping is called *bra* and denoted by $\langle v|$. It is easy to see that $\langle v|$ is an element of the *dual space* $\mathscr{H}^* := \mathrm{Hom}(\mathscr{H},\mathbb{C})$, namely the linear functional defined by

$$\langle v| : \quad u \mapsto (v,u), \tag{A.3}$$

for any $u \in \mathscr{H}$. Note, however, that bras and kets are not quite on equal footing, as the label of a bra is an element of $\mathscr{H}$, not $\mathscr{H}^*$. The reason we can do this is the *Riesz*[5] *representation theorem*, which states that every element of the dual space is of the form given in (A.3).

Using this notation, the concatenation $\langle u| \circ |v\rangle$ of a bra $\langle u| \in \mathrm{Hom}(\mathscr{H},\mathbb{C})$ with a ket $|v\rangle \in \mathrm{Hom}(\mathbb{C},\mathscr{H})$ 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 $u, v \in \mathscr{H}$,

$$\langle u| \circ |v\rangle \equiv (u,v). \tag{A.4}$$

Thus, in the following we will omit the $\circ$ and denote the scalar product by $\langle u|v\rangle$.

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

$$S = \sum_i |u_i\rangle\langle v_i| \tag{A.5}$$

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

$$\mathrm{id} = \sum_i |b_i\rangle\langle b_i| \tag{A.6}$$

for any orthonormal basis $\{b_i\}$ of $\mathscr{H}$. This is often called the *completeness relation* of the basis vectors.

A.3 Representations of operators by matrices

Given an orthonormalbasis $\{|b_k\rangle\}_{k=1}^d$, we can associate a matrix with any operator $S \in \mathrm{End}(\mathscr{H})$,

$$S \to S_{jk} = \langle b_j|S|b_k\rangle. \tag{A.7}$$

Here we are "overloading" the notation a bit, and referring to both the matrix components as well as the matrix itself as S_{jk}. In the study of relativity, this is referred to as *abstract index notation* or *slot-naming index notation*. We have chosen j to be the row index and k the column index, so that a product of operators like ST corresponds to the product of the corresponding matrices, but the other choice could have been made.

[5] Frigyes Riesz, 1880 – 1956, Hungarian mathematician.

It is important to realize that the representation of an operator by a matrix is not unique, but depends on the choice of basis. One way to see this is to use the completeness relation, equation (A.6), to write

$$\begin{aligned} S &= \mathrm{id}\, S\, \mathrm{id} && \text{(A.8)}\\ &= \sum_{j,k} |b_j\rangle\langle b_j|S|b_k\rangle\langle b_k| && \text{(A.9)}\\ &= \sum_{j,k} S_{j,k}|b_j\rangle\langle b_k|. && \text{(A.10)} \end{aligned}$$

Now the basis dependence is plain to see. Matrix representations can be given for more general operators $S \in \mathrm{Hom}(\mathcal{H},\mathcal{H}')$ by the same technique:

$$\begin{aligned} S &= \mathrm{id}_{\mathcal{H}'}\, S\, \mathrm{id}_{\mathcal{H}} && \text{(A.11)}\\ &= \sum_{j,k} |b'_j\rangle\langle b'_j|S|b_k\rangle\langle b_k| && \text{(A.12)}\\ &= \sum_{j,k} S_{j,k}|b'_j\rangle\langle b_k|. && \text{(A.13)} \end{aligned}$$

In our version of Dirac notation, $|v\rangle$ is itself an operator, so we can apply the above method to this case. Now, however, the input space is one-dimensional, so we drop the associated basis vector and simply write

$$|v\rangle = \sum_j v_j|b_j\rangle. \qquad \text{(A.14)}$$

According to the above convention, the representation of $|v\rangle$ is automatically a column vector, as it is the column index (which would take only one value) that has been omitted. Following our use of abstract index notation, the (vector) representative of $|v\rangle$ is called v_j, not $\vec{v}$ or similar.

In terms of matrix representatives, the inner product of two vectors u and v is given by $u_j^* \cdot v_j$, since the inner product is linear in the second argument, but antilinear in the first. We expect the representation of the adjoint of an operator to be the conjugate transpose of the matrix, but let us verify that this is indeed the case. The defining property of the adjoint is (A.1), or in Dirac notation

$$\langle u|Sv\rangle = \langle S^* u|v\rangle. \qquad \text{(A.15)}$$

In terms of matrix representatives, reading the above from right to left we have

$$\begin{aligned} (S^*u)_j^* \cdot v_j &= u_j^* \cdot (Sv)_j && \text{(A.16)}\\ &= \sum_{jk} u_j^* S_{jk} v_k && \text{(A.17)}\\ &= \sum_{jk} ([S_{jk}]^* u_j)^* v_k && \text{(A.18)}\\ &= \sum_{jk} ([S_{jk}]^\dagger u_k)^* v_j. && \text{(A.19)} \end{aligned}$$

Here † denotes the conjugate transpose of a matrix. Comparing the first expression with the last, it must be that $[S^*]_{jk} = [S_{jk}]^\dagger$, as we suspected.

A.4 Tensor products

Given vectors u and v from two Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, we may formally define their product $u \times v$, which is an element of the *Cartesian*[6] *product* $\mathcal{H}_A \times \mathcal{H}_B$. However, the Cartesian product does not respect the linearity of the underlying spaces. That is, while we may formally add $u \times v$ and $u' \times v$, the result is not $(u + u') \times v$; it is just $u \times v + u' \times v$. The idea behind the *tensor product* is to enforce this sort of linearity on $\mathcal{H}_A \times \mathcal{H}_B$. There are four combinations of vectors which we would expect to vanish by linearity:

$$\begin{aligned} &u \times v + u' \times v - (u + u') \times v, \\ &u \times v + u \times v' - u \times (v + v'), \\ &\alpha(u \times v) - (\alpha u) \times v, \\ &\alpha(u \times v) - u \times (\alpha v), \end{aligned} \tag{A.20}$$

for any $\alpha \in \mathbb{C}$. These vectors define an equivalence relation on $\mathcal{H}_A \times \mathcal{H}_B$ in that we can consider two elements of that space to be equivalent if they differ by some vector of the form in (A.20). These equivalence classes themselves form a vector space, and the resulting vector space is precisely the tensor product $\mathcal{H}_A \otimes \mathcal{H}_B$.

Since the construction enforces linearity of the products of vectors, we may consider the tensor product to be the space spanned by products of basis elements of each space. Furthermore, the inner product of $\mathcal{H}_A \otimes \mathcal{H}_B$ is defined by the linear extension of

$$(u \otimes v, u' \otimes v') = \langle u|u'\rangle\langle v|v'\rangle. \tag{A.21}$$

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

$$(S \otimes T)(u \otimes v) := (Su) \otimes (Tv) \tag{A.22}$$

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

$$\mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_A') \otimes \mathrm{Hom}(\mathcal{H}_B, \mathcal{H}_B') \simeq \mathrm{Hom}(\mathcal{H}_A \otimes \mathcal{H}_B, \mathcal{H}_A' \otimes \mathcal{H}_B'). \tag{A.23}$$

That is, the mapping defined by (A.22) is an isomorphism between these two vector spaces. This identification allows us to write, for instance,

$$|u\rangle \otimes |v\rangle = |u \otimes v\rangle, \tag{A.24}$$

for any $u \in \mathcal{H}_A$ and $v \in \mathcal{H}_B$.

A.5 Trace and partial trace

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

$$\mathrm{tr}(S) := \sum_i \langle b_i|S|b_i\rangle, \tag{A.25}$$

[6] René Descartes, 1596 – 1650, French philosopher and mathematician.

where $\{|b_i\rangle\}_i$ 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 is obviously linear,

$$\mathrm{tr}(\alpha S + \beta T) = \alpha \mathrm{tr}(S) + \beta \mathrm{tr}(T), \tag{A.26}$$

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

$$\mathrm{tr}(S^*) = \mathrm{tr}(S)^*, \tag{A.27}$$

since the adjoint of a complex number $\gamma \in \mathbb{C}$ is simply its complex conjugate. Furthermore, the trace is cyclic,

$$\mathrm{tr}(ST) = \mathrm{tr}(TS). \tag{A.28}$$

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

$$(S \geq 0) \wedge (T \geq 0) \implies \mathrm{tr}(ST) \geq 0. \tag{A.29}$$

The *partial trace* 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 $\mathcal{H}_A$. (Here and in the following, we will use subscripts to indicate the space on which an operator acts.) It is defined by the linear extension of the mapping.

$$\mathrm{tr}_B : \quad S \otimes T \mapsto \mathrm{tr}(T)S, \tag{A.30}$$

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 $T_A \in \mathrm{End}(\mathcal{H}_A)$. That is, for any operator $S_{AB} \in \mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$,

$$\mathrm{tr}_B\big(S_{AB}(T_A \otimes \mathrm{id}_B)\big) = \mathrm{tr}_B(S_{AB})T_A \tag{A.31}$$

and

$$\mathrm{tr}_B\big((T_A \otimes \mathrm{id}_B)S_{AB}\big) = T_A \mathrm{tr}_B(S_{AB}). \tag{A.32}$$

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. That is,

$$\mathrm{tr}(S_{AB}) = \mathrm{tr}(\mathrm{tr}_B(S_{AB})), \tag{A.33}$$

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

$$\mathrm{tr}_{AB}(S_{ABC}) = \mathrm{tr}_A(\mathrm{tr}_B(S_{ABC})). \tag{A.34}$$

A.6 Decompositions of operators and vectors

Singular value decomposition. Let $S \in \mathrm{Hom}(\mathcal{H}, \mathcal{H}')$ and let $\{b_i\}_i$ ($\{b'_i\}_i$) 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

$$S = UDV^*. \tag{A.35}$$

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

$$S = \sqrt{SS^*}U \tag{A.36}$$

and

$$S = U\sqrt{S^*S}. \tag{A.37}$$

Spectral decomposition. Let $S \in \mathrm{End}(\mathcal{H})$ be normal and let $\{|b_i\rangle\}_i$ 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 $\{|b_i\rangle\}_i$ such that

$$S = UDU^*. \tag{A.38}$$

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

$$S = \sum_i \alpha_i |b_i\rangle\langle b_i| \tag{A.39}$$

where $\alpha_i \in \mathbb{C}$ are the eigenvalues of S.

Equation (A.39) 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

$$f(S) := \sum_i f(\alpha_i)|b_i\rangle\langle b_i|. \tag{A.40}$$

A.7 Operator norms and the Hilbert-Schmidt inner product

The *Hilbert-Schmidt*[7] *inner product* between two operators $S, T \in \mathrm{End}(\mathcal{H})$ is defined by

$$(S, T) := \mathrm{tr}(S^*T). \tag{A.41}$$

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

$$||S||_2 = \sqrt{\sum_i |\alpha_i|^2}. \tag{A.42}$$

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

Lemma A.7.1. *Let $S, T \in \mathrm{End}(\mathcal{H})$. If $S \geq 0$ and $T \geq 0$ then*

$$\mathrm{tr}(ST) \geq 0. \tag{A.43}$$

[7] Erhard Schmidt, 1876 – 1959, German mathematician.

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

$$\mathrm{tr}(ST) = \mathrm{tr}(V^* V) \tag{A.44}$$

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 $\phi \in \mathcal{H}$,

$$\langle \phi | V^* V | \phi \rangle = \| V \phi \|^2 \geq 0. \tag{A.45}$$

□

The *trace norm* of S is defined by

$$\|S\|_1 := \mathrm{tr}|S| \tag{A.46}$$

where

$$|S| := \sqrt{S^* S}. \tag{A.47}$$

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

$$\|S\|_1 = \sum_i |\alpha_i|. \tag{A.48}$$

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

Lemma A.7.2. *For any* $S \in \mathrm{End}(\mathcal{H})$,

$$\|S\|_1 = \max_U |\mathrm{tr}(US)| \tag{A.49}$$

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

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

$$|\mathrm{tr}(US)| \leq \mathrm{tr}|S| \tag{A.50}$$

with equality for some appropriately chosen U.

Let $S = V|S|$ be the polar decomposition of S. Then, using the Cauchy-Schwarz[8] inequality

$$|\mathrm{tr}(Q^* R)| \leq \|Q\|_2 \|R\|_2, \tag{A.51}$$

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

$$\left|\mathrm{tr}(US)\right| = \left|\mathrm{tr}(UV|S|)\right| = \left|\mathrm{tr}(UV\sqrt{|S|}\sqrt{|S|})\right| \leq \sqrt{\mathrm{tr}(UV|S|V^* U^*)\mathrm{tr}(|S|)} = \mathrm{tr}(|S|), \tag{A.52}$$

which proves (A.50). Finally, it is easy to see that equality holds for $U := V^*$. □

[8] Karl Hermann Amandus Schwarz, 1843 – 1921, German mathematician.

A.8 The vector space of Hermitian operators

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

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

$$E_{i,j} := \begin{cases} \frac{1}{\sqrt{2}}|e_i\rangle\langle e_j| + \frac{1}{\sqrt{2}}|e_j\rangle\langle e_i| & \text{if } i < j \\ \frac{i}{\sqrt{2}}|e_i\rangle\langle e_j| - \frac{i}{\sqrt{2}}|e_j\rangle\langle e_i| & \text{if } i > j \\ |e_i\rangle\langle e_i| & \text{otherwise} \end{cases} \tag{A.53}$$

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

$$\dim \mathrm{Herm}(\mathscr{H}) = (\dim \mathscr{H})^2. \tag{A.54}$$

For two Hilbert spaces $\mathscr{H}_A$ and $\mathscr{H}_B$, we have in analogy to (A.23)

$$\mathrm{Herm}(\mathscr{H}_A) \otimes \mathrm{Herm}(\mathscr{H}_B) \cong \mathrm{Herm}(\mathscr{H}_A \otimes \mathscr{H}_B). \tag{A.55}$$

To see this, consider the canonical mapping from $\mathrm{Herm}(\mathscr{H}_A) \otimes \mathrm{Herm}(\mathscr{H}_B)$ to $\mathrm{Herm}(\mathscr{H}_A \otimes \mathscr{H}_B)$ defined by (A.22). It is easy to verify that this mapping is injective. Furthermore, because by (A.54) the dimension of both spaces equals $\dim(\mathscr{H}_A)^2 \dim(\mathscr{H}_B)^2$, it is a bijection, which proves (A.55).

Bibliography

[1] Rolf Landauer. Information is physical. *Physics Today*, 44(5):23, 1991.

[2] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, September 2000.

[3] Eleanor G. Rieffel and Wolfgang H. Polak. *Quantum Computing: A Gentle Introduction*. The MIT Press, 1 edition, March 2011.

[4] S. M. Barnett. *Quantum information*. Number 16 in Oxford master series in physics. Atomic, optical and laser physics. Oxford University Press, Oxford, 2009.

[5] Benjamin Schumacher and Michael Westmoreland. *Quantum Processes Systems, and Information*. Cambridge University Press, April 2010.

[6] A. Peres. *Quantum Theory: Concepts and Methods*, volume 72 of *Fundamental Theories of Physics*. Kluwer Academic Publishers, New York, 2002.

[7] Mark Wilde. *Quantum information theory*. Cambridge University Press, Cambridge, UK. ; New York, 2013.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, July 2006.

[9] David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 1st edition, June 2002.

[10] N. David Mermin. *Quantum Computer Science: An Introduction*. Cambridge University Press, 1 edition, September 2007.

[11] Scott Aaronson. *Quantum computing since Democritus*. Cambridge University Press, 2013.

[12] D. H Mellor. *Probability: a philosophical introduction*. Routledge, London; New York, 2005.

[13] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, July 1948.

[14] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, 7(3):535–563, 1928.

[15] Myron Tribus and Edward C. McIrvine. Energy and information. *Scientific American*, 224:178–184, September 1971.

[16] Michael George Luby and Avi Wigderson. *Pairwise independence and derandomization*. Now, Boston, 2006.

[17] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Mathematische Annalen*, 63(4):433–476, December 1907.

[18] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Number 2 in Investigations in Physics. Princeton University Press, 1955.

[19] John Preskill. Lecture notes for phys 229. 2004.

[20] Man-Duen Choi. Completely positive linear maps on complex matrices. *Linear Algebra and its Applications*, 10(3):285–290, June 1975.

[21] W. Forrest Stinespring. Positive functions on ðİŘű*-algebras. *Proceedings of the American Mathematical Society*, 6(2):211–216, 1955.

[22] Karl Kraus. *States, effects, and operations: fundamental notions of quantum theory: lectures in mathematical physics at the University of Texas at Austin*. Number 190 in Lecture notes in physics. Springer-Verlag, Berlin ; New York, 1983.

[23] A. JamioÅĆkowski. Linear transformations which preserve trace and positive semidefiniteness of operators. *Reports on Mathematical Physics*, 3(4):275–278, December 1972.

[24] Richard P Feynman, Robert B Leighton, and Matthew L Sands. *The Feynman lectures on physics, Vol. 3: Quantum Mechanics*. Addison-Wesley Pub. Co., Reading, Mass., 1963.

[25] Berthold-Georg Englert. Fringe visibility and which-way information: An inequality. *Physical Review Letters*, 77(11):2154, 1996.

[26] Nai-Le Liu, Li Li, Sixia Yu, and Zeng-Bing Chen. Duality relation and joint measurement in a mach-zehnder interferometer. *Physical Review A*, 79(5):052108, May 2009.

[27] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10):777, May 1935.

[28] Erwin SchrÃűdinger. Discussion of probability relations between separated systems. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(04):555–563, 1935.

[29] Asher Peres. Unperformed experiments have no results. *American Journal of Physics*, 46(7):745, 1978.

[30] Michael M. Wolf, David Perez-Garcia, and Carlos Fernandez. Measurements incompatible in quantum theory cannot be measured jointly in any other no-signaling theory. *Physical Review Letters*, 103(23):230402, December 2009.

[31] Fernando G. S. L. BrandÃčo, Matthias Christandl, and Jon Yard. Faithful squashed entanglement. *Communications in Mathematical Physics*, 306(3):805–830, September 2011.

[32] Douglas R. Stinson. *Cryptography: Theory and Practice, Third Edition*. Chapman and Hall/CRC, 3 edition, November 2005.

[33] C. E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656, October 1949.

[34] Stephen Wiesner. Conjugate coding. *SIGACT News*, 15(1):78–88, 1983.

[35] Charles H. Bennett and Gilles Brassard. Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of IEEE International Conference on Computers Systems and Signal Processing*, pages 175–179, New York, December 1984. IEEE.

[36] Dominic Mayers. Unconditional security in quantum cryptography. *Journal of the ACM*, 48(3):351âĂŞ406, May 2001. ACM ID: 382781.

[37] Peter W. Shor and John Preskill. Simple proof of security of the BB84 quantum key distribution protocol. *Physical Review Letters*, 85(2):441, July 2000.

[38] Eli Biham, Michel Boyer, P. Oscar Boykin, Tal Mor, and Vwani Roychowdhury. A proof of the security of quantum key distribution. *Journal of Cryptology*, 19(4):381–439, April 2006.

[39] Renato Renner. *Security of Quantum Key Distribution*. PhD thesis, ETH Zurich, September 2005. http://arxiv.org/abs/quant-ph/0512258v2.

[40] Artur K. Ekert. Quantum cryptography based on bell's theorem. *Physical Review Letters*, 67(6):661, 1991. Copyright (C) 2009 The American Physical Society; Please report any problems to prola@aps.org.

[41] Matthias Christandl, Robert KÃűnig, and Renato Renner. Postselection technique for quantum channels with applications to quantum cryptography. *Physical Review Letters*, 102(2):020504–4, January 2009.