Quantum Information Theory

Joseph M. Renes
With elements from the script of Renato Renner & Matthias Christandl
and contributions by Christopher Portmann

February 4, 2015

Lecture Notes ETH Zürich HS2014

Revision: f4e02a9be866 Branch: full Date: Wed, 04 Feb 2015 14:51:33 +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. For example, we might use the spin of a single electron to store information, rather than the magnetization of a small region of magnetic material (to say nothing of ink marks on a piece of paper). As this is a pretty broad question, the field of quantum information theory overlaps with many fields: physics, of course, but also computer science, electrical engineering, chemistry and materials science, mathematics, as well as philosophy.

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! Nevertheless, quantum and classical information theory are closely related. Because any such classical system can in principle be described in the language of quantum mechanics, classical information theory is actually a (practically significant) special case of quantum information theory.

The goal of this course is to provide a solid understanding of the mathematical foundations of quantum information theory, with which we can then 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 some alphabet.

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

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

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

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

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

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

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

But what we wanted was

$$|\psi\rangle|\psi\rangle = (a|0\rangle + b|1\rangle)(a|0\rangle + b|1\rangle) \tag{1.6}$$

$$= a^2|00\rangle + ab|01\rangle + ba|10\rangle + b^2|11\rangle, \tag{1.7}$$

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.

Instead, the cloning machine extends the superposition over two systems, producing an *entangled* state. As we will see, the superposition now manifests itself only in the two systems jointly, not in either system individually. Superposition of two states is often called *coherence*, for just as two classical waves are coherent if they have a definite phase relationship, a given superposition with weights a and b also has a definite phase relationship between the two states. It turns out that for a state like (1.5), the coherence of the first system has completely vanished; there is no more detectable phase relationship between the two states $|0\rangle$ and $|1\rangle$. Of course, the coherence isn't destroyed, since it can be restored by simply applying U^*.

The interplay between coherence, cloning, and entanglement already gives us an idea of the delicate nature of quantum information processing. Superposition, or coherence, is the hallmark of the quantum nature of an information processing device. The above example shows that mere copying of the state in one basis, which we think of as copying classical information encoded in this basis, is already enough to destroy the coherence. Thus, a truly quantum information processing device cannot leak any information whatsoever, it must operate completely isolated from its environment. This requirement is one of the daunting challenges of constructing quantum devices.

1.3 Measurement and disturbance

Even if a generic qubit is not definitely in one of the states $|0\rangle$ or $|1\rangle$, 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 Bennett[2] and Brassard[3] in 1984. The goal, as in any QKD protocol, is to create a secret key between the two parties, which may 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

[2] Charles Henry Bennett, born 1943, American physicist and information theorist.
[3] Gilles Brassard, born 1955, Canadian computer scientist.

signals, measures them randomly 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 $1/4$. This attack nets Eve the value of the bit sent by Alice with probability $1/2$.

But 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. Classical information, in contrast, is not so fragile and shows no evidence of it having been copied.

Even thought 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. Showing how to accomplish this task is in fact one of the goals of the course.

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 $|x\rangle$, $|z_1\rangle$, …, $|y\rangle$. 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 Notes & 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]. Advanced treatments are offered by Hayashi [6] and two volumes by Holevo [7, 8]. An inspiration for many of these books and early lecture notes is the book by Peres [9]. Wilde [10] 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 [11]. Mackay [12] treats information theory and many other interesting topics such as Bayesian inference and neural networks from a physics point of view. Mermin [13] 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 [14] (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.

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 natural 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. Being precise at this stage will allow us to better appreciate the differences to quantum mechanics and will be useful in formulating results of classical information theory.

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 event 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.

Example 2.2.1. The simplest, not utterly trivial example is perhaps given by two coins. When flipped, each lands either heads `H` or tails `T`. The sample space is $\Omega = \{\texttt{HH}, \texttt{HT}, \texttt{TH}, \texttt{TT}\}$. Events are any subset of the elements of Ω; the event corresponding to "the first coin shows heads" is $\{\texttt{HH}, \texttt{HT}\}$, and so forth.

Example 2.2.2. A more standard example in probability theory is $\Omega = \mathbb{R}$ and the events are formed by countable unions, intersections and complements of open sets. In contrast to the discrete case, here individual elements of Ω (points in $\mathbb{R}$) are *not* events themselves.

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

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

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.

Example 2.2.3. Returning to the coin example, valid probability measures include the uniform measure defined by $P[\{\omega\}] = 1/4$ for all ω or $P[\mathtt{HH}] = P[\mathtt{TT}] = 1/2$.

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 in this general context as a *measureable space*, and the uncountably infinite case is important in the mathematical study of 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 Conditional probability and measurement

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

$$P[E|E'] := \frac{P[E \cap E']}{P[E']} \quad \forall\, E \in \mathscr{E}. \tag{2.1}$$

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 or will occur. 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. If knowing E' is certain does not change the probability of E, then E and E' are called *mutually independent*. Formally, if $P[E|E'] = P[E]$, then $P[E \cap E'] = P[E] \cdot P[E']$.

In the Bayesian framework, the conditional probability rule (2.1) 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' has occurred or is certain to occur. 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 $P[\cdot|E']$.

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 $\{\omega\}$, but really any partition will do. The measurement then reports the kth outcome with probability $P[E_k]$ and the probability measure is updated from P to P' with $P'[E] = P[E|E']$.

Notice that if we average the new probability measure over the measurement outcomes themselves, we end up with the original (here we show this for the discrete case):

$$\begin{aligned} P'[E] &:= \sum_{k=1}^{M} P[E_k]P[E|E_k] = \sum_{k=1}^{M} P[E \cap E_k] \\ &= \sum_{k=1}^{M} \sum_{\omega \in E \cap E_k} P[\{\omega\}] = \sum_{\omega \in E} P[\{\omega\}] = P[E]. \end{aligned} \tag{2.2}$$

This calculation has an important physical interpretation. Suppose we describe a physical system by the probability measure P. If we then perform a measurement on the system, but forget the result, then our probabilistic description is unchanged. We could also imagine that someone else measures the system but does not tell us the result; knowing that they have performed a measurement does not change our description of the system.

2.2.3 Random variables

In the formal setting of probability theory, random variables are functions from Ω to the space of values taken by the 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 $\mathscr{X}$,

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

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. $X^{-1}(F) \in \mathscr{E}$. The space $(\mathscr{X}, \mathscr{F})$ is often called the *range* of the random variable X.

Thus, we may define events in terms of the random variables themselves. In doing so, the events $\mathscr{F}$ inherit a probability measure P_X from the probability space, like so:[4]

$$P_X[F] := P[X^{-1}(F)] \quad \forall F \in \mathscr{F}. \tag{2.4}$$

Analogously to (2.4), the conditional probability measure also gives rise to a conditional probability measure of any random variable X, $P[\cdot|E']$, i.e.,

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

Example 2.2.4. Suppose Ω is the sample space for the roll of three dice. Then possible random variables include the sum of the faces, their product, the product of the first two minus the third, etc. Calling these random variables X, Y, and Z and supposing that each die is equally likely to show any face, then it happens that the most likely values are $X = 10, 11$, $Y = 12, 24$ and $Z = 0$.

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

$$(X, Y): \quad \omega \mapsto X(\omega) \times Y(\omega). \tag{2.6}$$

[4] This is an instance of the general mathematical concept of a *pushforward*; here the probability measure P is pushed forward to P_X by the function X. In a different guise, the pushforward is familiar when changing variables in integrals.

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 $\mathscr{X} \times \mathscr{Y}$. Naturally, this construction extends to any (finite) number of random variables.

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 $(X_1, \ldots, X_n)$.

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 $P_{X_1 \cdots X_n}$.

Example 2.2.5. Returning to the example of three dice, X_j could simply be the value of the jth die, which is how we would label the ω themselves. Then $P_{X_1 X_2 X_3}$ is identical to P, and the probability spaces for the two are equivalent. The same holds if we instead choose $Y_1 = X_1$, $Y_2 = X_2$, and $Y_3 = X_1 + X_2 + X_3$, though the labelling of individual events ω by Y is not the same as by X. On the other hand, if we choose $\mathrm{Z}_1 = X_1$, $\mathrm{Z}_2 = X_2$, and $\mathrm{Z}_3 = X_1 + X_2 + X_3 \mod 2$, then the probability space associated with $P_{\mathrm{Z}_1 \mathrm{Z}_2 \mathrm{Z}_3}$ is distinct from Ω. The event $(\mathrm{Z}_1 = 1, \mathrm{Z}_2 = 4, \mathrm{Z}_3 = 1)$ does not correspond to a single sample space element ω, but rather the compound event $(X_1 = 1, X_2 = 4, X_3 = 2, 4, 6)$.

2.3 Discrete random variables

Discrete random variables are simply those for which $\mathscr{X}$ is a discrete space. In this case, we refer to it 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 $\sum_{x \in \mathscr{X}} P_X(x) = 1$. We will often call the probability mass function of X the *probability distribution* of X.

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, $P_X(x) = \frac{1}{|\mathrm{supp} P_X|}$ for all $x \in \mathscr{X}$, where $\mathrm{supp} P_X := \{x \in \mathscr{X} : P_X(x) > 0\}$ is the *support* of the function P_X. Furthermore, P_X is *uniform* if it is flat and has no zero probabilities, whence $P_X(x) = \frac{1}{|\mathscr{X}|}$ for all $x \in \mathscr{X}$.

2.3.1 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 $\mathscr{X}$ and $\mathscr{Y}$ 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

$$P_Y(y) = \sum_{x \in \mathscr{X}} P_{XY}(x, y) \quad \forall y \in \mathscr{Y}, \tag{2.7}$$

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

$$P_{X|Y=y}(x) = \frac{P_{XY}(x,y)}{P_Y(y)} \quad \forall x \in \mathscr{X}. \tag{2.8}$$

2.3.2 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 $(x,y) \in \mathscr{X} \times \mathscr{Y}$. Their joint distribution then satisfies $P_{XY}(x,y) = P_X(x)P_Y(y)$.

Related is the notion of *Markov*[5] *chains*. 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 $i \in \{1, \ldots, n-1\}$

$$P_{X_{i+1}|X_1=x_1,\ldots,X_i=x_i} = P_{X_{i+1}|X_i=x_i} \quad \forall x_1, \ldots, x_i. \tag{2.9}$$

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 (2.9) it also holds that X_{i-1} is independent of $X_{i+1}, \ldots, X_n$ given a fixed value of X_i.

2.3.3 Functions of random variables and Jensen's inequality

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

$$P_Y(y) = \sum_{x \in f^{-1}(y)} P_X(x). \tag{2.10}$$

For a real convex function f on a convex set $\mathscr{X}$, the expectation values of X and $f(X)$ are related by *Jensen's*[6] *inequality*:

$$\langle f(X) \rangle \geq f(\langle X \rangle). \tag{2.11}$$

The inequality is essentially a direct consequence of the definition of convexity, as depicted for binary random variables in Fig. 2.1.

2.3.4 I.i.d. distributions and their asymptotic behavior

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

$$P_{X_1 \cdots X_n} = P_X^{\times n} := P_X \times \cdots \times P_X. \tag{2.12}$$

[5] Andrey Andreyevich Markov, 1856 – 1922, Russian mathematician.
[6] Johan Ludwig William Valdemar Jensen, 1859 – 1925, Danish mathematician and engineer.

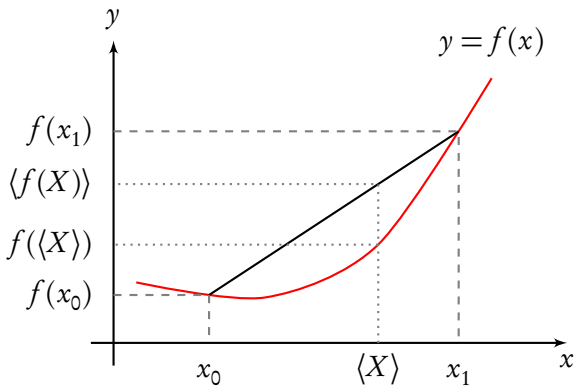


Figure 2.1: Depiction of Jensen's inequality for a convex function of a binary-valued random variable X. The random variable X can take two possible values, x_0 and x_1, with corresponding probabilities $P_X(x_0)$ and $P_X(x_1)$. A function f induces a new random variable $Y = f(X)$; for convex f it follows that $\langle Y \rangle \geq f(\langle X \rangle)$.

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, for example in repeated uses of a communication channel.

The *law of large numbers* and the *central limit theorem* characterize the "typical behavior" of real-valued i.i.d. random variables $X_1, \ldots, X_n$ in the limit of large n. The law of large numbers states that the sample mean of the X_i tends to the expectation value for large n. It usually comes in two versions, the *weak* and the *strong* law. As the names suggest, the latter implies the former.

More precisely, 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

$$Z_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.13}$$

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:

$$\lim_{n \to \infty} P\big[|Z_n - \mu| < \varepsilon\big] = 1 \quad \forall \varepsilon > 0. \tag{2.14}$$

The weak law of large numbers will be sufficient for our purposes, and is proven in the exercises. By contrast, the *strong law of large numbers* says that Z_n converges to μ with probability one,

$$P\big[\lim_{n \to \infty} Z_n = \mu\big] = 1. \tag{2.15}$$

Note that to properly interpret the strong law, we need the formal machinery of the underlying probablity space $(\Omega, \mathcal{E}, P)$, since the number of random variables is infinite. One way to remember the difference between the weak and strong laws is to realize that they are essentially saying the same thing, but using different notions of convergence. The weak law is a statement of *convergence in probability*, while the strong law is a statement of *almost-sure convergence*.

While the laws of large numbers tell us about the behavior of the sample mean, the *central limit theorem* gives some insight into the behavior of fluctuations around the mean, at least when the X_i

have bounded variance σ^2. In particular, let Φ be the cumulative distribution function of a standard normal distribution (zero-mean Gaussian with unit variance) and define the rescaled *fluctuation variable*

$$Y_n = \sqrt{n}(Z_n - \mu)/\sigma. \tag{2.16}$$

Then the central limit theorem asserts that the cumulative distribution of Y_n converges to that of the normal distribution:

$$\lim_{n\to\infty} P[Y_n \leq y] = \Phi(y). \tag{2.17}$$

This type of convergence is called *convergence in distribution*. It is weaker than either of the other two notions mentioned above.

The statements above only in the limit $n \to \infty$, but it is often more interesting to have bounds on the deviation of i.i.d. random variables from their typical behavior for finite n. For the deviation from the mean, such a statement is for instance provided by the *Hoeffding*[7] *bound*. Suppose that the random variables X_j takes values in a bounded range, from a to b. Then

$$P[|Z_n - \mu| \geq \varepsilon] \leq 2\exp\left[-\frac{2n\varepsilon^2}{(b-a)^2}\right]. \tag{2.18}$$

Meanwhile, the *Berry-Esseen*[8] *theorem* provides a bound on the speed of convergence in the central limit theorem. Supposing that $\rho := \langle |X_j|^3 \rangle < \infty$ and defining $F_n(y)$ to be the cumulative distribution of Y_n, the theorem states that there is a constant C (known to be less than one half) such that

$$|F_n(y) - \Phi(y)| \leq \frac{C\rho}{\sigma^3\sqrt{n}} \qquad \forall n, y. \tag{2.19}$$

2.4 Channels

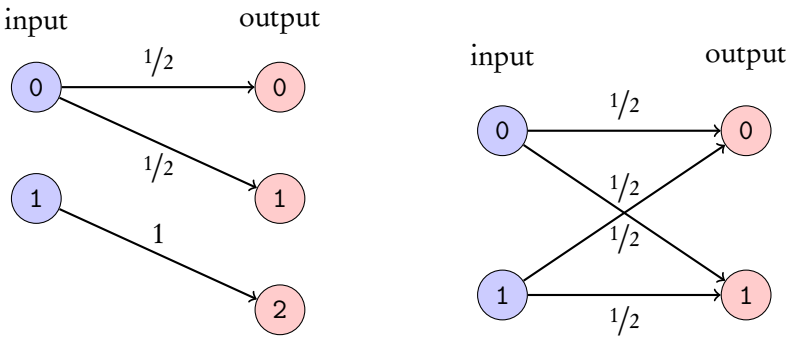
A *channel* W is a probabilistic mapping that assigns to each value of an *input alphabet* $\mathcal{X}$ a value of the *output alphabet* $\mathcal{Y}$. In doing so, it transforms the random variable X to the random variable $Y = W(X)$. It is specified by assigning a number $W(y|x)$ to each input-ouput pair (x, y) such that such that $W(\cdot|x)$ is a probability mass function for any $x \in \mathcal{X}$.

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 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 through 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.

Example 2.4.1. The channel depicted in Fig. 2.2(a) maps the input `0` to either `0` or `1` with equal probability; the input `1` is always mapped to `2`. The channel has the property that its input is uniquely determined by its output. Such a channel would allow the reliable transmission of one classical bit of information.

[7] Wassily Hoeffding, 1914 – 1991, Finnish statistician and probabilist.
[8] Carl-Gustav Esseen, 1918 – 2001, Swedish mathematician.



(a) A reliable channel (b) An unreliable channel

Figure 2.2: Examples of channels.

Example 2.4.2. The channel shown in Fig. 2.2(b) 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.

2.4.1 Formal definition and properties

Formally, a channel is a transformation of $\mathscr{X}$ to $\mathscr{Y}$ which preserves the probability structure. In other words, the channel must induce a valid probability distribution on $\mathscr{Y}$ from the distribution over $\mathscr{X}$. Given any fixed input $X = x$, $W(X = x)$ must be a probability mass function over $\mathscr{Y}$. Therefore, the earlier description using $W(y|x) = W(X = x)$ encompasses all possible channels. Indeed, we may regard $W(X = x)$ as the conditional distribution $P_{Y|X=x}$ associated with the joint distribution

$$P_{XY}(x,y) = P_X(x)W(y|x). \tag{2.20}$$

Moreover, channels can be seen as generalizations of functions (random variables). Indeed, if f is a function from $\mathscr{X}$ to $\mathscr{Y}$, its description as a channel W is given by

$$W(y|x) = \delta_{y,f(x)}. \tag{2.21}$$

Returning to the definition of Markov chains in (2.9), it is easy to see that a Markov chain is a sequence of random variables in which X_{j+1} is generated from X_j by some channel W_j.

2.4.2 Measurement as a channel

The process of measurement, described in §2.2.2 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 $\mathscr{X}$ into a set of disjoint events, i.e. a collection of sets E_y, $y = 1,\ldots,|\mathscr{Y}|$ 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

$$W(y|x) = \begin{cases} 1 & x \in E_y \\ 0 & \text{else} \end{cases}. \tag{2.22}$$

Now consider the joint distribution P_{XY}, given by (2.20). The marginal distribution for Y is simply

$$P_Y(y) = \sum_{x \in \mathscr{X}} P_X(x) W(y|x) = \sum_{x \in E_y} P_X(x), \tag{2.23}$$

the probability distribution of the measurement outcomes. Moreover, by (2.1), the conditional distribution of X given Y is just

$$P_{X|Y=y}(x) = \frac{1}{P_Y(y)} P_X(x) W(y|x). \tag{2.24}$$

Thus, the joint distribution (2.20) induced by the channel 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. This description of measurement is depicted in Fig. 2.3.

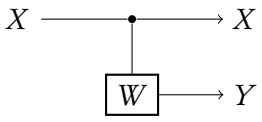


Figure 2.3: Depiction of measurement as a channel. Depending on the value of the input X, the measurement produces the result Y. This conditional action is represented by the dot and wire to the channel W.

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.

2.5 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.

2.5.1 Represention of the probability space

Since Ω is finite, we may define $N = |\Omega|$ and associate the elements ω with a basis $\{\vec{b}_\omega\}$ of $\mathbb{R}^N$. In particular, we could label ω with integers from 1 to N and define $\vec{b}_\omega$ to be the vector with a single 1 in the ωth component and all entries 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 $\vec{e}(E) \in \mathbb{R}^N$ associated with the event E is defined by

$$\vec{e}(E) = \sum_{\omega \in E} \vec{b}_\omega, \tag{2.25}$$

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

$$P(E) = \sum_{\omega\in E} P[\{\omega\}] = \sum_{\omega\in\Omega} \vec{e}(E)\cdot\vec{b}_\omega P[\{\omega\}] = \vec{e}(E)\cdot\left(\sum_{\omega\in\Omega}\vec{s}_\omega P[\{\omega\}]\right). \tag{2.26}$$

This suggests we define a vector $\vec{p} = \sum_{\omega\in\Omega}\vec{b}_\omega P[\{\omega\}] \in \mathbb{R}_+^N$, 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:

$$P[E] = \vec{e}(E)\cdot\vec{p}. \tag{2.27}$$

2.5.2 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 $\vec{x} = \sum_{\omega\in\Omega} x_\omega \vec{b}_\omega$. Then the *expected value* of X is just the average value under the probability distribution,

$$\langle X\rangle \coloneqq \sum_{\omega\in\Omega} P[\{\omega\}]X(\omega) = \vec{x}\cdot\vec{p}. \tag{2.28}$$

We can also succinctly represent the rule for conditional probability, (2.1), 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 (2.1) 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

$$\vec{p}\,' = \frac{1}{\vec{e}(E')\cdot\vec{p}}\sum_{\omega\in E'}\left(\vec{b}_\omega\cdot\vec{p}\right)\vec{b}_\omega = \frac{1}{\vec{e}(E')\cdot\vec{p}}\sum_{\omega\in\Omega}\left(\vec{b}_\omega\cdot\vec{p}\right)\left(\vec{e}(E')\cdot\vec{b}_\omega\right)\vec{b}_\omega. \tag{2.29}$$

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.

2.5.3 Transformations and dilations

We have seen in §2.4 that channels describe all transformations of probability distributions that preserve the probability structure. Inded, (2.23) shows that the transformation is *linear* and the transition probabilities $W(y|x)$ are the components of the matrix representation of the channel. Such matrices, called *stochastic matrices*, have positive entries and column-sums all equal to one by definition.

For given input and output alphabet sizes $n = |\mathcal{X}|$ and $m = |\mathcal{Y}|$, respectively, the set $\mathrm{St}(m,n)$ of all $m\times n$ stochastic matrices is convex, since the convex mixture of two channels is clearly also a valid channel. As the entries are bounded between 0 and 1 and each column sums to 1 (providing n linear

constraints), $\mathrm{St}(m,n)$ is a closed, compact subset of $\mathbb{R}^{m(n-1)}$. Indeed, it must be a convex polytope because the boundaries are specified by linear relations.

Among the stochastic matrices are the representations of deterministic transformations, i.e. usual functions as in (2.21). These are stochastic matrices with a single 1 in every column. Clearly, deterministic transformations cannot be nontrivially decomposed into other stochastic matrices, since the entries are bounded between 0 and 1. Thus, they are *extreme points* in $\mathrm{St}(m,n)$. In fact, they are the only extreme points, and every stochastic matrix can be expressed as a convex combination of deterministic transformations. To see why this is so, first denote by $D(j_1,\ldots,j_n)$ the matrix whose ith column has a 1 in the j_ith row, and zeros elsewhere. Then, an arbitrary $T \in \mathrm{St}(m,n)$ with components T_{ij} can be expressed as

$$T = \sum_{j_1,\ldots,j_n=1}^{m} T_{j_1,1}\cdots T_{j_n,n} D(j_1,\ldots,j_n). \tag{2.30}$$

The coefficient for a given $D(j_1,\ldots,j_n)$ is simply the product of the transition probabilities for $X=1$ to be mapped to $Y=j_1$ and so forth. The coefficients make up a convex combination, since they are positive real numbers whose sum is unity (it is the product of the column sums of T). Summing over the coefficients of $D(j_1,\ldots,j_{\ell-1},k,j_{\ell+1},\ldots,j_n)$ for arbitrary $j_1,\ldots,j_{\ell-1},j_{\ell+1},\ldots j_n$, we recover $T_{k,\ell}$ as intended.

The number of vertices of the $\mathrm{St}(m,n)$ polytope is m^n, the number of distinct deterministic matrices from $\mathscr{X}$ to $\mathscr{Y}$. However, any given T can be expressed as a combination of just $m(n-1)+1$ vertices (one plus the dimension of the space in which the polytope lives) by *Carathéodory's*[9] *theorem*, though we will not make use of this fact here. Altogether we have the following proposition.

Proposition 2.5.1: Convex decomposition of stochastic matrices

Any $T \in \mathrm{St}(m,n)$ can be expressed as a convex combination of no greater than $m(n-1)+1$ deterministic transformations.

Using this representation, we can construct a *dilation* of any stochastic transformation of single random variables to a deterministic transformation on a pair of random variables. This is but the first example of a dilation we will meet in this course. Loosely speaking, the idea of a dilation is to regard any given element of a convex set as the image under some fixed map of an extreme point of the associated convex set in a larger space. In the present case, suppose that we have a channel W from $\mathscr{X}$ to $\mathscr{Y}$ and a representation $W = \sum_{z=1}^{|\mathscr{Z}|} \lambda_z D_z$ in terms of deterministic maps. Then, let Z be a random variable over alphabet $\mathscr{Z}$, with $P_Z(z) = \lambda_z$, and define the new transformation $W_{\mathrm{det}} : \mathscr{X} \times \mathscr{Z} \to \mathscr{Y} \times \mathscr{Z}$ which applies D_z to $\mathscr{X}$ when $Z = z$. In other words, if D_z is the representation of the function f_z, W_{det} deterministically applies f_z conditioned on the value of Z. The original transformation W is recovered by marginalizing over the additional random variable Z. This is depicted in Fig. 2.4. Formally, we have shown the following.

Proposition 2.5.2: Dilation of channels to deterministic functions

For any channel $W : \mathscr{X} \to \mathscr{Y}$ there exists a deterministic channel $W_{\mathrm{det}} : \mathscr{X} \times \mathscr{Z} \to \mathscr{Y} \times \mathscr{Z}$ and

[9] Constantin Carathéodory, 1873 – 1950, Greek mathematician.

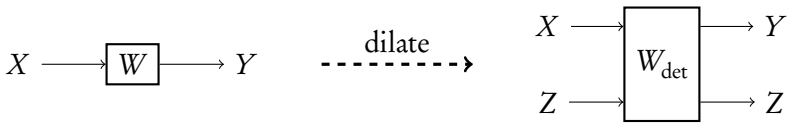


Figure 2.4: Dilation of a channel W to a deterministic channel W_{det}. Essentially, the randomness inherent in the channel is extracted to the random variable Z, which is then used to control which deterministic function is applied to X.

a random variable Z (probability distribution P_Z) such that

$$P_{W(X)}(y) = \sum_{z \in \mathscr{Z}} P_{W_{\text{det}}(X,Z)}(y,z). \tag{2.31}$$

If the row-sums of a stochastic matrix are also all unity, the matrix is called *doubly stochastic*. In this case, it follows that the input and output random variables or probability spaces have the same size, since the sum of all entries equals both the number of rows and the number of columns. Equivalently, a doubly stochastic matrix is a stochastic matrix which maps the uniform distribution to itself. Doubly stochastic $m \times n$ matrices also form a convex set.

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*. Similarly to the case of stochastic matrices, doubly stochastic matrices can be expressed as convex combinations of permutations, a fact known as *Birkhoff's*[10] *theorem*. We may therefore dilate any doubly stochastic transformation to a reversible transformation by following the same construction as above.

2.6 Notes & Further reading

Mellor gives a nice introduction to the philosophy of probability theory in [15]. Jaynes describes how Bayesian probability "ought" to be applied in science in [16]; to say that opinions vary among mathematical and statistical researchers is an understatement. For more on the mathematical structure of probability theory itself, see the introductory text by Ross [17], an intermediate approach by Gut [18], and recent in-depth treatments by Durrett [19] and Fristedt and Gray [20]. The discussion of stochastic matrices is adapted from Davis [21].

[10] Garrett Birkhoff, 1911 – 1996, American mathematician.

2.7 Exercises

Exercise 2.1. Statistical distance → solution

The *statistical distance*, or total variational distance, between two probability distributions P and Q over an alphabet $\mathcal{X}$ is generally defined by

$$\delta(P,Q) \coloneqq \sup_{S\subseteq X} |P[S]-Q[S]|, \tag{2.32}$$

where the maximization is over all events $S \subseteq X$. For finite alphabets that we are considering in this course, the supremum can be replaced by a maximum.

a) Show that $\delta(\cdot,\cdot)$ is a good measure of distance by proving that $0 \leq \delta(P,Q) \leq 1$ and the triangle inequality $\delta(P,R) \leq \delta(P,Q)+\delta(Q,R)$ for arbitrary probability distributions P, Q and R.

b) Suppose that P and Q represent the probability distributions of the outcomes of two dice, which we can also label P and Q. You are allowed to throw one of them once and then have to guess which it was. What is your best strategy? What is the probability that you guess correctly and how can you relate that to the statistical distance $\delta(P,Q)$?

c) Show that for a finite alphabet the statistical distance can also be expressed as

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

Exercise 2.2. Jensen's inequality → solution

For any convex function f and probability distribution $\{p_1,..,p_n\}$, prove Jensen's inequality:

$$f\Big(\sum_{k=1}^{n} p_k x_k\Big) \leq \sum_{k=1}^{n} p_k f(x_k). \tag{2.34}$$

Regarding the x_k as defining a random variable X, this can also be written as

$$f(\langle X\rangle) \leq \langle f(X)\rangle. \tag{2.35}$$

Exercise 2.3. Weak law of large numbers → solution

Let A be a positive random variable with expectation value $\langle A\rangle = \sum_a a\ P_A(a)$. Let $P[A \geq \varepsilon]$ denote the probability of an event $\{A \geq \varepsilon\}$.

a) Prove Markov's inequality

$$P[A \geq \varepsilon] \leq \frac{\langle A\rangle}{\varepsilon}. \tag{2.36}$$

b) Use Markov's inequality to prove Chebyshev's inequality

$$P\big[(X-\mu)^2 \geq \epsilon\big] \leq \frac{\sigma^2}{\epsilon}, \tag{2.37}$$

where $\mu = \langle X\rangle$ and σ denotes the standard deviation of X.

c) Use Chebyshev's inequality to prove the weak law of large numbers for i.i.d. X_i with expectation value μ and variance $\sigma^2 \leq \infty$:

$$\lim_{n\to\infty} P\left[\left(\frac{1}{n}\sum_i X_i - \mu\right)^2 \geq \varepsilon\right] = 0 \quad \text{for any } \varepsilon > 0. \tag{2.38}$$

Exercise 2.4. Conditional probabilities: Knowing more does not always help → solution

You and your grandfather are trying to guess if it will rain tomorrow. All he knows is that it rains on 80% of the days. You know that and you also listen to the weather forecast and know that it is right 80% of the time and is always correct when it predicts rain.

a) What is the optimal strategy for your grandfather? And for you?

b) Both of you keep a record of your guesses and the actual weather for statistical analysis. After some time, i.e. enough that you can apply the weak law of large numbers, who will have guessed correctly more often?

3

Quantum Mechanics

In this chapter we present the formalism of quantum mechanics and investigate the similarities to and differences from classical mechanics and classical probability theory.

3.1 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[1] and von Neumann[2] regarding the states of physical systems as well as their evolution. 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[3] space $\mathscr{H}$. In particular, any physical state can be represented by a *normalized vector* $|\phi\rangle \in \mathscr{H}$ which is unique up to a phase factor. In the following, we will call $\mathscr{H}$ the *state space* of the system.

2) *Dynamics:*
For any possible evolution of an isolated physical system with state space $\mathscr{H}$ and for any fixed time interval $[t_0, t_1]$ there exists a *unitary* U describing the mapping of states $|\phi\rangle \in \mathscr{H}$ at time t_0 to the state $|\phi'\rangle = U|\phi\rangle$ at time t_1. The unitary U is unique up to a phase factor. This is the *Schrödinger*[4] *picture*, and the unitary is determined from the Hamiltonian[5] of the system by the Schrödinger equation.

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 $\mathscr{H}$. Each eigenvalue x of an observable O corresponds to a possible value of the observable. Since O is self-adjoint, it takes the form $O = \sum_x x\Pi_x$, where Π_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 $|\phi\rangle \in \mathscr{H}$, then the probability of observing outcome x is given by the *Born*[6] *rule*:

$$P_X(x) = \mathrm{Tr}[\Pi_x|\phi\rangle\langle\phi|]. \tag{3.1}$$

[1] Paul Adrien Maurice Dirac, 1902 – 1984, English physicist.
[2] John von Neumann, 1903 – 1957, Hungarian-American mathematician and polymath.
[3] David Hilbert, 1862 – 1943, German mathematician.
[4] Erwin Rudolf Josef Alexander Schrödinger, 1887 – 1961, Austrian physicist.
[5] William Rowan Hamilton, 1805 – 1865, Irish physicist, astronomer, and mathematician.
[6] Max Born, 1882 – 1970, German physicist and mathematician.

The state $|\phi'_x\rangle$ of the system after the measurement, conditioned on the event that the outcome is x, is given by

$$|\phi'_x\rangle := \sqrt{\frac{1}{P_X(x)}}\Pi_x|\phi\rangle. \tag{3.2}$$

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

$$|\Psi\rangle = |\phi\rangle \otimes |\phi'\rangle \in \mathscr{H}_A \otimes \mathscr{H}_B. \tag{3.3}$$

3.2 Qubits

The simplest quantum system, the qubit, has just two levels, a state space of $\mathscr{H} = \mathbb{C}^2$. We typically denote a "standard basis" for a qubit by the states $|0\rangle$ and $|1\rangle$. A qubit is any system, or more precisely degree of freedom, whose state vector $|\psi\rangle$ can be written as

$$|\psi\rangle = a|0\rangle + b|1\rangle, \tag{3.4}$$

with $a, b \in \mathbb{C}^2$ and $|a|^2 + |b|^2 = 1$. Any two vectors $|\psi\rangle$ and $|\psi'\rangle$ such that $|\psi'\rangle = c|\psi\rangle$ for some $c \in \mathbb{C}$ with $|c| = 1$ represent the same state. The following table lists several examples of qubit systems.

Degree of freedom	Possible basis states $\|0\rangle$ and $\|1\rangle$	
Spin-1/2	$\|m = 1/2\rangle$	$\|m = -1/2\rangle$
Photon polarization	$\|$horizontal$\rangle$	$\|$vertical$\rangle$
"Two-level" atom	$\|$groundstate$\rangle$	$\|$excitedstate$\rangle$
Position in a deep double well potential	$\|$left$\rangle$	$\|$right$\rangle$

Table 3.1: Examples of qubit systems

A useful parameterization of states comes from the spin-1/2 picture. Any state $|\psi\rangle$ can be associated with a point on the unit sphere described by spherical coordinates (θ, φ) via the relation

$$|\psi\rangle = \cos\tfrac{\theta}{2}|0\rangle + e^{i\varphi}\sin\tfrac{\theta}{2}|1\rangle. \tag{3.5}$$

This sphere of states is called the *Bloch*[7] *sphere*, as depicted in Fig. 3.1.

Equivalently, we can label states by *Bloch vectors*, unit vectors $\hat{n} = \hat{x}\sin\theta\cos\varphi + \hat{y}\sin\theta\sin\varphi + \hat{z}\cos\theta$. Then it is easy to see that the states $|\hat{n}\rangle$ and $|-\hat{n}\rangle$ are orthogonal. The states along the six cardinal directions ($\pm\hat{x}$, $\pm\hat{y}$, and $\pm\hat{z}$) form three orthogonal bases, and the states $|\pm\hat{x}\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)$

[7] Felix Bloch , 1905 – 1983, Swiss physicist.

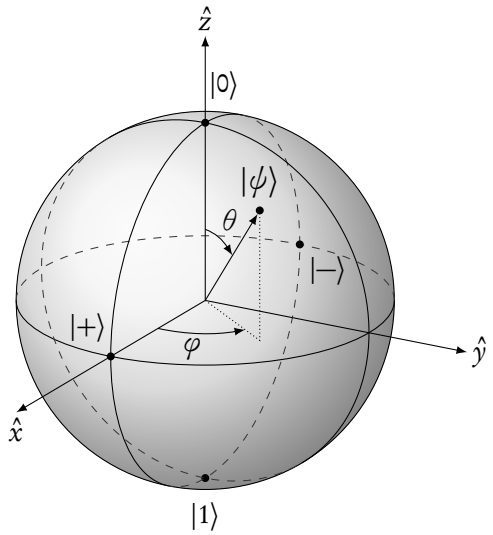


Figure 3.1: The Bloch sphere. Every qubit state can be associated with a point on the unit sphere.

are usually just denoted $|\pm\rangle$. These three bases are the eigenbases of the three *Pauli*[8] *operators*:

$$\sigma_x = |0\rangle\langle 1| + |1\rangle\langle 0| = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3.6}$$

$$\sigma_y = -i|0\rangle\langle 1| + i|1\rangle\langle 0| = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tag{3.7}$$

$$\sigma_z = |0\rangle\langle 0| - |1\rangle\langle 1| = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{3.8}$$

here the matrices are the representations of the respective operators in the basis $\{|0\rangle, |1\rangle\}$. A linear combination of Pauli operators with real coefficients leads to a Hermitian[9] operator.

These three operators, together with the identity operator $\mathbb{1}$, form a very convenient basis for operators on $\mathbb{C}^2$, i.e. a basis for $\mathrm{End}(\mathbb{C}^2)$. This follows because we can very easily construct the matrices $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, etc. from the Pauli operators, and the latter is evidently a basis for $\mathrm{End}(\mathbb{C}^2)$. Writing $A = a_0\mathbb{1} + \vec{a}\cdot\vec{\sigma}$ for an operator A, with $\vec{\sigma} = \hat{x}\sigma_x + \hat{y}\sigma_y + \hat{z}\sigma_z$, it is straightforward to verify that $|\pm\hat{a}\rangle$ are the eigenstates of A, with eigenvalues $\lambda_\pm = a_0 \pm ||\vec{a}||_2$. Here $\hat{a}$ is the normalized version of $\vec{a}$.

Using this relation, we can immediately infer that the projection operators $\Pi_{\hat{n}} := |\hat{n}\rangle\langle\hat{n}|$ take the form

$$\Pi_{\hat{n}} = \tfrac{1}{2}(\mathbb{1} + \hat{n}\cdot\sigma). \tag{3.9}$$

Then it is simple to verify that for the state $|\hat{m}\rangle$, the probability of obtaining $\Pi_{\hat{n}}$ in a measurement is just

$$P[\Pi_{\hat{n}}|\hat{m}] = \tfrac{1}{2}(1 + \hat{n}\cdot\hat{m}). \tag{3.10}$$

[8] Wolfgang Ernst Pauli, 1900 – 1958, Austrian-born Swiss theoretical physicist.
[9] Charles Hermite, 1822 – 1901, French mathematician.

Note that even though $|+\rangle$ looks like a probabilistic mixture of $|0\rangle$ and $|1\rangle$, and does give $P[\pm\hat{z}|\hat{x}] = \frac{1}{2}$, it is not a mixture at all, since $P[\hat{x}|\hat{x}] = 1$. This distinction is what is meant by saying that $|+\rangle$ is a *coherent* combination of $|0\rangle$ and $|1\rangle$. An incoherent combination is simply the probabilistic mixture, which would have $P[\hat{x}] = \frac{1}{2}$. In classical probability, this cannot occur: if a classical bit has probability $1/2$ to be in either state, then there is no other elementary event, like $\Pi_{\hat{x}}$, for which the probability is one. Quantum-mechanically we can observe the relative phase between the two elementary states.

Another way to describe this phenomena is to say that σ_x and σ_z are *complementary* observables. We have just seen that an eigenstate of one of these observables has no definite value of the other, as a measurement in the other basis is completely random. But the situation is even stranger than this. Since measurement disturbs the state, alternately measuring the two observables can lead to a situation in which a once-certain outcome is made random. After a σ_z measurement of $|+\rangle$, for instance, the state is either $|0\rangle$ or $|1\rangle$, both of which have only a probability of $1/2$ of returning $+$ in a measurement of σ_x. Without the intervening σ_z measurement, the result would of course have been $+$ with certainty. Needless to say, this is also impossible in classical probability theory. While measurement changes the probability distribution $\vec{p}$, it does not disturb the underlying states $\vec{b}_j$. Events $\vec{e}_k$ which are certain (i.e. contain the actual state $\vec{b}_j$), remain so after the measurement.

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

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 pairing" 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 $\Pi_\phi = |\phi\rangle\langle\phi|$ and $\Pi_{\overline{\phi}} = \mathbb{1} - |\phi\rangle\langle\phi|$. 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 from the table: the sample space. Actually, we have implicitly used a classical sample space for measurement outcomes, in defining the probability rule. But this gives many sample spaces for a quantum system, one for every possible measurement, and it is not clear how these are related to one another. What could be the

analog of *the* sample space in quantum theory? One is tempted to say that the $|\psi\rangle$ are the quantum version of the $\vec{b}_\omega$, since sharp distributions $\vec{p}$ are essentially equivalent to $\vec{b}_\omega$. But then comes the famous measurement problem. We are looking for a sample space to model the "real, physical state" of the system. In that case, however, why does the state (now a physical thing) evolve unitarily under "normal" dynamics but differently (collapse) 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 only the jumble of sample spaces associated with measurement outcomes to refer the probability to. What is it about a quantum system that a measurement is supposed to reveal? Surely more than just "the event that the measurement outcome is ...". If there is a useful underlying sample space that relates all of the measurement spaces, why don't we just formulate quantum mechanics directly in these terms? As we'll see when discussing the Bell[10] inequalities in §3.9, 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 $\vec{b}_\omega$, or just as a nice way to encode the probability of various measurement events, like $\vec{p}$? As far as I know, there's no satisfactory answer to this question, though many are convinced by their particular approaches to solve the riddle. One could also hope that the very 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.

In quantum information theory it is useful and common to take the latter approach above. This is an operational or instrumental approach to interpreting the formal objects in quantum mechanics. That is, we view the job of the theory as describing experimental setups, and the formal objects in the theory therefore refer to the different parts of an experiment. In the most abstract setting, we can think of an experiment as consisting of two parts, first *preparation*, in which we have set up the experimental apparatus in some way, and then *measurement*, in which we run the experiment and record the results. Most importantly, this point of view shifts our conception of the 'state' of a system: Instead of referring to the "real degrees of freedom", it instead refers to the preparation. The Born rule now plays the central role, as it tells us the probability of that a given measurement will result in a particular output, given a particular preparation. The analogy presented above reveals that classical probability theory and quantum mechanics are very similar from this vantage point.

3.4 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 written as

$$|\Psi\rangle = \sum_{j=1}^{d_A} \sum_{k=1}^{d_B} \Psi_{jk} |b_j\rangle \otimes |b'_k\rangle. \tag{3.11}$$

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

Here d_A (d_B) is the dimension of $\mathscr{H}_A$ ($\mathscr{H}_B$). In fact, we can always adapt the bases of A and B such that the summation only contains "diagonal" elements, where the indices of the basis elements are equal. This is called the *Schmidt*[11] *decomposition* of the state. Formally, we have

Proposition 3.4.1: Schmidt decomposition

Given any bipartite state $|\Psi\rangle_{AB} \in \mathscr{H}_A \otimes \mathscr{H}_B$, there exist orthonormal bases $\{|\xi_j\rangle\}_A$ and $\{|\eta_j\rangle\}_B$ for $\mathscr{H}_A$ and $\mathscr{H}_B$, respectively, such that

$$|\Psi\rangle_{AB} = \sum_{j=1}^{d_{\min}} \lambda_j |\xi_j\rangle_A \otimes |\eta_j\rangle_B, \tag{3.12}$$

where $d_{\min} = \min\{d_A, d_B\}$ and the λ_j are such that $\lambda_j \geq 0$ for all j and $\sum_j \lambda_j^2 = 1$.

Proof. 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. 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

$$\begin{aligned} |\Psi\rangle &= \sum_{j=1}^{d_A} \sum_{k=1}^{d_B} \sum_{\ell}^{d_{\min}} U_{j,\ell} \lambda_\ell [V^*]_{\ell,k} |b_j\rangle \otimes |b'_k\rangle = \sum_{\ell}^{d_{\min}} \lambda_\ell \Big(\sum_{j=1}^{d_A} U_{j,\ell} |b_j\rangle \Big) \otimes \Big(\sum_{k=1}^{d_B} V^*_{k,\ell} |b'_k\rangle \Big) \\ &= \sum_{\ell}^{d_{\min}} \lambda_\ell |\xi_\ell\rangle \otimes |\eta_\ell\rangle, \end{aligned} \tag{3.13}$$

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. Since the singular values are positive and the state is assumed to be normalized, $\sum_j \lambda_j^2 = 1$. □

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*. Choosing a basis $\{|b_k\rangle\}$ for $\mathscr{H}_A \simeq \mathscr{H}_B$, the *canonical maximally entangled* state is given by

$$|\Phi\rangle_{AB} \coloneqq \frac{1}{\sqrt{d_A}} \sum_k |b_k\rangle_A \otimes |b_k\rangle_B. \tag{3.14}$$

3.5 No cloning & no deleting

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 §1.2. Attempting to clone a general qubit state $|\psi\rangle = a|0\rangle + b|1\rangle$ results in the entangled state $a|0\rangle\otimes|0\rangle + b|1\rangle\otimes|1\rangle$. This argument works for systems of any dimension, so we have the following

[11] Erhard Schmidt, 1876 – 1959, German mathematician.

Proposition 3.5.1: No cloning

There exists no unitary operator U_{AB} on $\mathscr{H}_A \otimes \mathscr{H}_B$ such that, for fixed $|\varphi\rangle_B$ and all $|\psi\rangle_A$,

$$U_{AB}|\psi\rangle_A \otimes |\varphi\rangle_B = |\psi\rangle_A \otimes |\psi\rangle_B, \tag{3.15}$$

A similar argument shows that it is also not possible to *delete* arbitrary quantum states, i.e. turn two copies into one. Here we need three systems to include the state of the deleting apparatus.

Proposition 3.5.2: No deleting

There exists no unitary operator U_{ABC} on $\mathscr{H}_A \otimes \mathscr{H}_B \otimes \mathscr{H}_C$ such that

$$U_{ABC}|\psi\rangle_A \otimes |\psi\rangle_B \otimes |\eta\rangle_C = |\psi\rangle_A \otimes |\varphi\rangle_B \otimes |\eta'\rangle_C \tag{3.16}$$

for fixed $|\varphi\rangle_B$, $|\eta\rangle_C$, $|\eta'\rangle_C$ and all $|\psi\rangle_A$.

Proof. Consider two different input states $|\psi\rangle$ and $|\psi'\rangle$ to the deleting machine. The overlap between the initial states is simply $\langle\psi|\psi'\rangle^2$, but at the output it is $\langle\psi|\psi'\rangle$, since the other states $|\varphi\rangle$, $|\eta\rangle$, and $|\eta'\rangle$ are fixed. A unitary operation would however leave the overlap invariant. □

3.6 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. Each is constructed using a basis of maximally-entangled states of two qubits, called the Bell basis.

3.6.1 The Bell basis

The canonical maximally entangled state of two qubits is

$$|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{AB}\,. \tag{3.17}$$

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

$$|\Phi_x\rangle_{AB} = (\mathbb{1}_A \otimes (\sigma_x)_B)|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}(|01\rangle + |10\rangle)_{AB}\,. \tag{3.18}$$

If the qubit is a spin-$\frac{1}{2}$ particle, σ_x corresponds to a rotation by π around the x-axis. Clearly this state is orthogonal to $|\Phi\rangle$. What about σ_z?

$$|\Phi_z\rangle_{AB} = (\mathbb{1}_A \otimes (\sigma_z)_B)|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}(|00\rangle - |11\rangle)_{AB}\,. \tag{3.19}$$

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

$$|\Phi_y\rangle_{AB} = \big(\mathbb{1}_A \otimes (-i\sigma_y)_B\big)|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}(|01\rangle - |10\rangle)_{AB}\,, \tag{3.20}$$

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 starting with $|\Phi_y\rangle$ and noticing that since σ_y commutes with rotations about the y axis, it cannot change the value of J_y.

3.6.2 Superdense coding

Consider again our two separated parties from §1.4, Alice and Bob. Alice would like to send a message to Bob, a message composed of two bits (sell stocks? buy gold?), but she only has 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 with projectors $\Pi_j = |\Phi_j\rangle\langle\Phi_j|$

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).

3.6.3 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 $|\Phi_j\rangle$ 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 $|\Phi\rangle$?

$$
\begin{aligned}
{}_{A'A}\langle\Phi|\psi\rangle_{A'}|\Phi\rangle_{AB} &= {}_{A'A}\langle\Phi|\tfrac{1}{\sqrt{2}}(a|000\rangle + a|011\rangle + b|100\rangle + b|111\rangle)_{A'AB} && (3.21)\\
&= \tfrac{1}{2}(\langle 00| + \langle 11|)_{A'A}(a|000\rangle + a|011\rangle + b|100\rangle + b|111\rangle)_{A'AB} && (3.22)\\
&= \tfrac{1}{2}(a|0\rangle + b|1\rangle)_B = \tfrac{1}{2}|\psi\rangle_B. && (3.23)
\end{aligned}
$$

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 1/4. And what about the other results?

Since $|\Phi_x\rangle_{AA'} = (\sigma_x)_{A'}|\Phi\rangle_{AA'}$, it follows that

$$_{A'A}\langle\Phi_x|\psi\rangle_{A'}|\Phi\rangle_{AB} =_{A'A} \langle\Phi|(\sigma_x)_{A'}|\psi\rangle_{A'}|\Phi\rangle_{AB} = \tfrac{1}{2}(\sigma_x)_B|\psi\rangle_B, \tag{3.24}$$

by repeating the above argument with $|\psi\rangle$ replaced with $\sigma_x|\psi\rangle$. This works similarly for the other two outcomes. Thus, if Alice communicates the result of the Bell basis measurement to Bob, he can apply the corresponding Pauli operator to obtain the input state $|\psi\rangle$. Alice needs 2 bits to describe which outcome occurred, and since each projected state has the same weight, the probability of every outcome is $1/4$. The fact that the probability distribution does not depend on the input state is important, otherwise information about the state would essentially leak into other degrees of freedom, and the state could not be properly reconstructed by Bob.

3.7 Complementarity

Complementarity of the particle and wave nature of light in the double slit experiment is one of the most well-known examples of the difference between classical and quantum mechanics. Indeed, Feynman[12] 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.[22]

3.7.1 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[13]-Zehnder[14] interferometer using polarizing beamsplitters (PBS), depicted in Fig. 3.2.

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 mode 0 and mode 1 in accord with Fig. 3.2. These two modes also form a two-dimensional state space $\mathscr{H}_M$ with basis states $|0\rangle_M$ and $|1\rangle_M$.

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

$$U_{\text{PBS}} \coloneqq \sum_{z=0}^{1} |z\rangle\langle z|_P \otimes |z\rangle_M. \tag{3.25}$$

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.

[12] Richard Phillips Feynman, 1918 – 1988, American physicist.
[13] Ludwig Mach, 1868 – 1951, Austrian inventor (son of physicist Ernst Mach).
[14] Ludwig Louis Albert Zehnder, 1854 – 1949, Swiss physicist.

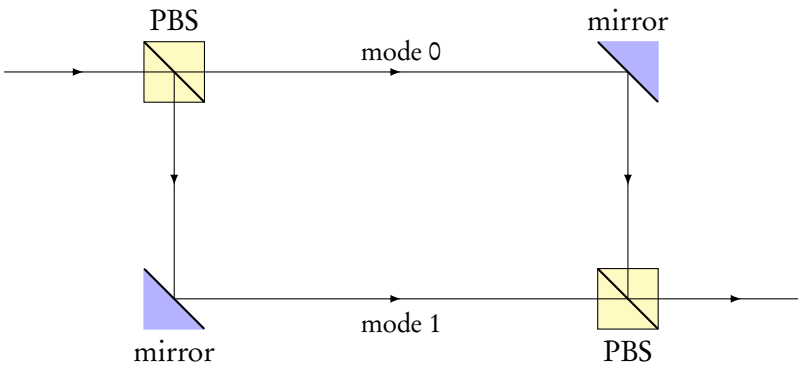


Figure 3.2: A Mach-Zehnder interferometer. The first polarizing beamsplitter (PBS) transmits input light into the two modes 0 and 1 according to its polarization state. The interferometer is constructed to that the propagating light acquires the same phase in either mode. Then the second PBS reverses the action of the first.

The action of the second polarizing beamsplitter is to reverse this process. 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 reproduce the input state, since $U_{\text{PBS}}^* U_{\text{PBS}} = \mathbb{1}_P$.

The polarization observable σ_z associated with horizontal and vertical is the "particle" property of the photon, since a photon in a definite state (eigenstate) of this observable takes a definite path through the interferometer. The first PBS is analogous to light being forced through the two slits in the double slit experiment. The observable σ_x associated with polarization at $\pm 45°$ is the "wave" property of the photon. A photon in a definite state of σ_x does not take a definite path through the interferometer (indeed, any $\sigma_{\hat{n}}$ for $\hat{n}$ in the $x-y$ plane defines a wave property), but produces a kind of "interference pattern" at the output. In an interference pattern, the intensity at any point depends on the relative phases of the paths superimposed at that point. Here, there are two paths and the relative phase at the output can be detected by measuring σ_x: The $+$ outcome signifies zero relative phase, $-$ a relative phase of π. The second PBS mimics the interference of the two beams in the double slit experiment, and a measurement of σ_x at the output is akin to the screen or film used to record the interference pattern.

Moreover, just as in the double slit experiment, if we try to determine the particle property, we inevitably destroy the wave property. Interference is possible only if no information about the path has been acquired. To see why, observe that the first PBS enables us to measure σ_z of the photon by measuring in which arm of the interferometer the photon is located. This is part of the *von Neumann picture of measurement*, which we shall examine in more detail in §4.2.1. 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 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

$$U_{\text{arm}} \coloneqq \sum_z |z\rangle\langle z|_M \otimes |\varphi_z\rangle_A, \tag{3.26}$$

where the $|\varphi_z\rangle_A$ are the two states of the atom produced in the measurement process. Altogether, the action of the interferometer and indirect measurement is described by the isometry

$$W_{P\to PA} \coloneqq U_{\text{PBS}}^* U_{\text{arm}} U_{\text{PBS}} = \sum_z |z\rangle\langle z|_P \otimes |\varphi_z\rangle_A \tag{3.27}$$

Now suppose the photon is initially polarized with a definite value of σ_x, say $+45°$, so that its quantum state is

$$|\psi_0\rangle_P = |+\rangle_P = \tfrac{1}{\sqrt{2}}(|0\rangle + |1\rangle)_P\,. \tag{3.28}$$

At the output, the experimental setup produces the state

$$|\psi_1\rangle_{PA} = W_{P\to PA}|\psi_0\rangle_P = \tfrac{1}{\sqrt{2}}\sum_{z=0}^{1}|z\rangle_P|\varphi_z\rangle_A. \tag{3.29}$$

We then measure σ_x on P, obtaining outcomes according to the distribution P_X. If the atomic states are identical, $|\varphi_0\rangle = |\varphi_1\rangle$, then P_X is deterministic: $P_X(+) = 1$. However, if the atomic states are orthogonal, $|\varphi_z\rangle = |z\rangle$, then P_X is uniform:

$$P_X(\pm) = \text{Tr}[(|\pm\rangle\langle\pm|_P \otimes \mathbb{1}_M)|\psi_1\rangle\langle\psi_1|_{PM}] = \tfrac{1}{2}. \tag{3.30}$$

The interference pattern is completely washed out, and one says the *coherence* has been destroyed, because there is no longer any way to determine the relative phase of the input state.

Notice that the output state $|\psi_1\rangle_{PM}$ in this case is maximally entangled; the loss of the photon's coherence is due to its entanglement with the atom. However, the coherence is still present, but only in the combined system. This follows because it is in principle possible to restore the photon's coherence, simply by applying the inverse isometry. This may be difficult in practice, but nothing prohibits it in principle.

One reason it may be difficult is that we must completely erase *all* information about the particle property, σ_z. And a more realistic description of the indirect measurement would include the fact that the measurement result is stored in *many* degrees of freedom, such as many magnetized spins in a region of a hard drive, not solely in one atom by itself. That is, the output state of a more realistic description of the indirect measurement is of the form

$$|\psi_1\rangle_{PA_1,\ldots,A_n} = \tfrac{1}{\sqrt{2}}(|0,\ldots,0\rangle + |1,\ldots,1\rangle)_{PA_1,\ldots,A_n}, \tag{3.31}$$

for some large n. For the present purposes, though, it suffices to consider $n = 2$. Even then, no unitary action on the photon polarization P and the first measurement record A_1 can restore the coherence and allow us to infer the relative phase of the input polarization state. This is easily seen by examining the *density operator* for the joint PA_1 system. Density operators will be examined in more detail in §4.1, but for the present purposes we can argue as follows.

Since we are ignoring A_2, we could imagine that someone measures it in the $|z\rangle$ basis, but does not tell us the measurement result. If the outcome were 0, then the state would be $|00\rangle_{PA_1}$, while outcome 1 leads to the state $|11\rangle_{PA_1}$. Each of these outcomes is equally likely. But this is precisely the same state of affairs that would result had the original polarization state had been $|-\rangle_P$. In that case, the state of PA_1A_2 would have been $\frac{1}{\sqrt{2}}(|000\rangle - |111\rangle)_{PA_1A_2}$, which indeed leads back to the same two equally-likely possible states conditioned on the measurement outcome, $|00\rangle$ and $|11\rangle$. Therefore,

there is no way to distinguish between the two possible original relative phases $|\pm\rangle_P$. In the language of density operators, coherence cannot be recovered because in either case the state of PA_1 is given by

$$\rho_{PA_1} = \mathrm{Tr}_{A_2}[|\psi_1\rangle\langle\psi_1|_{PA_1A_2}] = \tfrac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)_{PA_1}. \tag{3.32}$$

Interestingly, returning to the case $n = 1$, it is not actually necessary to apply the inverse of V to restore coherence. We can determine the relative phase of the input by comparing local measurements made on the photon and the atom separately. Observe that the maximally entangled state also takes the same form in the σ_x basis:

$$\tfrac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{PA} = \tfrac{1}{\sqrt{2}}(|++\rangle + |--\rangle)_{PA}. \tag{3.33}$$

Thus, the interferometer produces $|\psi_1^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{PA}$ when the input is $|\psi_0\rangle = |+\rangle$. On the other hand, if the input had been $|-\rangle$, an easy calculation shows the output would be

$$|\psi_1^-\rangle_{PA} = \tfrac{1}{\sqrt{2}}(|00\rangle - |11\rangle)_{PA} = \tfrac{1}{\sqrt{2}}(|+-\rangle + |-+\rangle)_{PA}. \tag{3.34}$$

Therefore, by measuring σ_x of the atom, the polarization state of the photon is transformed into either $|+\rangle$ or $|-\rangle$. The photon will then display interference in the form of a biased distribution P_X, and we can infer whether the input was $|+\rangle$ or $|-\rangle$ by comparing the measurements of the photon and the atom. If the measurement results are identical, the original phase was $+1$, if they differ, -1. This phenomenon is called the *quantum eraser*, as we have erased the original σ_z measurement record in the atom by measuring σ_x.

3.7.2 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 (which produces the random variable Z) by measuring the ancilla system to learn the value of z. From Exercise 3.2, when $P_Z(z) = \frac{1}{2}$, the probability of correctly guessing the outcome of the hypothetical measurement is given by $p_{\text{guess}}(Z|A) = \frac{1}{2}(1 + \sqrt{1 - |\langle\varphi_0|\varphi_1\rangle|^2})$. This motivates the definition of the *distinguishability* of the two paths by

$$D := \sqrt{1 - |\langle\varphi_0|\varphi_1\rangle|^2}. \tag{3.35}$$

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, we denote the probability of X given the input state $|\psi\rangle$ as $P_{X|\psi}$. Then the above motivates the definition of the *visibility* as

$$V := \max_{|\psi\rangle} |P_{X|\psi}(+) - P_{X|\psi}(-)|. \tag{3.36}$$

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.

Proposition 3.7.1: Wave–Particle Complementarity Relation

$$D^2 + V^2 = 1. \tag{3.37}$$

Proof. Supposing the input state is $|\psi\rangle_P = \sum_z \psi_z |z\rangle_P$ for $\psi_z \in \mathbb{C}$ such that $|\psi_0|^2 + |\psi_1|^2 = 1$, the total output state is just

$$|\psi'\rangle_{PA} = \sum_z \psi_z |z\rangle_P |\varphi_z\rangle_A. \tag{3.38}$$

Then we have

$$\begin{aligned}
\left|P_{X|\psi}(+) - P_{X|\psi}(-)\right| &= \left|\mathrm{Tr}[(|+\rangle\langle+|_P \otimes \mathbb{1}_A)|\psi'\rangle\langle\psi'|_{PA}] - \mathrm{Tr}[(|+\rangle\langle+|_P \otimes \mathbb{1}_A)|\psi'\rangle\langle\psi'|_{PA}]\right| \\
&= \left|\langle\psi'|(\sigma_x)_P \otimes \mathbb{1}_A|\psi'\rangle_{PA}\right| = \Big|\sum_{z,z'=0}^{1} \psi^*_{z'}\psi_z \langle z'|z \oplus 1\rangle\langle\varphi_{z'}|\varphi_z\rangle\Big| \\
&= \Big|\sum_{z=0}^{1} \psi_z \psi^*_{z\oplus 1}\langle\varphi_{z\oplus 1}|\varphi_z\rangle\Big| \leq |\psi_0\psi_1^*\langle\varphi_1|\varphi_0\rangle| + |\psi_1\psi_0^*\langle\varphi_0|\varphi_1\rangle| \\
&= 2|\psi_0\psi_1^*| \cdot |\langle\varphi_0|\varphi_1\rangle| \leq |\langle\varphi_0|\varphi_1\rangle|.
\end{aligned} \tag{3.39}$$

The first inequality is the triangle inequality for complex numbers, while the second is the fact that $|\psi_0\psi_1^*| \leq \frac{1}{2}$. 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 $0 \leq p \leq 1$ and two arbitrary angles θ_0 and θ_1. Thus $|\psi_0\psi_1^*| = |\sqrt{p}\sqrt{1-p}| \leq \frac{1}{2}$.

Choosing $\psi_0 = \psi_1 = \frac{1}{\sqrt{2}}$ saturates this bound, and therefore $V = |\langle\varphi_0|\varphi_1\rangle|$. From the expression for the distinguishability, $|\langle\varphi_0|\varphi_1\rangle|^2 = 1 - D^2$, completing the proof. □

3.8 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 what we mean by "real properties" in the first place. A very concise notion is given by Einstein[15], Podolsky[16], and Rosen[17] (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.[23]

[15] Albert Einstein, 1879 – 1955, German-born theoretical physicist and philosopher of science.
[16] Boris Yakovlevich Podolsky, 1896 – 1966, Russian-American physicist.
[17] Nathan Rosen, 1909 – 1995, American-Israeli physicist.

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 two different expansions of a given bipartite state

$$|\Psi\rangle_{AB} = \sum_k |\psi_k\rangle_A \otimes |u_k\rangle_B = \sum_s |\varphi_s\rangle_A \otimes |v_s\rangle_B, \tag{3.40}$$

where the $|\psi_k\rangle$ and $|\varphi_s\rangle$ are arbitrary states, while the $|u_k\rangle$ and $|v_s\rangle$ are unnormalized, but mutually orthogonal states. According to the postulates, 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, by measuring system B in yet another basis, still other post-measurement states can be prepared in 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.[24]

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 does not know the measurement result, and thus the probability of any particular experiment she might do is unchanged by the fact that Bob has or has not measured system B. For a measurement with projection operators Π_x, Alice anticipates that outcome x will occur with probability $P_X(x) = \mathrm{Tr}[((\Pi_x)_A \otimes \mathbb{1}_B)|\Psi\rangle\langle\Psi|_{AB}]$ in either case. Since the probability distribution 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, recall the Robertson[18] uncertainty relation,

$$\Delta A \Delta B \geq \tfrac{1}{2} |\mathrm{Tr}[[A,B]|\psi\rangle\langle\psi|]|, \tag{3.41}$$

for ΔA (ΔB)) the standard deviation of observable A (B) in state $|\psi\rangle$ and $[A,B]$ the commutator. It implies that noncommuting observables cannot simultaneously take on well-defined values, in contradiction to the EPR argument. 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

[18] Howard Percy Robertson, 1903 – 1961, American mathematician and physicist.

> 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.[19]

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 *locally realistic*. If a locally realistic description of quantum phenomena is possible, it is not to be 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.

3.9 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.

[19] "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."[25, 26]

3.9.1 The CHSH inequality

A simplified version of Bell's argument was put forth by Clauser[20], Horne[21], Shimony[22], and Holt[23], 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 that a_0 and a_1 (and b_0 and b_1) cannot be simultaneously measured (this is an instance of the reasoning from counterfactuals described above). From this, it immediately follows that

$$C = (a_0 + a_1)b_0 + (a_0 - a_1)b_1 = \pm 2. \tag{3.42}$$

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

$$P(a_0 = \mathsf{a}_0, a_1 = \mathsf{a}_1, b_0 = \mathsf{b}_0, b_1 = \mathsf{b}_1 | \lambda) P_{\mathrm{HV}}(\lambda). \tag{3.43}$$

But since (3.42) is an equality, averaging over λ like so will only lead to

$$|\langle C \rangle| = |\langle a_0 b_0 \rangle + \langle a_1 b_0 \rangle + \langle a_0 b_1 \rangle - \langle a_1 b_1 \rangle| \leq 2. \tag{3.44}$$

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

$$\langle \Psi | (\vec{\sigma}_A \cdot \hat{a})(\vec{\sigma}_B \cdot \hat{b}) | \Psi \rangle_{AB} = -\hat{a} \cdot \hat{b}. \tag{3.45}$$

To see this, compute

$$(\vec{\sigma}_A \cdot \hat{a})(\vec{\sigma}_B \cdot \hat{b}) |\Psi\rangle_{AB} = \sum_{jk} a_j b_k (\sigma_j \otimes \sigma_k) |\Psi\rangle_{AB} = -\sum_{jk} a_j b_k (\mathbb{1} \otimes \sigma_k \sigma_j) |\Psi\rangle_{AB}. \tag{3.46}$$

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 $\mathbb{1} \otimes \sigma_k |\Psi\rangle_{AB}$ 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

$$\langle a_0 b_0 \rangle = \langle a_1 b_0 \rangle = \langle a_0 b_1 \rangle = -\tfrac{1}{\sqrt{2}} \qquad \text{and} \qquad \langle a_1 b_1 \rangle = \tfrac{1}{\sqrt{2}}, \tag{3.47}$$

[20] John Francis Clauser, born 1942, American physicist.
[21] Michael A. Horne, American physicist
[22] Abner Shimony, born 1928, American physicist and philosopher of science.
[23] Richard A. Holt, American physicist.

so that $|\langle C\rangle| = 2\sqrt{2} \not\leq 2$. Therefore, Einstein's goal of a locally realistic version of quantum mechanics is impossible.

However, it is important to point out that this does not immediately rule out all "classical" descriptions of the whole experiment, just those most similar to our description of classical mechanical systems. The locally realistic model leaves out several aspects of the experiment, presuming them to be irrelevant. One is the fact that Alice and Bob need to transmit their measurement results to a common location in order to compare them. Another is that they need to have access to synchronized reference frames in order to properly carry out their measurements. Perhaps by including these additional features one can recover a more classical understanding of the CHSH experiment.

The use of entangled states is necessary in the CHSH 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.[24]

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[24] put it best in the title of a paper on the subject of Bell inequalities: *Unperformed experiments have no results.*[27]

3.9.2 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[25] 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

$$C^2 = 4\mathbb{1} - [a_0, a_1][b_0, b_1]. \tag{3.48}$$

Now compute the infinity norm of C^2, which is defined by

$$||C^2||_\infty := \sup_{|\psi\rangle}\left(\frac{||C^2|\psi\rangle||}{|||\psi\rangle||}\right). \tag{3.49}$$

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

$$||C^2||_\infty = ||4\mathbb{1} - [a_0, a_1][b_0, b_1]||_\infty \leq 4 + ||[a_1, a_0]||_\infty + ||[b_0, b_1]||_\infty \tag{3.50}$$

$$\leq 4 + ||a_1||_\infty (||a_0||_\infty + ||-a_0||_\infty) + ||b_0||_\infty (||b_1||_\infty + ||-b_1||_\infty) = 8. \tag{3.51}$$

In the last step we used the fact that $||\pm c||_\infty = 1$ for c having eigenvalues ± 1.

[24] Asher Peres, 1934 – 2005, Israeli physicist,
[25] Boris Semyonovich Tsirelson, Russian-Israeli mathematician.

3.9.3 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

$$\sum_{xy} a_x \oplus b_y = 2 \qquad \text{while} \qquad \sum_{xy} x \cdot y = 1. \tag{3.52}$$

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

$$P(a \oplus b = x \cdot y) = \tfrac{1}{4} \sum_{a,b,x,y} \delta_{a \oplus b = x \cdot y} \sum_{\lambda} p(\lambda) P(a|x,\lambda) P(b|y,\lambda) \leq \tfrac{3}{4}, \tag{3.53}$$

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

$$\langle a'_0 b'_1 \rangle = \langle (-1)^{a_0 + b_1} \rangle = p_{01} - (1 - p_{01}) = 2p_{01} - 1, \tag{3.54}$$

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

$$|\langle C \rangle| = 2 \sum_{xy} p_{xy} - 4 = 2 \cdot 4 p_{\text{win}}^{\text{DLHV}} - 4 \leq 2, \tag{3.55}$$

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.

3.10 Notes & Further reading

The axioms of quantum mechanics were laid out by Dirac in 1930 [28] and von Neumann in 1932 [29]. A more recent and very lucid treatment of quantum mechanics as a whole is given by Ballentine [30]; more mathematical treatments are offered by Hall [31] and Takhtajan [32]. Here we follow the treatment by Nielsen and Chuang [2], tailored to quantum information theory's focus on finite-dimensional systems.

For more on interpretations of quantum mechanics, the reader could do worse than consulting the Stanford Encyclopedia of Philosophy, the Compendium of Quantum Physics by Greenberger, Hentschel, and Weinert [33], or the book of Hughes [34]. The recent books of Jaeger [35] and Timpson [36] discuss issues of interpretation in the context of quantum information theory.

The wave-particle complementarity relation is adapted from Englert [37].

3.11 Exercises

Exercise 3.1. Hadamard gate → solution

An important qubit transformation in quantum information theory is the Hadamard gate. In the basis of $\sigma_{\hat{z}}$, it takes the form

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

That is to say, if $|0\rangle$ and $|1\rangle$ are the $\sigma_{\hat{z}}$ eigenstates, corresponding to eigenvalues $+1$ and -1, respectively, then

$$H = \frac{1}{\sqrt{2}}(|0\rangle\langle 0| + |0\rangle\langle 1| + |1\rangle\langle 0| - |1\rangle\langle 1|).$$

a) Show that H is unitary.

b) What are the eigenvalues and eigenvectors of H?

c) What form does H take in the basis of $\sigma_{\hat{x}}$? $\sigma_{\hat{y}}$?

d) Give a geometric interpretation of the action of H in terms of the Bloch sphere.

Exercise 3.2. State distinguishability → solution

One way to understand the cryptographic abilities of quantum mechanics is from the fact that non-orthogonal states cannot be perfectly distinguished.

a) In the course of a quantum key distribution protocol, suppose that Alice randomly chooses one of the following two states and transmits it to Bob:

$$|\phi_0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \qquad \text{or} \qquad |\phi_1\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle).$$

Eve intercepts the qubit and performs a measurement to identify the state. The measurement consists of the orthogonal states $|\psi_0\rangle$ and $|\psi_1\rangle$, and Eve guesses the transmitted state was $|\phi_0\rangle$ when she obtains the outcome $|\psi_0\rangle$, and so forth. What is the probability that Eve correctly guesses the state, averaged over Alice's choice of the state for a given measurement? What is the optimal measurement Eve should make, and what is the resulting optimal guessing probability?

b) Now suppose Alice randomly chooses between two states $|\phi_0\rangle$ and $|\phi_1\rangle$ separated by an angle θ on the Bloch sphere. What is the measurement which optimizes the guessing probability? What is the resulting probability of correctly identifying the state, expressed in terms of θ? In terms of the states?

Exercise 3.3. Fidelity → solution

a) Given a qubit prepared in a completely unknown state $|\psi\rangle$, what is the *fidelity* F^2 of a random guess $|\phi\rangle$, where $F(|\phi\rangle, |\psi\rangle)^2 = |\langle\phi|\psi\rangle|^2$? The fidelity (squared) can be thought of as the probability that an input state (the guess) $|\phi\rangle$ passes the "ψ" test, which is the measurement in the basis $|\psi\rangle$, $|\psi^\perp\rangle$.

b) In order to improve the guess, we might make a measurement of the qubit, say along the $\hat{z}$ axis. Given the result $k \in \{0, 1\}$, our guess is then the state $|k\rangle$. What is the average fidelity of the guess after the measurement, i.e. the probability of passing the "ψ" test?

Exercise 3.4. Indirect measurement → solution

Suppose a quantum system is prepared in one of two nonorthogonal states $|\phi_1\rangle$ or $|\phi_2\rangle$. We would like to make a measurement to determine which state was prepared, but do so without disturbing the state. To this end, we could consider making an indirect measurement in which we also prepare an auxiliary state $|\text{blank}\rangle$, apply a unitary U_{AB} which has the action

$$|\phi_j\rangle_A|\text{blank}\rangle_B \to U_{AB}|\phi_j\rangle_A|\text{blank}\rangle_B = |\phi_j\rangle_A|\beta_j\rangle_B,$$

and then measure system B in some way. This scheme evidently does not disturb the state of system A. What is the most we can learn about which state was prepared? What if the two states $|\phi_j\rangle$ are orthogonal?

Exercise 3.5. Broken measurement → solution

Alice and Bob share a state $|\Psi\rangle_{AB}$, and Bob would like to perform a measurement described by projectors Π_j on his part of the system, but unfortunately his measurement apparatus is broken. He can still perform arbitrary unitary operations, however. Meanwhile, Alice's measurement apparatus is in good working order. Show that there exist projectors Π'_j and unitaries U_j and V_j so that

$$|\Psi_j\rangle = \left(\mathbb{1}\otimes\Pi_j\right) = \left(U_j\otimes V_j\right)\left(\Pi'_j\otimes\mathbb{1}\right)|\Psi\rangle.$$

(Note that the state is unnormalized, so that it implicitly encodes the probability of outcome j.) Thus Alice can assist Bob by performing a related measurement herself, after which they can locally correct the state.
Hint: Work in the Schmidt basis of $|\Psi\rangle$.

Exercise 3.6. Remote copy → solution

Alice and Bob would like to create the state $|\Psi\rangle_{AB} = a|00\rangle_{AB} + b|11\rangle_{AB}$ from Alice's state $|\psi\rangle_A = a|0\rangle_A + b|1\rangle_A$, a "copy" in the quantum-mechanical sense. Additionally, they share the canonical entangled state $|\Phi\rangle$. Can they create the desired state by performing only local operations (measurements and unitary operators), provided Alice can only send *one* bit of classical information to Bob?

Exercise 3.7. Measurements on a bipartite state → solution

Consider a 2-qubit Hilbert space $\mathscr{H}_{AB} = \mathscr{H}_A\otimes\mathscr{H}_B$ with basis $\{|00\rangle, |01\rangle|10\rangle, |11\rangle\}$ in the Bell state

$$|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B). \tag{3.56}$$

Two parties, called Alice and Bob, get half of the state $|\Phi\rangle$ so that Alice has qubit A and Bob has qubit B. Alice then performs a measurement $\mathscr{M}_A^\theta := \{|\theta\rangle\langle\theta|, |\tfrac{\pi}{2}-\theta\rangle\langle\tfrac{\pi}{2}-\theta|\}$, with $|\theta\rangle := \cos\theta|0\rangle + \sin\theta|1\rangle$, on her qubit.

a) What description does Alice give to system B, given the outcome of her measurement?

b) If Bob performs the measurement $\mathscr{M}_B^0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}$ on B, what is the probability distribution for his outcomes? How would Alice describe his probability distribution?

Exercise 3.8. The Hilbert-Schmidt inner product → solution

Suppose R and Q are two quantum systems with the same Hilbert space. Let $\{|i\rangle_R\}_i$ and $\{|i\rangle_Q\}_i$ be two orthonormal basis sets for R and Q. Let A be an operator on R and B an operator on Q. Define $|\Omega\rangle = \sum_i |i\rangle_R|i\rangle_Q$.

a) Show that $A \otimes \mathbb{1}|\Omega\rangle = \mathbb{1} \otimes A^T|\Omega\rangle$.

b) Show that $\mathrm{Tr}(A^T B) = \langle\Omega|A \otimes B|\Omega\rangle$. This is the Hilbert-Schmidt inner product of operators on R (or Q), and the result shows that it can be thought of as the inner product between the states $A \otimes \mathbb{1}|\Omega\rangle$ and $\mathbb{1} \otimes B|\Omega\rangle$.

Exercise 3.9. Teleportation redux → solution

a) Show that for the canonical entangled state $|\Phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ and any unitary operator U

$$\left(U_A \otimes \overline{U}_B\right)|\Phi\rangle_{AB} = |\Phi\rangle_{AB},$$

where $\overline{U}$ denotes complex conjugation in the $|0\rangle, |1\rangle$ basis.

b) Show that for any state $|\psi\rangle$

$${}_A\langle\psi|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}|\overline{\psi}\rangle_B.$$

c) What happens if Alice and Bob use the state $(\mathbb{1}_A \otimes U_B)|\Phi\rangle_{AB}$ for teleportation? Or if Alice measures in the basis $U^*_{A'}|\Phi_j\rangle_{A'A}$?

d) Instead of a single system state $|\psi\rangle_{A'}$, Alice has a bipartite state $|\psi\rangle_{A_1A_2}$. What happens if she performs the teleportation protocol on system A_2?

Exercise 3.10. "All-or-nothing" violation of local realism → solution

Consider the three qubit state $|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}(|000\rangle - |111\rangle)_{123}$, the Greenberger-Horne-Zeilinger state.

a) Show that $|\mathrm{GHZ}\rangle$ is a simultaneous eigenstate of $X_1Y_2Y_3$, $Y_1X_2Y_3$, and $Y_1Y_2X_3$ with eigenvalue $+1$, where X and Y are the corresponding Pauli operators.

b) Use the results of part (a) to argue by Einstein locality that each qubit has well-defined values of X and Y. For qubit j, denote these values by x_j and y_j. We say that these values are *elements of reality*. What would local realism, i.e. the assumption of realistic values that are undisturbed by measurements on other qubits, predict for the product of the outcomes of measurements of X on each qubit?

c) What does quantum mechanics predict for the product of the outcomes of measurements of X on each qubit?

4

Quantum States, Measurements, and Channels

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 we 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. Nevertheless, we will also find that the more general notions of quantum states, measurements, and dynamics are consistent with simply treating a non-isolated system as part of a larger isolated system and applying the postulates.

4.1 Quantum states

4.1.1 Mixtures of states

Consider a quantum system $\mathscr{H}_A$ whose state depends on a classical value (random variable) Z and let $|\phi_z\rangle\langle\phi_z|_A \in \mathscr{S}(\mathscr{H}_A)$ 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 $\{P_Z(z), |\phi_z\rangle\}$.

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

$$P_{X|Z=z}(x) = \mathrm{Tr}[\Pi_x U_A |\phi_z\rangle\langle\phi_z|_A U_A^*]. \tag{4.1}$$

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

$$P_X(x) = \sum_z P_Z(z) P_{X|Z=z}(x). \tag{4.2}$$

By linearity, this can be rewritten as

$$P_X(x) = \mathrm{Tr}[\Pi_x U_A \rho_A U_A^*]. \tag{4.3}$$

where we have implicitly defined

$$\rho_A = \sum_z P_Z(z) |\phi_z\rangle\langle\phi_z|_A. \tag{4.4}$$

Observe that ρ_A is positive, $\rho_A \geq 0$, and has $\mathrm{Tr}[\rho_A] = 1$. These two conditions are all we really need to get sensible results from the Born probability rule $P_X(x) = \mathrm{Tr}[\Pi_x \rho]$, and operators satisfying these conditions are termed *density operators*. We can then generalize the quantum state postulate as follows.

Definition 4.1.1: Quantum states

The state of a system associated with Hilbert space $\mathscr{H}$ is given by a *density operator*, a linear operator ρ on $\mathscr{H}$ satisfying

$$\rho \geq 0, \tag{4.5}$$
$$\mathrm{Tr}[\rho] = 1. \tag{4.6}$$

We denote the set of density operators by $\mathscr{S}(\mathscr{H})$; observe that it is convex. 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 $\mathrm{Tr}\rho^2 = 1$. They are extreme points in the set $\mathscr{S}(\mathscr{H})$. States with more than one nonzero eigenvalue are called *mixed states* since they are mixtures (convex combinations) of their eigenvectors.

Alternatively, expression (4.3) can be obtained by applying the postulates of Section 3.1 directly to the density operator ρ_A defined above. In particular, by replacing $|\phi\rangle\langle\phi|$ with the density operator ρ in (3.1). In other words, from the point of view of an observer with no access to Z, the situation is consistently characterized by ρ_A.

4.1.2 "Classical" states

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 Z distributed according to P_Z, the associated density operator is

$$\rho_Z = \sum_{z \in \mathscr{Z}} P_Z(z) |b_z\rangle\langle b_z|, \tag{4.7}$$

for some orthonormal basis $\{|b_z\rangle\}$. When the state of a different system A depends on the value of Z, the overall state is called a classical-quantum state.

Definition 4.1.2: Classical-quantum states

A state ρ_{ZA} is called a classical-quantum (CQ) state with Z classical if it is of the form

$$\rho_{ZA} = \sum_z P_Z(z) |b_z\rangle\langle b_z|_Z \otimes (\varphi_z)_A, \tag{4.8}$$

with $\{|b_z\rangle\}_z$ a family of orthonormal vectors on $\mathscr{H}_Z$ and φ_z a set of arbitrary density operators.

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:

$$\sigma = \sum_k p_k \rho_k \otimes \varphi_k. \tag{4.9}$$

Any quantum state of the form (4.9) is called *separable* and any state which is not separable is said to be *entangled*.

4.1.3 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 $\mathscr{H}_A \otimes \mathscr{H}_B$ is that, to an observer with no access to B, the state of A does not correspond to a fixed vector $|\phi\rangle \in \mathscr{H}_A$, 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 $|\Psi\rangle \in \mathscr{H}_A \otimes \mathscr{H}_B$. The expectation value of O_A is given by

$$\langle O_A \rangle_\Psi = \mathrm{Tr}[O_A \otimes \mathbb{1}_B |\Psi\rangle\langle\Psi|] = \mathrm{Tr}[O_A \mathrm{Tr}_B[|\Psi\rangle\langle\Psi|]], \tag{4.10}$$

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:

$$\begin{aligned}\langle O_A \rangle_\Psi &= \langle\Psi|(O_A \otimes \mathbb{1}_B)|\Psi\rangle = \sum_{jk} \lambda_j \lambda_k \langle\xi_j| \otimes \langle\eta_j|(O_A \otimes \mathbb{1}_B)|\xi_k\rangle \otimes |\eta_k\rangle \\ &= \sum_{jk} \lambda_j \lambda_k \langle\xi_j|O_A|\xi_k\rangle\langle\eta_j|\eta_k\rangle = \sum_k \lambda_k^2 \langle\xi_k|O_A|\xi_k\rangle = \mathrm{Tr}[O_A \sum_k \lambda_k^2 |\xi_k\rangle\langle\xi_k|].\end{aligned} \tag{4.11}$$

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 (4.5) and (4.6) and is therefore a density operator. Moreover, since the $|\xi\rangle$ are orthonormal, this expression is in fact the eigendecomposition of ρ_A.

4.1.4 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. Formally, we make the following definition.

Definition 4.1.3: Purification

Given a density operator ρ_A on Hilbert space $\mathscr{H}_A$, a purification $|\Psi\rangle_{AB}$ is a state on $\mathscr{H}_A \otimes \mathscr{H}_B$ for some $\mathscr{H}_B$ such that

$$\rho_A = \mathrm{Tr}_B[|\Psi\rangle\langle\Psi|_{AB}] \tag{4.12}$$

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". It is an instance of a dilation as described in §2.5.3.

Given an ensemble decomposition of a density operator $\rho = \sum_{z=1}^{n} P_Z(z)|\phi_z\rangle\langle\phi_z|$ as in (4.4), it is easy to construct a purification of ρ. Simply invent an additional system B of dimension at least n and define

$$|\Psi\rangle_{AB} = \sum_{z=1}^{n} \sqrt{P_Z(z)}|\phi_z\rangle_A \otimes |b_z\rangle_B, \tag{4.13}$$

where $|b_z\rangle_B$ is a basis for $\mathscr{H}_B$. This also works for CQ states, like $\rho_{AZ} = \sum_z P_Z(z)(\rho_z)_A \otimes |b_z\rangle\langle b_z|_Z$ in (4.8). Now invent two additional systems B and Z' and define

$$|\Psi\rangle_{ABZZ'} = \sum_z \sqrt{P_Z(z)}|\varphi_z\rangle_{AB} \otimes |b_z\rangle_Z \otimes |b_z\rangle_{Z'}, \tag{4.14}$$

where $|\varphi_z\rangle_{AB}$ is a purification of $(\rho_z)_A$. System Z' is identical to Z and is responsible for making the AZ state classical. Equally well, then, the state of AZ' is a CQ state, with the classical information stored in Z'. Thus, classical information is in a sense information that can be, and indeed has already been, copied to another system.

The Schmidt decomposition of a purification $|\Psi\rangle_{AB}$ of ρ_A is directly related to the eigendecomposition of ρ itself as we saw in (4.11). Moreover, the Schmidt decomposition immediately implies that any two purifications of a state ρ must be related by a partial isometry connecting the respective purifying systems. Suppose $|\Psi\rangle_{AB}$ and $|\Psi'\rangle_{AB'}$ are two purifications of ρ_A, with $\dim(\mathscr{H}_A) = d$ and $\dim(\mathscr{H}_B') \geq \dim(\mathscr{H}_B) \geq d$. Note that $\mathscr{H}_B$ is free to have dimension larger than d and indeed must have as in (4.13). In view of the relation to the eigendecomposition, the Schmidt forms of the two states must be

$$|\Psi\rangle_{AB} = \sum_{k=1}^{d} \sqrt{p_k}|\xi_k\rangle_A \otimes |\eta_k\rangle_B \qquad \text{and} \qquad |\Psi'\rangle_{AB'} = \sum_{k=1}^{d} \sqrt{p_k}|\xi_k\rangle_A \otimes |\eta_k'\rangle_{B'}. \tag{4.15}$$

Both $\{|\eta_k\rangle\}$ and $\{|\eta_k'\rangle\}$ are orthonormal bases for $\mathscr{H}_B$ and $\mathscr{H}_B'$. Therefore, the operation $V_{B\to B'}$ defined by $V_{B\to B'}|\eta_j\rangle_B = |\eta_j'\rangle_{B'}$ is a partial isometry since $V^*V = \mathbb{1}$. This means it preserves the inner product; see §A.1. Therefore, we have shown

Proposition 4.1.1: Unitary relation of purifications

For any two purifications $|\Psi\rangle_{AB}$ and $|\Psi'\rangle_{AB'}$ of a state ρ_A with $\dim(\mathscr{H}_B) \leq \dim(\mathscr{H}_{B'})$, there exists a partial isometry $V_{B\to B'}$ such that $|\Psi'\rangle_{AB'} = (\mathbb{1}_A \otimes V_{B\to B'})|\Psi\rangle_{AB}$. By embedding $\mathscr{H}_B$ into $\mathscr{H}_{B'}$, the partial isometry can be extended to a unitary.

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 (4.4) 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 $|b_k\rangle$ with equal probability. However, the decomposition is not unique, as the same state could be written as

$$\rho_A = \frac{1}{2}|\tilde{b}_0\rangle\langle\tilde{b}_0| + \frac{1}{2}|\tilde{b}_1\rangle\langle\tilde{b}_1| \tag{4.16}$$

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

$$\rho = \sum_{k=1}^{n} p_k|\phi_k\rangle\langle\phi_k| = \sum_{j=1}^{m} q_j|\psi_j\rangle\langle\psi_j| \tag{4.17}$$

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

$$|\Psi_1\rangle_{AB} = \sum_{k=1}^{n} \sqrt{p_k}|\phi_k\rangle_A \otimes |b'_k\rangle_B \qquad \text{and} \qquad |\Psi_2\rangle_{AB} = \sum_{j=1}^{m} \sqrt{q_j}|\psi_j\rangle_A \otimes |b'_j\rangle_B. \tag{4.18}$$

Here we have chosen $\mathscr{H}_B$ such that its dimension is the greater of n and m. As these pure states are purifications of the same density operator, there must be a unitary U such that $\mathbb{1}_A \otimes U_B|\Psi_1\rangle_{AB} = |\Psi_2\rangle_{AB}$. But then we have

$$\begin{aligned}\sqrt{q_k}|\psi_k\rangle &= \sum_j \sqrt{q_j}|\psi_j\rangle\langle b'_k|b'_j\rangle = {}_B\langle b'_k|\Psi_2\rangle_{AB} \\ &= \sum_j \sqrt{p_j}|\phi_j\rangle\langle b'_k|U|b'_j\rangle = \sum_j U_{kj}\sqrt{p_j}|\phi_j\rangle.\end{aligned} \tag{4.19}$$

Thus, we have shown the following statement.

Proposition 4.1.2: 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

$$\sqrt{q_k}|\psi_k\rangle = \sum_j U_{kj}\sqrt{p_j}|\phi_j\rangle. \tag{4.20}$$

Moreover, this argument establishes that that any particular ensemble decomposition can be realized by appropriate measurement of the purifying system. That is, measuring system B of (4.18) with projectors $\Pi_k = |b_k\rangle\langle b_k|$ produces the CQ state

$$\rho_{AZ} = \sum_{z=1}^{n} p_z|\phi_z\rangle\langle\phi_z|_A \otimes |b_z\rangle\langle b_z|_Z. \tag{4.21}$$

But measuring system B with projectors $\Pi_k = U^*|b_k\rangle\langle b_k|U$ produces the CQ state

$$\rho'_{AZ} = \sum_{z=1}^{n} q_z |\psi_z\rangle\langle\psi_z|_A \otimes |b_z\rangle\langle b_z|_Z. \tag{4.22}$$

Since all ensemble decompositions are related by unitaries, any of them can be obtained in this way. This phenomenon is sometimes called *steering*: someone with access to the purifying system can steer the state of A into any one of the possible ensembles. Steering provides more evidence for the fact that one cannot attach any physical importance to a particular ensemble decomposition of a system.

4.1.5 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 pure states. Pure states are pure and 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.

4.2 Generalized measurements

We have seen in the previous section that, as long as we are only interested in the observable quantities of subsystem $\mathcal{H}_A$ of a larger state space $\mathcal{H}_A \otimes \mathcal{H}_B$, 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 $\mathcal{H}_A$ and $\mathcal{H}_B$. The basic mathematical objects to be introduced in this context are *completely positive maps (CPMs)* and *positive operator valued measures (POVMs)*.

4.2.1 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 [29].

We have said that measurements are described by a set of projection operators $\{\Pi_x\}$, one Π_x for every outcome x. Given a state ρ, we saw in §4.1 that the xth outcome occurs with probability

$P_X(x) = \mathrm{Tr}[\Pi_x \rho]$. 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 $\rho = \sum_z P_Z(z)|\phi_z\rangle\langle\phi_z|$ for some pure states $|\phi_z\rangle$. For each z, the measurement produces the state $|\psi_{x,z}\rangle = \Pi_x|\phi_z\rangle / \sqrt{\langle\phi_z|\Pi_x|\phi_z\rangle}$ with probability $P_{X|Z=z}(x) = \langle\phi_z|\Pi_x|\phi_z\rangle$. From the discussion in §4.1.1, 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

$$\begin{aligned}\rho_x &= \sum_z P_{Z|X=x}(z)|\psi_{x,z}\rangle\langle\psi_{x,z}| = \sum_z \frac{P_{Z|X=x}(z)}{P_{X|Z=z}(x)}\Pi_x|\phi_z\rangle\langle\phi_z|\Pi_x \\ &= \sum_z \frac{P_Z(z)}{P_X(x)}\Pi_x|\phi_z\rangle\langle\phi_z|\Pi_x = \frac{1}{P_X(x)}\Pi_x\rho\Pi_x. \end{aligned} \tag{4.23}$$

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

The above calculation is only for one outcome x, but of course the measurement produces an entire ensemble of states, ρ_x with probability $P_X(x)$. We may describe the ensemble of post-measurement states and their probabilities with the cq state

$$\rho'_{XQ} = \sum_x P_X(x)|b_x\rangle\langle b_x|_X \otimes (\rho_x)_Q, \tag{4.24}$$

where Q is the original quantum system and X is a system that stores the measurement result. To an observer without access to the measurement result, the description of the state after the measurement is given by

$$\rho'_Q = \sum_x P_X(x)\rho_x = \sum_x \Pi_x\rho\Pi_x = \mathrm{Tr}_X[\rho'_{XQ}]. \tag{4.25}$$

This is often called a *non-selective* measurement. Generally, $\rho'_Q \neq \rho_Q$: 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 (4.25). Does it have any physical meaning? A purification is given by

$$|\Psi\rangle_{QX} = \sum_x \Pi_x|\phi\rangle_Q \otimes |b_x\rangle_X. \tag{4.26}$$

The interesting thing is that we can describe the transformation

$$|\phi\rangle_Q \otimes |b_0\rangle_X \mapsto |\Psi\rangle_{QX} \tag{4.27}$$

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

$$UU^* = \sum_{xx'} \Pi_{x'}\Pi_x^* \otimes V_{x'}V_x^* = \sum_x \Pi_x \otimes V_x V_x^* = \sum_x \Pi_x \otimes \mathbb{1} = \mathbb{1} \otimes \mathbb{1}. \tag{4.28}$$

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 (Q) and a system

storing the measurement result (X). Actually, this procedure does more than correlate Q and X, it *entangles* them.

The measurement process is not quite finished though, since $|\Psi\rangle_{QX}$ describes a coherent superposition of *all* possible outcomes. To realize a particular outcome, we have to assume that X is somehow itself measured in the $\{|b_x\rangle\}$ basis. So how does this really solve the measurement problem? In order to measure X, we need to correlate it with X', and then we will need to measure X', requiring correlation with X'', 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 (4.26), and

$$|\xi\rangle_{QXX'} = \sum_x \Pi_x|\phi\rangle_Q \otimes |b_x\rangle_X \otimes |b_x\rangle_{X'} \tag{4.29}$$

if taken to the next step. In the former case, tracing out system X leaves the density operator $\rho'_Q = \sum_x \Pi_x|\phi\rangle\langle\phi|\Pi_x$, while in the latter case tracing out X' leaves the classical-quantum state $\rho'_{QX} = \sum_x \Pi_x|\phi\rangle\langle\phi|\Pi_x \otimes |b_x\rangle\langle b_x|$.

4.2.2 Mixtures of measurements & POVMs

Measurements can themselves be "mixed" in the way we saw quantum states can be mixed in §4.1.1. 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 $\{\Pi_0 = |0\rangle\langle 0|, \Pi_1 = |1\rangle\langle 1|\}$ or $\{\Pi_\pm = |\pm\rangle\langle\pm|\}$ 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 $\Pi_{x,y}$ the corresponding projector, then the probability distribution when the state is ρ is given by

$$P_{XY}(x,y) = \tfrac{1}{2}\mathrm{Tr}[\Pi_{x,y}\rho] = \mathrm{Tr}[\Lambda_{x,y}\rho]. \tag{4.30}$$

Here we have implicitly defined the operators $\Lambda_{x,y}$. Observe that these sum to $\mathbb{1}$, just as we insisted for any projective measurement, although they are no longer disjoint.

Just as with quantum states, this example suggests that we should allow arbitrary operators $\{\Lambda_x\}_{x\in\mathscr{X}}$ as long as they yield sensible results in the probability rule $P_X(x) = \mathrm{Tr}[\Lambda_x\rho]$. For this we need the Λ_x to be positive and sum to identity. We thus modify the quantum measurement postulate as follows

Definition 4.2.1: Quantum measurements

The outcome of a measurement on a quantum system is a random variable X on the set of outcomes $\mathscr{X}$. Each particular outcome x is associated with an operator Λ_x, the set of which must obey the conditions

$$\Lambda_x \geq 0 \qquad \forall x \in \mathscr{X}, \tag{4.31}$$

$$\sum_{x\in\mathscr{X}} \Lambda_x = \mathbb{1}. \tag{4.32}$$

The probability of outcome x for quantum state ρ is given by $P_X(x) = \mathrm{Tr}[\Lambda_x\rho]$.

Such a set of Λ_x 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.

4.2.3 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*, another instance of a dilation from §2.5.3.

In fact, we have already met the Naimark extension in §4.2.1: One method of realizing a POVM is the von Neumann approach. For a set of projectors Π_x we saw that $U_{AB} = \sum_x (\Pi_x)_A \otimes (V_x)_B$ is a unitary operator taking $|\psi\rangle_A \otimes |0\rangle_B$ to $|\psi'\rangle_{AB}$ such that measuring B with $(\Pi_x)_B$ realizes measurement of $(\Pi_x)_A$ on A. To extend this to an arbitrary POVM with elements Λ_x, define U_{AB} implicitly by the action

$$U_{AB}|\psi\rangle_A \otimes |0\rangle_B = \sum_x \sqrt{\Lambda_x}|\psi\rangle_A \otimes |b_x\rangle_B = |\psi'\rangle_{AB}. \tag{4.33}$$

The probability of outcome x when measuring B with Π_x is

$$P_X(x) = \mathrm{Tr}[\psi'_{AB}\mathbb{1}_A \otimes (\Pi_x)_B] = \mathrm{Tr}[U_{AB}(|\psi\rangle\langle\psi|_A \otimes |0\rangle\langle 0|_B)U^*_{AB}(\mathbb{1}_A \otimes (\Pi_x)_B] = \mathrm{Tr}[\psi\Lambda_x], \tag{4.34}$$

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 a partial isometry. Letting $|\phi'\rangle_{AB} = U_{AB}|\phi\rangle_A \otimes |0\rangle_B$, it follows that

$$\langle\phi'|\psi'\rangle = \sum_{x,x'}(\langle\phi|\sqrt{\Lambda_{x'}} \otimes \langle b_{x'}|)(\sqrt{\Lambda_x}|\psi\rangle \otimes |b_x\rangle) = \sum_x \langle\phi|\Lambda_x|\psi\rangle = \langle\phi|\psi\rangle. \tag{4.35}$$

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 $\mathscr{H}' = \mathscr{H}_A \otimes \mathscr{H}_B$.

Returning to (4.34), the fact that U_{AB} can be seen as a partial isomtery means that the original POVM elements are essentially equivalent to a set of projection operators as in the following theorem.

Theorem 4.2.1: Naimark extension

A Naimark extension of a POVM $\{\Lambda_x\}$ on $\mathscr{H}_A$ is given by

$$(\Pi_x)_{AB} = U^*_{AB}(\mathbb{1}_A \otimes |b_x\rangle\langle b_x|_B)U_{AB}, \tag{4.36}$$

for U_{AB} satisfying $U_{AB}|\psi\rangle_A \otimes |0\rangle_B = \sum_x \sqrt{\Lambda_x}|\psi\rangle_A \otimes |b_x\rangle_B$ with any $|\psi\rangle \in \mathscr{H}_A$.

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* (extra) 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 $\mathbb{1}_A \otimes |b_x\rangle\langle b_x|$. 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 [9, §9-6] or [38, §3.1.4].

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

4.2.4 Post-measurement states and quantum instruments

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

$$\sum_k M_{jk}^* M_{jk} = \Lambda_j. \tag{4.37}$$

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

$$V_{ABC}|\psi\rangle_A|b_0\rangle_B|b_0\rangle_C = \sum_{jk} M_{jk}|\psi\rangle_A|b_j\rangle_B|b_k\rangle_C, \tag{4.38}$$

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

$${}_{BC}\langle b_0, b_0|_A\langle\psi|V_{ABC}^*(\Pi_j)_B V_{ABC}|\psi\rangle_A|b_0, b_0\rangle_{BC} = \sum_k \langle\psi|M_{jk}^* M_{jk}|\psi\rangle = \langle\psi|\Lambda_j|\psi\rangle. \tag{4.39}$$

This is just a reflection of the fact that there are many appropriate "square-roots" of the operator Λ_x in (4.33). However, the output of the two implementations is different:

$$|\psi\rangle \xrightarrow{U} \rho_j = \frac{\sqrt{\Lambda_j}|\psi\rangle\langle\psi|\sqrt{\Lambda_j}}{p_j}, \tag{4.40}$$

$$|\psi\rangle \xrightarrow{V} \rho_j' = \frac{\sum_k M_{jk}|\psi\rangle\langle\psi|M_{jk}^*}{p_j}. \tag{4.41}$$

The specification of the postmeasurement state (along with the measurement outcome) of a POVM is called a *quantum instrument*. The formal definition is as follows.

Definition 4.2.2: Quantum instruments

Consider a quantum measurement with measurement operators M_{xy} and POVM elements $\Lambda_x = \sum_y M_{xy}^* M_{xy}$. For a system Q initially in state ρ_Q, the measurement process produces the state

$$\rho_{XQ}' = \sum_{x\in\mathcal{X}} |b_x\rangle\langle b_x|_X \otimes \sum_{y\in\mathcal{Y}} M_{x,y}\rho M_{x,y}^*, \tag{4.42}$$

which encodes both the outcome in X and the conditional state in Q. The probability of each outcome is given by $P_X(x) = \mathrm{Tr}[\sum_{y\in\mathcal{Y}} M_{x,y}\rho M_{x,y}^*]$.

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.

4.3 Quantum operations

4.3.1 Superoperators

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. From the ensemble interpretation of mixtures of states, 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, as in

$$\mathscr{E}\big(\lambda\rho_0 + (1-\lambda)\rho_1\big) = \lambda\mathscr{E}(\rho_0) + (1-\lambda)\mathscr{E}(\rho_1), \tag{4.43}$$

for any $\rho_0, \rho_1 \in \mathscr{S}(\mathscr{H}_A)$ and any $\lambda \in [0,1]$. If we do not require convexity in this manner, the trouble is that the transformation of a 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!

Any map $\mathscr{E}$ from $\mathscr{S}(\mathscr{H}_A)$ to $\mathscr{S}(\mathscr{H}_B)$ can be uniquely extended to a linear map $\mathscr{E}_{\text{ext}}$ taking $\text{End}(\mathscr{H}_A)$ to $\text{End}(\mathscr{H}_B)$, which agrees with $\mathscr{E}$ on $\mathscr{S}(\mathscr{H}_A)$. Therefore, we broaden our focus to linear maps from operators to operators, often called *superoperators*, *quantum operations*, or just *channels*.

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) $\text{Tr}[\mathscr{E}(\rho)] = 1$ for $\text{Tr}[\rho] = 1$.

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

$$\mathscr{T} : S \mapsto S^T, \tag{4.44}$$

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

$$\langle\phi|S^T|\phi\rangle = \langle\phi|\overline{S}^*\phi\rangle = \langle\overline{S}\phi|\phi\rangle = \overline{\langle\phi|\overline{S}\phi\rangle} = \langle\overline{\phi}|S|\overline{\phi}\rangle \geq 0, \tag{4.45}$$

from which we conclude $S^T \geq 0$. Here $|\overline{\phi}\rangle$ denotes the vector formed from $|\phi\rangle$ by taking the complex conjugate of 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

$$(\mathscr{E} \otimes \mathscr{F})(S \otimes T) := \mathscr{E}(S) \otimes \mathscr{F}(T). \tag{4.46}$$

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 in (3.14). This state is a purification of the maximally-mixed state. The state resulting from the map is simply

$$\rho'_{AB} = \mathscr{I}_A \otimes \mathscr{T}_B(|\Phi\rangle\langle\Phi|_{AB}) = \frac{1}{d}\sum_{jk}|k\rangle\langle j|_A \otimes |j\rangle\langle k|_B. \tag{4.47}$$

Direct calculation reveals that $U_{AB} = d\,\rho'_{AB}$ is the *swap operator*, i.e. $U_{AB}|\psi\rangle_A|\phi\rangle_B = |\phi\rangle_A|\psi\rangle_B$. 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 $\rho'_{AB} \ngeq 0$.

4.3.2 Completely positive maps (CPMs)

In order to ensure compatibility with purification, we must demand that quantum operations be *completely positive*: positive on ρ and all its purifications:

Definition 4.3.1: Completely positive superoperator

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 $\mathrm{CPTP}(\mathscr{H}_A, \mathscr{H}_B)$ the set of completely positive, trace-preserving maps from $\mathrm{End}(\mathscr{H}_A)$ to $\mathrm{End}(\mathscr{H}_B)$.

We have already encountered an example of a CPTP map in §4.2. Performing a measurement described by measurement operators $\{M_k\}$ with $\sum_k M_k^* M_k = \mathbb{1}$ 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

$$\mathscr{E}(\rho) = \sum_k M_k \rho M_k^*. \tag{4.48}$$

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

$$U_{AB}|\psi\rangle_A|0\rangle_B = \sum_k M_k|\psi\rangle_A|k\rangle_B. \tag{4.49}$$

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

In fact, all CPTP maps are of the form (4.48), 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*. The Kraus form implies the existence of a unitary as in (4.49), called the *Stinespring*[4] *dilation*. The fact that any CPTP map has a Stinespring dilation is the content of the *Stinespring representation theorem*, and by appealing to the postulates on unitary dynamics, it shows that CPTP maps encompass all possible physical transformations that could be performed on a quantum system. Historically, the Stinespring representation theorem was established first (as a generalization of the Naimark extension, it so happens), but we shall follow the route via the Choi isomorphism and Kraus representation theorem, as this is simpler for finite-dimensional vector spaces.

[2] Karl Kraus, 1938 – 1988, German physicist.
[3] Man-Duen Choi, Canadian mathematician.
[4] William Forrest Stinespring, American mathematician.

4.3.3 The Choi isomorphism

The Choi isomorphism is a mapping that relates superoperators to operators and CPMs to density operators. It gives rise to a representation of the action of superoperators as operator multiplication and partial trace, and 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*.

Actually, we have already encountered the Choi isomorphism in (4.47). In general, the *Choi map* takes a given channel $\mathcal{E}_{A\to B}$ to the bipartite operator on $\mathcal{H}_A \otimes \mathcal{H}_B$ which results from applying $\mathcal{E}_{A\to B}$ to one subsystem of a maximally-entangled state. Eq. (4.47) shows that the Choi map of the transpose superoperator is proportional to the swap operator. In the following definition we make use of a "copy" of the state space $\mathcal{H}_A$, called $\mathcal{H}_{A'}$. Note that the Choi map depends on the choice of basis used to define the state $|\Phi\rangle_{A'A}$ of (3.14).

Definition 4.3.2: Choi map

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

$$\mathsf{C} : \mathcal{E}_{A\to B} \mapsto (\mathcal{I}_A \otimes \mathcal{E}_{A'\to B})(|\Phi\rangle\langle\Phi|_{AA'}). \tag{4.50}$$

The Choi map gives us a way to represent the action of any given superoperator in terms of multiplication of ordinary bipartite operators and taking partial traces. This is formalized in the following.

Theorem 4.3.1: Choi representation

For any superoperator $\mathcal{E} \in \mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$, there exists an operator $O_{AB} \in \mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that

$$\mathcal{E}_{A\to B}(S_A) = d_A \cdot \mathrm{Tr}_A\Big[\big(\mathcal{T}_A(S_A) \otimes \mathbb{1}_B\big) O_{AB}\Big]. \tag{4.51}$$

Indeed, $O_{AB} = \mathsf{C}(\mathcal{E}_{A\to B})$.

Proof. First, recall that $S_A^T \otimes \mathbb{1}_{A'}|\Phi\rangle_{AA'} = \mathbb{1}_A \otimes S_{A'}|\Phi\rangle_{AA'}$ for arbitrary $S_A \in \mathrm{End}(\mathcal{H}_A)$, where the transpose is defined in the basis defining $|\Phi\rangle_{AA'}$, and that $\mathrm{Tr}_{A'}[|\Phi\rangle\langle\Phi|_{AA'}] = \mathbb{1}_A/d_A$. For convenience, let us simply write $\Phi_{AA'}$ for $|\Phi\rangle\langle\Phi|_{AA'}$. Then we have

$$\begin{aligned}
\mathcal{E}_{A\to B}(S_A) &= \mathcal{E}_{A'\to B}(S_{A'}) = \mathcal{E}_{A'\to B}\big(S_{A'}\mathrm{Tr}_A[\Phi_{A'A}]\big) \\
&= d_A \cdot \mathrm{Tr}_A\Big[\mathcal{E}_{A'\to B} \otimes \mathcal{I}_A\big((S_{A'} \otimes \mathbb{1}_A)\Phi_{A'A}\big)\Big] \\
&= d_A \cdot \mathrm{Tr}_A\Big[\mathcal{E}_{A'\to B} \otimes \mathcal{I}_A\big((\mathbb{1}_{A'} \otimes S_A^T)\Phi_{A'A}\big)\Big] \\
&= d_A \cdot \mathrm{Tr}_A\Big[\big(\mathcal{T}(S_A) \otimes \mathbb{1}_B\big)\big(\mathcal{E}_{A'\to B} \otimes \mathcal{I}_A(\Phi_{A'A})\big)\Big].
\end{aligned} \tag{4.52}$$

We recognize $\mathsf{C}(\mathcal{E}_{A\to B})$ as the second factor in the final expression, completing the proof. □

Using the Choi representation it is easy to see that the Choi map is an isomorphism.

Corollary 4.3.1: Choi isomorphism of superoperators and bipartite operators

The Choi map C is an isomorphism from superoperators $\mathrm{Hom}(\mathrm{End}(\mathscr{H}_A), \mathrm{End}(\mathscr{H}_B))$ to operators $\mathrm{End}(\mathscr{H}_A \otimes \mathscr{H}_B)$. Its inverse C^{-1} takes any O_{AB} to the map $\mathscr{E}_{A\to B}$ defined by (4.51).

Proof. The Choi map is linear, so it is a homomorphism of vector spaces. To be an isomorphism, it must also be both one-to-one (injective) and onto (surjective). The latter holds since any bipartite operator O_{AB} is the image of some superoperator under C, namely, the superoperator defined by the Choi representation (4.51). For the former, observe that, as a linear map, C is one-to-one if and only if $\mathsf{C}(\mathscr{E}) = 0$ implies $\mathscr{E} = 0$. This is clearly the case, again by (4.51).

Alternately, we may infer that C is an isomorphism by showing that $\mathsf{C}^{-1} \circ \mathsf{C}(\mathscr{E}_{A\to B}) = \mathscr{E}_{A\to B}$ for all $\mathscr{E}_{A\to B}$ and $\mathsf{C} \circ \mathsf{C}^{-1}(O_{AB}) = O_{AB}$ for all O_{AB}. The former is established in the proof of the Choi representation. For the latter, we compute

$$\begin{aligned}
\mathsf{C} \circ \mathsf{C}^{-1}(O_{AB}) &= \sum_{jk} |b_j\rangle\langle b_k|_A \cdot \mathrm{Tr}_{A'}\Big[(|b_k\rangle\langle b_j|_{A'} \otimes \mathbb{1}_B) O_{A'B}\Big] \\
&= \sum_{jk} (|b_j\rangle_A\langle b_j|_{A'} \otimes \mathbb{1}_B) O_{A'B} (|b_k\rangle_{A'}\langle b_k|_A \otimes \mathbb{1}_B) \\
&= \Big(\sum_j |b_j\rangle_A\langle b_j|_{A'} \otimes \mathbb{1}_B\Big) O_{A'B} \Big(\sum_k (|b_k\rangle_{A'}\langle b_k|_A \otimes \mathbb{1}_B)\Big) = O_{AB},
\end{aligned} \tag{4.53}$$

which establishes the claim. □

Choi's interest in considering the 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 $\mathscr{E}_{A\to B}$ is completely positive only if $\mathsf{C}(\mathscr{E}_{A\to B})$ is positive. In this case, we call $\mathsf{C}(\mathscr{E}_{A\to B})$ the *Choi state*. We shall return to the 'if' condition later, in Prop. 4.3.2.

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

$$\mathsf{J} : \mathscr{E}_{A\to B} \mapsto (\mathscr{T}_A \otimes \mathscr{E}_{A'\to B})(|\Phi\rangle\langle\Phi|_{AA'}). \tag{4.54}$$

Despite the appearance of the transpose map, this isomorphism is actually basis independent, owing to the fact that $\mathscr{T}_A \otimes \mathscr{I}_{A'}(\Phi_{AA'})$ 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 $\mathscr{E} = \mathsf{J}^{-1}(O_{AB})$ whose action is specified by

$$\mathscr{E} : S_A \mapsto d_A \cdot \mathrm{Tr}_A\Big[(S_A \otimes \mathbb{1}_B) O_{AB}\Big]. \tag{4.55}$$

4.3.4 The Kraus representation theorem

Now we are ready to establish the Kraus representation theorem.

[5] Andrzej Jamiołkowski, born 1946, Polish physicist.

Theorem 4.3.2: Kraus representation

For any $\mathcal{E} \in \mathrm{CPTP}(\mathcal{H}_A, \mathcal{H}_B)$ there exists a family $\{M_\ell\}_\ell$ of operators $M_\ell \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B)$ such that

$$\mathcal{E} : S_A \mapsto \sum_\ell M_\ell S_A M_\ell^* \tag{4.56}$$

and $\sum_\ell M_\ell^* M_\ell = \mathbb{1}_A$. Conversely, any mapping $\mathcal{E}$ of the form (4.56) is in $\mathrm{CPTP}(\mathcal{H}_A, \mathcal{H}_B)$.

Proof. The converse follows from the discussion surrounding (4.49), and will be formally shown later, using the Stinespring representation.

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

$$M_\ell : |\phi\rangle \mapsto \sqrt{\lambda_\ell d_A}\, {}_A\langle\overline{\phi}|\lambda_\ell\rangle_{AB}, \tag{4.57}$$

where complex conjugation is taken in the basis which defines the Choi map. The map is linear, since

$$\begin{aligned} M_\ell\Big(\sum_k \phi_k |b_k\rangle\Big) = M_\ell|\phi\rangle &= \sqrt{\lambda_\ell d_A}\, {}_A\langle\overline{\phi}|\lambda_\ell\rangle_{AB} \\ &= \sqrt{\lambda_\ell d_A}\sum_k \phi_k\, {}_A\langle b_k|\lambda_\ell\rangle_{AB} = \sum_k \phi_k M_\ell |b_k\rangle. \end{aligned} \tag{4.58}$$

Using the eigendecomposition of ρ_{AB} in the Choi representation (4.51) gives, for an arbitrary S_A,

$$\mathcal{E}_{A\to B}(S_A) = d_A \cdot \mathrm{Tr}_A\Big[(S_A^T \otimes \mathbb{1}_B)\sum_\ell \lambda_\ell |\lambda_\ell\rangle\langle\lambda_\ell|_{AB}\Big] \tag{4.59}$$

$$= d_A \sum_\ell \lambda_\ell \mathrm{Tr}_A\Big[\sum_{jk}\langle b_k|S|b_j\rangle\,(|b_j\rangle\langle b_k|_A \otimes \mathbb{1}_B)|\lambda_\ell\rangle\langle\lambda_\ell|_{AB}\Big] \tag{4.60}$$

$$= d_A \sum_\ell \lambda_\ell \sum_{jk}\langle b_k|S|b_j\rangle\, {}_{AB}\langle\lambda_\ell|b_j\rangle_A\, {}_A\langle b_k|\lambda_\ell\rangle_{AB} \tag{4.61}$$

$$= \sum_{jk\ell}\langle b_k|S|b_j\rangle\, M_\ell|b_k\rangle\langle b_j|M_\ell^* = \sum_\ell M_\ell S_A M_\ell^*. \tag{4.62}$$

Since $\mathcal{E}_{A\to B}$ is trace preserving, the following holds for arbitrary ρ:

$$\mathrm{Tr}[\mathcal{E}_{A\to B}(\rho)] = \sum_\ell \mathrm{Tr}[M_\ell \rho M_\ell^*] = \mathrm{Tr}[\sum_\ell M_\ell^* M_\ell \rho]. \tag{4.63}$$

This implies that $\sum_\ell M_\ell^* M_\ell = \mathbb{1}$, 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, Prop. 4.1.2, 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

$$\sqrt{\lambda'_\ell}\langle\overline{\phi}|\lambda'_\ell\rangle = \sum_m \sqrt{\lambda_m}U_{\ell m}\langle\overline{\phi}|\lambda_\ell\rangle, \tag{4.64}$$

and we therefore have the following.

Proposition 4.3.1: Unitary relation of Kraus decompositions

Given any two operator-sum representations of $\mathcal{E}_{A\to B}$ involving Kraus operators M_m and M'_ℓ, there exists a unitary U such that

$$M'_\ell = \sum_m U_{\ell m}M_m. \tag{4.65}$$

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

$$M_\ell|\phi\rangle = \sqrt{\eta_\ell d_A}\,{}_{AB}\langle\eta_\ell|\phi\rangle_A, \tag{4.66}$$

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, but they map $\mathcal{H}_A$ to its dual $\mathcal{H}_A^*$. This removes the need for transposition in the representation of the channel: compare (4.55) with (4.51). 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.

4.3.5 The Stinespring representation theorem

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

Theorem 4.3.3: Stinespring representation

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

$$\mathcal{E}_{A\to B} : S_A \mapsto \mathrm{Tr}_R(U_{A\to BR}S_AU^*_{A\to BR}). \tag{4.67}$$

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

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

$$U_{A\to BR}|\psi\rangle_A|0\rangle_R = \sum_k M_k|\psi\rangle_A|k\rangle_R, \tag{4.68}$$

just as in (4.49). That this is an isometry was already established in (4.35), but we repeat the calculation here for completeness:

$$\langle\phi'|\psi'\rangle = \sum_{\ell,\ell'}(\langle\phi|M_{\ell'}^*\otimes\langle b_{\ell'}|)(M_\ell|\psi\rangle\otimes|b_\ell\rangle) = \sum_\ell\langle\phi|M_\ell^*M_\ell|\psi\rangle = \langle\phi|\psi\rangle. \tag{4.69}$$

Since at most $d_A d_B$ Kraus operators are needed, $\dim(\mathscr{H}_R)$ 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 $\mathscr{E}_{A\to A}$ can be dilated to an isometry $U_{A\to AR}$, which can be extended to a unitary on AR. Unitaries describe dynamical evolution according to the postulates, and therefore CPTP maps describe all possible physical operations on quantum systems.

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 §4.3.3. We have the following

Proposition 4.3.2: Condition for complete positivity

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

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 $\mathscr{E}$ has a Kraus representation and hence a Stinespring dilation $U_{A\to BR}$. Therefore, for any $\mathscr{H}_{R'}$ we have

$$\mathscr{E}_{A\to B}\otimes\mathscr{I}_{R'}(\rho_{AR'}) = \mathrm{Tr}_R[(U_{A\to BR}\otimes\mathbb{1}_{R'})\rho_{AR'}(U_{A\to BR}\otimes\mathbb{1}_{R'})^*], \tag{4.70}$$

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.

Proposition 4.3.3: Choi isomorphism of CPMs and certain bipartite states

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

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. □

4.4 Everything is a quantum operation

Although we have focussed separately on states, measurements, and channels in this chapter, there is really no mathematical distinction between them. For measurements, we have already met the channel description, namely the formulation of quantum instruments. States, on the other hand, can be regarded as channels from $\mathbb{C}$ to $\mathscr{H}$. Specifically, consider the density operator ρ, with eigendecomposition $\rho = \sum_{k=1}^{d} \lambda_k |\lambda_k\rangle\langle\lambda_k|$. Absorbing the eigenvalue into the eigenvector by defining $|\lambda'_k\rangle = \sqrt{\lambda_k}|\lambda_k\rangle$, the channel associated with the density operator ρ has d Kraus operators: $M_k = |\lambda'_k\rangle$. As described in §A.2, $|\lambda'_k\rangle$ can be regarded as a linear map from $\mathbb{C}$ to $\mathscr{H}$. Table 4.4 lists the Kraus operators for different channels.

Operation	Formal equivalent	Instance	Kraus operators
Prepare A	Density operator	$\rho_A = \sum_{k=1}^{d} \lvert\lambda'_k\rangle\langle\lambda'_k\rvert_A$	$\{\lvert\lambda'_k\rangle_A\}_{k=1}^{d}$
Ignore A	Partial trace	Tr_A	$\{\langle b_k\rvert_A\}_{k=1}^{d}$
Measure A	POVM	$\{(\Lambda_x)_A\}_{x=1}^{n}$	$\left\{\lvert x\rangle_X \otimes \langle b_x\rvert_A \sqrt{\Lambda_x}_A\right\}_{x=1}^{n}$
Nonselective mmt	Mmt operators	$\{M_x\}_{x=1}^{n}$	$\{(M_x)_A\}_{x=1}^{n}$
Postmmt	Quantum instrument	$\{M_x\}_{x=1}^{n}$	$\{\lvert x\rangle_X \otimes (M_x)_A\}_{x=1}^{n}$

Table 4.1: Kraus operators associated with different experimental operations.

Finally, it is worth mentioning a fifth representation of superoperators (and operators), the *natural* representation in which superoperator composition is expressed as matrix multiplication. We start by first describing an important property of the maximally entangled state. Consider an arbitrary linear operator $M : \mathscr{H}_A \to \mathscr{H}_B$. Denote by $|\Phi_{d_A}\rangle_{AA'}$ the canonical maximally entangled state on two state spaces $\mathscr{H}_A \simeq \mathscr{H}_{A'}$ of dimension d_A. Then, by using components, it is straightforward to show that

$$\mathbb{1}_A \otimes O_{A'} |\Phi_{d_A}\rangle_{AA'} = (O^T)_{B'} \otimes \mathbb{1}_B |\Phi_{d_B}\rangle_{B'B}, \tag{4.71}$$

where, as usual, the transpose of O is relative to the bases used to define the maximally entangled states of AA' and BB'. Now apply $\mathscr{E}_{A\to B}(\rho_A) = \sum_k M_k \rho M_k^*$ to half of an entangled state:

$$\begin{aligned} \mathbb{1}_{B'} \otimes \mathscr{E}_{A\to B}(\rho_A)|\Phi\rangle_{B'B} &= (\mathbb{1}_{B'} \otimes \sum_k M_k \rho M_k^*)|\Phi\rangle_{B'B} = \sum_k \overline{M}_k \otimes M_k \rho |\Phi\rangle_{A'A} \\ &= \sum_k \overline{M}_k \sqrt{\rho} \otimes M_k \sqrt{\rho} |\Phi\rangle_{A'A} = \sum_k (\overline{M}_k \otimes M_k)(\sqrt{\rho} \otimes \sqrt{\rho})|\Phi\rangle_{A'A}. \end{aligned} \tag{4.72}$$

Applying a subsequent channel with Kraus operators L_j would result in an additional left multiplication by $\sum_j \overline{L}_j \otimes L_j$. Thus we have the desired representation, formalized in the following definition.

Definition 4.4.1: Natural representation

For a superoperator $\mathcal{E}_{A\to B}$ with Kraus operators M_k, the *natural representation* $\mathsf{N}(\mathcal{E}_{A\to B})$ is defined by

$$\mathsf{N}(\mathcal{E}_{A\to B}) = \sum_k \overline{M}_k \otimes M_k. \tag{4.73}$$

It satisfies the relation

$$\mathsf{N}(\mathcal{E}_2 \circ \mathcal{E}_1) = \mathsf{N}(\mathcal{E}_2)\mathsf{N}(\mathcal{E}_2). \tag{4.74}$$

4.5 Notes & Further Reading

The phrase "church of the larger Hilbert space" was coined by John Smolin and is far from the only pun in the terminology of quantum information theory. The mathematical structure relevant for open quantum systems that we have traced here developed in the works of Naimark [39], Stinespring [40], Hellwig and Kraus [41, 42], Jamiołkowski [43] (building on work of de Pillis [44]) and Choi [45]. For more detailed treatments, see the books of Davies [46] and especially Kraus [47].

4.6 Exercises

Exercise 4.1. The Bloch ball → solution

In this exercise we will show that qubit density operators can always be expressed as

$$\rho = \tfrac{1}{2}(\mathbb{1} + \vec{r}\cdot\vec{\sigma}), \tag{4.75}$$

where $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ and $\vec{r} = (r_x, r_y, r_z)$, $|\vec{r}| \leq 1$ is the Bloch vector, specifying a point in the unit ball. The surface of the ball is the Bloch sphere.

a) Find and diagonalize the states represented by Bloch vectors $\vec{r}_1 = (\frac{1}{2}, 0, 0)$ and $\vec{r}_2 = (\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$.

b) Show that the operator ρ defined in (4.75) is a valid density operator for any vector $\vec{r}$ with $|\vec{r}| \leq 1$.

c) Now show the converse: Any two-level density operator may be written as (4.75).

d) Check that the surface of the ball is formed by all the pure states.

Exercise 4.2. Partial trace → solution

Let ρ_{AB} be a density matrix on the bipartite Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ and $\rho_A = \mathrm{Tr}_B[\rho_{AB}]$.

a) Show that ρ_A is a valid density operator.

b) Suppose ρ_{AB} is a pure state, $\rho_{AB} = |\Psi\rangle\langle\Psi|_{AB}$ for $|\Psi\rangle_{AB} = \sum_{jk}^{d_A,d_B} C_{jk}|b_j\rangle_A \otimes |b'_k\rangle_B$, where $\{|b_j\rangle\}_j$ and $\{|b'_j\rangle\}_j$ are orthonormal bases for $\mathcal{H}_A$ and $\mathcal{H}_B$. Let C be the $d_A \times d_B$ matrix with entries C_{jk}. Show that $\rho_A = CC^\dagger$ and $\rho_B = C^\dagger C$, where $C^\dagger$ is the conjugate transpose of C.

c) Calculate the reduced density matrix of system A in any of the Bell states.

d) Consider a classical probability distribution P_{XY}. Calculate the marginal distribution P_X for

$$P_{XY}(x,y) = \begin{cases} \frac{1}{2} & \text{for } (x,y) = (0,0), \\ \frac{1}{2} & \text{for } (x,y) = (1,1), \\ 0 & \text{else,} \end{cases}$$

with alphabets $\mathcal{X}, \mathcal{Y} = \{0,1\}$. How can we represent P_{XY} in the form of a quantum state? Calculate the partial trace of P_{XY} in its quantum representation.

Exercise 4.3. Canonical purifications → solution

Given a state ρ on $\mathcal{H}_A$, consider the state $|\psi\rangle_{AB}$ on $\mathcal{H}_A \otimes \mathcal{H}_B$, defined as

$$|\psi\rangle_{AB} = \left(\sqrt{\rho_A} \otimes U_B\right)|\Omega\rangle, \qquad |\Omega\rangle_{AB} = \sum_k |k\rangle_A \otimes |k\rangle_B, \tag{4.76}$$

where U_B is any unitary on $\mathcal{H}_B \simeq \mathcal{H}_A$.

a) Show that $|\psi\rangle_{AB}$ is a purification of ρ_A.

b) Show that every purification of ρ can be written in this form.

Exercise 4.4. Decompositions of density matrices → solution

Consider a mixed state ρ with two different pure state decompositions

$$\rho = \sum_{k=1}^{d} \lambda_k |k\rangle\langle k| = \sum_{\ell=1}^{n} p_\ell |\phi_\ell\rangle\langle\phi_\ell|,$$

the former being the eigendecomposition so that $\{|k\rangle\}$ is an orthonormal basis.

a) Show that the probability vector $\vec{\lambda}$ *majorizes* the probability vector $\vec{p}$, which means that there exists a doubly stochastic matrix T_{jk} such that $\vec{p} = T\vec{\lambda}$. The defining property of doubly stochastic, or bistochastic, matrices is that $\sum_k T_{jk} = \sum_j T_{jk} = 1$.
Hint: Observe that for a unitary matrix U_{jk}, $T_{jk} = |U_{jk}|^2$ is doubly stochastic.

b) The uniform probability vector $\vec{u} = (\frac{1}{n}, \ldots, \frac{1}{n})$ is invariant under the action of an $n \times n$ doubly stochastic matrix. Is there an ensemble decomposition of ρ such that $p_\ell = \frac{1}{n}$ for all ℓ?
Hint: Try to show that $\vec{u}$ is majorized by any other probability distribution.

Exercise 4.5. Generalized measurement by direct (tensor) product → solution

Consider an apparatus whose purpose is to make an indirect measurement on a two-level system, A, by first coupling it to a three-level system, B, and then making a projective measurement on the latter. B is initially prepared in the state $|0\rangle$ and the two systems interact via the unitary U_{AB} as follows:

$$|0\rangle_A|0\rangle_B \to \tfrac{1}{\sqrt{2}}(|0\rangle_A|1\rangle_B + |0\rangle_A|2\rangle_B),$$
$$|1\rangle_A|0\rangle_B \to \tfrac{1}{\sqrt{6}}(2|1\rangle_A|0\rangle_B + |0\rangle_A|1\rangle_B - |0\rangle_A|2\rangle_B).$$

a) Calculate the measurement operators acting on A corresponding to a measurement on B in the canonical basis $|0\rangle, |1\rangle, |2\rangle$.

b) Calculate the corresponding POVM elements. What is their rank? Onto which states do they project?

c) Suppose A is in the state $|\psi\rangle_A = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)_A$. What is the state after a measurement, averaging over the measurement result?

Exercise 4.6. Geometry of POVMs → solution

In this exercise we will show that the set of 2-outcome POVMs is a convex set with orthogonal measurements as extremal points.

Let $F = \{F_1, F_2\}$ and $G = \{G_1, G_2\}$ be two two-outcome POVMs. We define an element-wise convex combination of F and G as $\alpha F + (1-\alpha)G := \{\alpha F_1 + (1-\alpha)G_1, \alpha F_2 + (1-\alpha)G_2\}$, with $0 \leq \alpha \leq 1$.

a) Consider a POVM with two outcomes and respective measurement operators E and $\mathbb{1} - E$. Suppose that E has an eigenvalue λ such that $0 < \lambda < 1$. Show that the POVM is not extremal by expressing it as a nontrivial convex combination of two POVMs.
Hint: Consider the spectral decomposition of E and rewrite it as a convex combination of two POVM elements.

b) Suppose that E is an orthogonal projector. Show that the POVM cannot be expressed as a nontrivial convex combination of POVMs.

c) What is the operational interpretation of an element-wise convex combination of POVMs?

Exercise 4.7. Some common quantum channels → solution

Find Kraus representations for the following qubit channels

- The dephasing channel: $\rho \to \rho' = \mathcal{E}(\rho) = (1-p)\rho + p\,\mathrm{diag}(\rho_{00}, \rho_{11})$ (the off-diagonal elements are annihiliated with probability p).
- The depolarizing channel: $\rho \to \rho' = \mathcal{E}(\rho) = (1-p)\rho + \frac{p}{2}\mathbb{1}$.
- The amplitude damping (damplitude) channel, defined by the action $|00\rangle \to |00\rangle$, $|10\rangle \to \sqrt{1-p}|10\rangle + \sqrt{p}|01\rangle$.

What is the minimal number of Kraus operators in each case? What happens to the Bloch vector? In each case, can the Kraus operators be chosen to be unitaries? Or projection operators?

Exercise 4.8. Classical channels as CPTP maps → solution

In this exercise we will see how to represent classsical channels as CPTP maps.

a) Consider the binary symmetric channel, defined by

$$W(y|x) = \begin{cases} 1-p & y = x \\ p & y \neq x \end{cases},$$

for $x, y \in \{0, 1\}$. Recall that we can represent the probability distributions on both ends of the channel as quantum states in a given basis: for instance, if $P_X(0) = q, P_X(1) = 1-q$, we may express this as the 1-qubit mixed state $\rho_X = q\,|0\rangle\langle 0| + (1-q)\,|1\rangle\langle 1|$.

What is the quantum state ρ_Y that represents the final probability distribution P_Y in the computational basis?

b) Next we want to represent the channel as a map

$$\begin{aligned} \mathcal{E}_W : \mathcal{S}(\mathcal{H}_X) &\mapsto \mathcal{S}(\mathcal{H}_Y) \\ \rho_X &\to \rho_Y. \end{aligned}$$

Find a Kraus representation of $\mathcal{E}_W$.

Hint: think of each Kraus operator $M_k = M_{xy}$ as the representation of the branch that maps input x to output y.

c) Now we have a representation of the classical channel in terms of the evolution of a quantum state. What happens if the initial state ρ_X is not diagonal in the standard basis?

d) Consider an arbitrary classical channel $W(y|x)$ from an n-bit space X to an m-bit space Y, defined by the conditional probabilities $\left\{P_{Y|X=x}(y)\right\}_{xy}$. Express W as a map $\mathcal{E}_W : \mathcal{S}(\mathcal{H}_X) \mapsto \mathcal{S}(\mathcal{H}_Y)$ in the Kraus representation.

Exercise 4.9. Unital channels → solution

A superoperator $\mathcal{E}$ is unital if $\mathcal{E}(\mathbb{1}) = \mathbb{1}$, or in terms of Kraus operators, $\sum_k A_k A_k^* = \mathbb{1}$. Show that the eigenvalues of the output ρ' of a unital superoperator majorize the eigenvalues of the input ρ. *Hint:* Express the input (output) as $\rho = U\Lambda U^*$ ($\rho' = V\Lambda' V^*$) for U, V unitary and Λ, Λ' diagonal.

Exercise 4.10. The Choi isomorphism → solution

Consider the family of mappings between operators on two-dimensional Hilbert spaces

$$\mathcal{E}_\alpha : \rho \mapsto \tfrac{1}{2}\mathbb{1} + \alpha(X\rho Z + Z\rho X), \qquad 0 \leq \alpha \leq 1,$$

where X and Z are the Pauli operators.

a) Use the Bloch representation to determine for what range of α these mappings are positive. What happens to the Bloch sphere?

b) Calculate the Choi state of $\mathcal{E}_\alpha$. For what range of α is the mapping a CPTP map?

c) Find a Kraus representation of $\mathcal{E}_\alpha$ for $\alpha = 1/4$.

Exercise 4.11. The Stinespring isometry

Show the existence of the Stinespring isometry directly from the Choi state.
Hint: purify the Choi state and use Proposition 4.3.3 and (4.71).

Exercise 4.12. The natural representation

What is the complete positivity condition in the natural representation? What is the condition on trace preservation? Having a unital channel?

5

Quantum Detection Theory

In this chapter we study various distinguishability measures of quantum states and channels. These are interesting in their own right and will later be used for to give a quantitative means of saying that one protocol approximates another, as well as in constructing various information processing protocols.

5.1 Distinguishing states & channels

5.1.1 State preparation

Imagine someone builds a device which either produces the state ρ or the state σ, but does not tell us which. Nor can we look inside the device to figure out which state is produced. How easily can we tell the difference between the two devices? How should we measure the distinguishability of the two states?

A direct approach is to ask how well we could determine which device we have by performing an actual experiment on it. The probability of correctly guessing can be regarded as a measure of the distinguishability of the two states or devices. Our *prior* probability of which device we have is given by some distribution P_X, where $X = 0$ is associated with ρ and $X = 1$ with σ. Then, given the results of an experiment we perform, we would like the probability of correctly guessing X to be as large as possible.

Since we cannot look inside the state preparation device, our only option is to measure the output state in some way. We can describe any measurement with a POVM, and in this case the POVM need only have two outcomes, one associated with ρ and the other with σ. A seemingly more general setup would be to perform a measurement with many outcomes and then base our decision on these outputs, possibly in a probabilistic way. For instance, given the outcome of the measurement we may we might flip a biased coin to generate our guess as to the true state. But this entire procedure can be described with an initial POVM, say with elements Γ_y, followed by a channel $W : \mathscr{Y} \to \mathscr{X}' = \{0, 1\}$. The total probability that the test indicates ρ for an arbitrary input state ξ is simply

$$P(X' = 0|\xi) = \sum_y W(0|y)\mathrm{Tr}[\xi\Gamma_y] \tag{5.1}$$

$$= \mathrm{Tr}[\xi \sum_y W(0|y)\Gamma_y]. \tag{5.2}$$

Thus we may define the POVM elements $\Lambda_x = \sum_y W(x|y)\Gamma_y$ and simply use these directly. These operators form a valid POVM since $W(x|y)$ is a conditional probability distribution for each y.

Let us label the output state space of the device by B. For a given probability $P_X(x)$ that the actual device produces $\xi_0 = \rho$ or $\xi_1 = \sigma$, the average probability of correctly guessing when using the POVM with elements Λ_x is simply

$$p_{\text{guess}}(X|B) = \sum_x P_X(x)\mathrm{Tr}[\Lambda_x \xi_x]. \tag{5.3}$$

When P_X is uniform, so that we believe the actual device is equally-likely to either of the two possibilities, the average guessing probability is simply

$$p_{\text{guess}}(X|B) = \tfrac{1}{2}\sum_x \mathrm{Tr}[\Lambda_x \xi_x] = \tfrac{1}{2}(\mathrm{Tr}[\Lambda_0\rho] + \mathrm{Tr}[\Lambda_1\sigma]) \tag{5.4}$$

$$= \tfrac{1}{2}(\mathrm{Tr}[\Lambda_0\rho] + \mathrm{Tr}[(\mathbb{1} - \Lambda_0)\sigma]) = \tfrac{1}{2}(1 + \mathrm{Tr}[\Lambda_0(\rho - \sigma)]). \tag{5.5}$$

We can regard the second term as a measure of distinguishability. If the two states are disjoint (have disjoint support), then we can guess perfectly, which is reflected in the fact that the second term is unity. On the other hand, if the states are identical, we may as well guess, and the second term is zero. By optimizing over all POVM elements we can increase the guessing probability, and thus the distinguishability. This leads to the following definition.

Definition 5.1.1: State Distinguishability

The *distinguishability* of any two quantum states ρ and σ is defined as

$$\delta(\rho,\sigma) := \max \operatorname{Tr}[\Lambda(\rho-\sigma)] \quad \text{s.t. } 0 \leq \Lambda \leq \mathbb{1}. \tag{5.6}$$

Clearly the state distinguishability is symmetric in its arguments and $\delta(\rho,\sigma) \geq 0$, since $\Lambda = 0$ is a possible POVM element.

The difference $\rho-\sigma$ is a Hermitian operator, and therefore has real eigenvalues; the projector onto the positive part of $\rho-\sigma$ is clearly the optimal POVM element. Denote the projection onto the positive part by $\{\rho-\sigma \geq 0\}$, observe that

$$0 = \operatorname{Tr}[\rho-\sigma] = \operatorname{Tr}[\{\rho-\sigma \geq 0\}(\rho-\sigma)] + \operatorname{Tr}[\{\rho-\sigma < 0\}(\rho-\sigma)], \tag{5.7}$$

while, if λ_j are the eigenvalues of $\rho-\sigma$,

$$\sum_j |\lambda_j| = \operatorname{Tr}[\{\rho-\sigma \geq 0\}(\rho-\sigma)] - \operatorname{Tr}[\{\rho-\sigma < 0\}(\rho-\sigma)]. \tag{5.8}$$

From (A.48), the sum of the absolute values of eigenvalues of a Hermitian operator is just the trace norm, so we have found that

$$\delta(\rho,\sigma) = \tfrac{1}{2}\|\rho-\sigma\|_1. \tag{5.9}$$

This form immediately implies that $\delta(\rho,\sigma) = 0$ iff $\rho = \sigma$.

For commuting states, i.e. classical probability distributions, we can write $\rho = \sum_x P(x)|x\rangle\langle x|$ and $\sigma = \sum_x Q(x)|x\rangle\langle x|$. The optimal measurement is to observe the value of x and then report ρ if it is such that $P(x) \geq Q(x)$ and report σ otherwise. In the general case of noncommuting ρ and σ, we saw above that all that really matters in the problem is the operator $\rho-\sigma$. Thus, to optimally distinguish the states we can simply measure in the eigenbasis of $\rho-\sigma$ and report ρ for all outcomes j associated with $\lambda_j \geq 0$ and report σ otherwise. This is just the measurement found in Exercise 5.1.

Often (5.9) is taken as the definition of the distinguishability and the operational meaning is investigated subsequently. But it is important to observe that the following three useful properties of the distinguishability follow directly from the operational (variational) definition.

The first is the triangle inequality:

Proposition 5.1.1: Triangle inequality for state distinguishability

For any quantum states ρ, σ, and τ,

$$\delta(\rho,\sigma) \leq \delta(\rho,\tau) + \delta(\tau,\sigma). \tag{5.10}$$

Proof.

$$\max_{\Lambda} \mathrm{Tr}[\Lambda(\rho-\sigma)] = \max_{\Lambda} \mathrm{Tr}[\Lambda(\rho-\tau+\tau-\sigma)] \tag{5.11}$$
$$= \max_{\Lambda} (\mathrm{Tr}[\Lambda(\rho-\tau)] + \mathrm{Tr}[\Lambda(\tau-\sigma)]) \tag{5.12}$$
$$\leq \max_{\Lambda} \mathrm{Tr}[\Lambda(\rho-\tau)] + \max_{\Lambda} \mathrm{Tr}[\Lambda(\tau-\sigma)]. \tag{5.13}$$

□

The distinguishability forms a *metric* or distance measure on the space of density operators since it is symmetric, positive semidefinite, nondegenerate (zero iff the arguments are equal), and obeys the triangle inequality.

The second important property is monotonicity under all CPTP maps: The distinguishability can only decrease after applying a physical map to the states.

Proposition 5.1.2: Monotonicity of the state distinguishability

For any CPTP map $\mathcal{E}$ and states ρ and σ,

$$\delta(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \leq \delta(\rho, \sigma). \tag{5.14}$$

Proof. The proof rests on the fact that applying $\mathcal{E}$ and then performing a distinguishing experiment is a particular kind of experiment, and is therefore generally suboptimal. To state this formally it is easiest to use the adjoint map $\mathcal{E}^*$, defined via the inner product as usual:

$$\mathrm{Tr}[\mathcal{E}^*(\Lambda)\rho] = \mathrm{Tr}[\Lambda\mathcal{E}(\rho)]. \tag{5.15}$$

Observe that if $\mathcal{E}$ has a Kraus representation with operators M_k, then $\mathcal{E}^*$ has a Kraus representation with operators M_k^*. Therefore, $\mathcal{E}^*$ is therefore completely positive. Furthermore, the trace preserving condition of $\mathcal{E}$ implies that $\mathcal{E}^*$ is unital, meaning $\mathcal{E}^*(\mathbb{1}) = \mathbb{1}$. This implies that $\mathcal{E}^*(\Lambda)$ is a valid POVM element if Λ is, since $\mathbb{1} - \Lambda \geq 0$ gives $\mathcal{E}^*(\Lambda) \leq \mathbb{1}$. Thus, we have

$$\delta(\mathcal{E}(\rho), \mathcal{E}(\sigma)) = \max_{\Lambda' : 0 \leq \Lambda' \leq \mathbb{1}} \mathrm{Tr}[\mathcal{E}^*(\Lambda')(\rho-\sigma)] \tag{5.16}$$
$$= \max_{\Lambda : \Lambda = \mathcal{E}^*(\Lambda'), 0 \leq \Lambda' \leq \mathbb{1}} \mathrm{Tr}[\Lambda(\rho-\sigma)] \tag{5.17}$$
$$\leq \max_{\Lambda : 0 \leq \Lambda \leq \mathbb{1}} \mathrm{Tr}[\Lambda(\rho-\sigma)] \tag{5.18}$$
$$= \delta(\rho, \sigma). \tag{5.19}$$

□

Note that by discussion after (5.9), there exists a channel such that the distinguishability does not decrease: the measurement in the eigenbasis of $\rho - \sigma$.

The third property is convexity of the distinguishability measure:

Proposition 5.1.3: Joint convexity of the state distinguishability

For $0 \leq \lambda \leq 1$, states ρ_1, ρ_2, σ_1, and σ_2 with $\rho = \lambda\rho_1 + (1-\lambda)\rho_2$ and similarly for σ,

$$\delta(\rho,\sigma) \leq \lambda\delta(\rho_1,\sigma_1) + (1-\lambda)\delta(\rho_2,\sigma_2). \tag{5.20}$$

Proof. The claim follows from monotonicity under partial trace for suitably-chosen states. Let $\rho'_{XA} = \lambda|1\rangle\langle 1|_X \otimes \rho_1 + (1-\lambda)|2\rangle\langle 2|_X \otimes \rho_2$ and similarly for σ'_{XA}. Then choosing $\mathcal{E}$ to be the partial trace over X, we have $\rho = \mathcal{E}(\rho')$ and $\sigma = \mathcal{E}(\sigma')$. To distinguish between ρ' and σ', we may simply look at the X system and then perform the corresponding ρ_j versus σ_j experiment. This leads to the quantity on the righthand side of the bound (and is in fact the optimal test). □

It is interesting to observe that we could restrict the set of POVM elements and still obtain a distinguishability measure that obeys the triangle inequality. For instance, in the setting of bipartite systems, we may only be able to perform measurements on one system at a time. The proof of Proposition 5.1.1 evidently does not change if we restrict the set of POVM elements Λ over which the maximization is performed. Moreover, such distinguishability measures also inherit monotonicity from the proof of Proposition 5.1.2, at least for those channels $\mathcal{E}$ which preserve the restricted set of POVM elements. In the bipartite example this would include operations on the two systems individually.

5.1.2 General quantum channels

To distinguish two quantum channels we can adopt the same approach: ask for the best possible probability in correctly guessing the actual channel using the results of an experiment on the device. The only difference with the previous scenario is that in an experiment on a channel we are free to choose the input state as well as the measurement on the output. This fits neatly with the view that a state is actually a channel from a one-dimensional system, i.e. a channel with a fixed input.

Since we are free to choose the input, we can also contemplate entangled inputs. That is, we may test a channel $\mathcal{E}_{A'\to B}$ by allowing it to act on one part of a bipartite input state $\rho_{AA'}$ and then perform a joint measurement on the output systems A and B. This leads to the following definition.

Definition 5.1.2: Channel distinguishability

The distinguishability of two CPTP channels $\mathcal{E}_{A'\to B}$ and $\mathcal{F}_{A'\to B}$ is defined as

$$\begin{aligned} \delta(\mathcal{E},\mathcal{F}) \coloneqq \max\ & \mathrm{Tr}[\Lambda_{AB}(\mathcal{I}_A \otimes \mathcal{E}_{A'\to B}(\rho_{AA'}) - \mathcal{I}_A \otimes \mathcal{F}_{A'\to B}(\rho_{AA'}))]. \\ \text{s.t. } & 0 \leq \Lambda \leq \mathbb{1} \\ & \rho_{AA'} \geq 0, \mathrm{Tr}[\rho_{AA'}] = 1 \end{aligned} \tag{5.21}$$

We have seen that it is generally important to consider the possibility of entangled inputs to channels, and not allowing such inputs could limit the power of the experimental tests. Indeed, we will see an example of a channels for which this is precisely the case in Exercise 5.4.

Since the channel distinguishability is just the distinguishability of the states that the two channels can produce from the optimal, fixed input, we can also write (omitting $\mathcal{I}_A$)

$$\delta(\mathcal{E},\mathcal{F}) = \max_{\rho_{AA'}} \delta(\mathcal{E}_{A'\to B}(\rho_{AA'}), \mathcal{F}_{A'\to B}(\rho_{AA'})). \tag{5.22}$$

In much the same way as in the previous section, we can easily show the following properties of the channel distinguishability.

Proposition 5.1.4: Properties of the channel distinguishability

The following properties hold for arbitrary CPTP maps $\mathcal{E}_{A\to B}$, $\mathcal{E}'_{A\to B}$, $\mathcal{E}''_{A\to B}$, $\mathcal{F}_{A'\to A}$, and $\mathcal{G}_{B\to B'}$:

1) Positivity: $\delta(\mathcal{E},\mathcal{E}')\geq 0$,
2) Triangle inequality: $\delta(\mathcal{E},\mathcal{E}'')\leq \delta(\mathcal{E},\mathcal{E}')+\delta(\mathcal{E}',\mathcal{E}'')$,
3) Monotonicity: $\delta(\mathcal{E}\circ\mathcal{F},\mathcal{E}'\circ\mathcal{F})\leq\delta(\mathcal{E},\mathcal{E}')$ and $\delta(\mathcal{G}\circ\mathcal{E},\mathcal{G}\circ\mathcal{E}')\leq\delta(\mathcal{E},\mathcal{E}')$.

Proof. The first claim follows by choosing $\Lambda=0$ and any input ρ in the optimization. The second is essentially the same as the proof of triangle inequality for state distinguishability, Proposition 5.1.1. The third follows for reasons similar to the proof of Proposition 5.1.2. Namely, in the first relation the channel $\mathcal{F}$ maps the set of possible input states onto a possibly smaller set, restricting the optimization. In the second relation $\mathcal{G}^*$ restricts the set of POVM elements. □

5.2 Fidelity

In §5.1.1 we motivated the notion of distinguishability of state prepraration by asking how well we could distinguish two different state preparations in any possible experiment. What if we had access to the purification while doing so? Investigating this question leads to the notion of fidelity.

5.2.1 Fidelity of quantum states

First, recall from Exercise 3.2 that for already-pure states we have the relation

$$\delta(\phi,\psi)=\sqrt{1-|\langle\phi|\psi\rangle|^2}. \tag{5.23}$$

This motivates the following definition of fidelity of pure states:

$$F(\phi,\psi):=|\langle\phi|\psi\rangle|. \tag{5.24}$$

A word of warning: many authors define fidelity as the *squared* overlap. This has an appealing interpretation as a probability for state ψ to pass a test (POVM element) of the form $|\phi\rangle\langle\phi|$ (or *vice versa*). However, the fidelity as we have defined it here also shows up in classical information theory and we will follow this convention.

If mixed states can be compared by also examining their purifications, then (5.23) also provides a link between the overlap of the purifications and the distinguishability. Since a mixed state has many purifications, we consider the worst case distinguishability and define the fidelity as follows.

Definition 5.2.1: Fidelity

Given two quantum states ρ and σ on $\mathscr{H}_A$, the fidelity $F(\rho,\sigma)$ is defined as

$$F(\rho,\sigma) \coloneqq \max_{\psi_{AB},\phi_{AB}} |\langle\psi|\phi\rangle_{AB}|, \tag{5.25}$$

where the maximization is taken over all purifications of ρ and σ.

From the definition it is immediately clear that $0 \leq F(\rho,\sigma) \leq 1$ and that it is invariant under unitaries or partial isometries. Monotonicity of fidelity under arbitrary quantum operations is also straightforward.

Proposition 5.2.1: Montonicity of fidelity

For a CPTP map $\mathscr{E}_{A\to B}$ and two quantum states ρ and σ,

$$F(\mathscr{E}(\rho),\mathscr{E}(\sigma)) \geq F(\rho,\sigma). \tag{5.26}$$

Proof. Let $V_{A\to AB}$ be a Stinespring isometry for $\mathscr{E}_{A\to B}$, and define $\hat{\rho}_{AB} = V_{A\to AB}\rho_A V_{A\to AB}^*$ and similarly for $\hat{\sigma}_{AB}$. Then $\mathscr{E}_{A\to B}(\rho_A) = \mathrm{Tr}_A[\hat{\rho}_{AB}] = \hat{\rho}_B$. By the invariance under partial isometries we have $F(\rho_A,\sigma_A) = F(\hat{\rho}_{AB},\hat{\sigma}_{AB})$. Thus, it remains to establish that the fidelity can only increase under partial trace. Suppose that $\hat{\rho}_{ABC}$ and $\hat{\sigma}_{ABC}$ are the optimal purifications such that

$$F(\hat{\rho}_{AB},\hat{\sigma}_{AB}) = F(\hat{\rho}_{ABC},\hat{\sigma}_{ABC}). \tag{5.27}$$

Trivially, $\hat{\rho}_{ABC}$ and $\hat{\sigma}_{ABC}$ are also purifications of $\hat{\rho}_B$ and $\hat{\sigma}_B$, respectively. Hence they are feasible for the optimization in $F(\hat{\rho}_B,\hat{\sigma}_B)$ and thus $F(\hat{\rho}_B,\hat{\sigma}_B) \geq F(\hat{\rho}_{AB},\hat{\sigma}_{AB})$. □

Our definition is in the spirit of the original by Bures and Uhlmann[1], who used the overlap squared, but these days fidelity is more commonly defined as the quantity in the following proposition, known as Uhlmann's theorem.

Theorem 5.2.1: Uhlmann's theorem

For any two quantum states ρ and σ,

$$F(\rho,\sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1. \tag{5.28}$$

Proof. Let $|\phi\rangle_{AB}$ be a purification of ρ_A and $|\psi\rangle_{AB}$ a purification of σ_A. (If they do not have identical purification spaces, embed the smaller in the larger so that they do.) Defining the unnormalized state $|\Omega\rangle_{AA'} = \sum_x |b_x\rangle_A \otimes |b_x\rangle_{A'}$, we can write the purifications as $|\phi\rangle_{AB} = \sqrt{\rho}_A \otimes U_{A'\to B}|\Omega\rangle_{AA'}$ and similarly $|\psi\rangle_{AB} = \sqrt{\sigma}_A \otimes V_{A'\to B}|\Omega\rangle_{AA'}$ for some isometries U and V.

By Proposition 4.1.1, for some choice of unitaries U and V we can write

$$\begin{aligned} F(\phi_{AB},\psi_{AB}) &= |\langle\phi|\psi\rangle_{AB}| && (5.29)\\ &= |\langle\Omega|\sqrt{\rho_A}\sqrt{\sigma_A} \otimes V_{A'\to B}^* U_{A'\to B}|\Omega\rangle_{AA'}| && (5.30)\\ &= |\mathrm{Tr}[\sqrt{\rho}\sqrt{\sigma}U^T(V^*)^T]|. && (5.31) \end{aligned}$$

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

Since U and V are isometries, the product V^*U is unitary and therefore $\max_{\phi_{AB},\psi_{AB}} F(\phi_{AB},\psi_{AB})$ amounts to $\max_U |\mathrm{Tr}[\sqrt{\rho}\sqrt{\sigma}U]|$. By Lemma A.7.2 we then have

$$\max_{\phi_{AB},\psi_{AB}} F(\phi_{AB},\psi_{AB}) = \max_U |\mathrm{Tr}[U\sqrt{\rho_A}\sqrt{\sigma_A}]| \tag{5.32}$$

$$= \|\sqrt{\rho_A}\sqrt{\sigma_A}\|_1, \tag{5.33}$$

completing the proof. □

Using this expression for the fidelity, we can infer that there exists a measurement which does not decrease the fidelity. This was also the case for the distinguishability, though for fidelity the situation is more involved.

Proposition 5.2.2: Measurement achieving the fidelity

For any two states ρ and σ, there exists a POVM $\{\Lambda_x\}$ with outcome distributions $P(x) = \mathrm{Tr}[\Lambda_x\rho]$ and $Q(x) = \mathrm{Tr}[\Lambda_x\sigma]$, respectively, such that

$$F(P,Q) = F(\rho,\sigma). \tag{5.34}$$

Proof. The proof is a careful application of the Cauchy-Schwartz inequality. Let U be the optimal unitary in (5.32). Then the fidelity is just

$$F(\rho,\sigma) = \mathrm{Tr}[U\sqrt{\rho}\sqrt{\sigma}] \tag{5.35}$$

$$= \sum_x \mathrm{Tr}[U\sqrt{\rho}\sqrt{\Lambda_x}\sqrt{\Lambda_x}\sqrt{\sigma}], \tag{5.36}$$

where we have used the condition $\sum_x \Lambda_x = \mathbb{1}$ in the second equation. We can regard each term in this expression as the inner product $\langle u_x, v_x\rangle = \mathrm{Tr}[u_x^* v_x]$ for $u_x^* = U\sqrt{\rho}\sqrt{\Lambda_x}$ and $v_x = \sqrt{\Lambda_x}\sqrt{\sigma}$. Then, by the Cauchy-Schwartz inequality,

$$F(\rho,\sigma) = \sum_x \langle u_x, v_x\rangle \leq \sum_x \sqrt{\langle u_x,u_x\rangle\langle v_x,v_x\rangle} = \sum_x \sqrt{P(x)Q(x)}. \tag{5.37}$$

For equality we need to satisfy $u_x \propto v_x$ for all x, i.e.

$$\sqrt{\Lambda_x}\sqrt{\rho}U^* \propto \sqrt{\Lambda_x}\sqrt{\sigma}. \tag{5.38}$$

Recalling Lemma A.7.2, the optimal U satisfies $\sqrt{\rho}\sqrt{\sigma} = U^*|\sqrt{\rho}\sqrt{\sigma}|$ and therefore $\sqrt{\rho} = U^*|\sqrt{\rho}\sqrt{\sigma}|\sigma^{-1/2}$ for invertible σ, or equivalently $\sqrt{\rho}U^* = \sigma^{-1/2}|\sqrt{\rho}\sqrt{\sigma}|$. This gives the condition

$$\sqrt{\Lambda_x}\sigma^{-1/2}|\sqrt{\rho}\sqrt{\sigma}|\sigma^{-1/2} \propto \sqrt{\Lambda_x}. \tag{5.39}$$

Choosing Λ_x to be projectors onto the eigenbasis of $\sigma^{-1/2}|\sqrt{\rho}\sqrt{\sigma}|\sigma^{-1/2}$ satisfies this condition and completes the proof. □

Importantly, the fidelity and state distinguishability are related to each other, as given in the following bounds. Note that since the dimension of the state space does not appear, the two quantities can be thought of as essentially the same (as are two norms on a space which are related by dimension-independent bounds).

Proposition 5.2.3: Bounds relating fidelity and distinguishability

For any two quantum states ρ and σ,

$$\delta(\rho,\sigma)+F(\rho,\sigma)\geq 1, \qquad \text{and} \tag{5.40}$$
$$\delta(\rho,\sigma)^2+F(\rho,\sigma)^2\leq 1. \tag{5.41}$$

Proof. For the former, let P and Q be the distributions arises from the measurement achieving the fidelity. By monotonicity of the distinguishability we have

$$\delta(\rho,\sigma)\geq\delta(P,Q)=\tfrac{1}{2}\sum_x|P(x)-Q(x)| \tag{5.42}$$
$$=\tfrac{1}{2}\sum_x|\sqrt{P(x)}-\sqrt{Q(x)}|(\sqrt{P(x)}+\sqrt{Q(x)}) \tag{5.43}$$
$$\geq\tfrac{1}{2}\sum_x(\sqrt{P(x)}-\sqrt{Q(x)})^2=1-F(P,Q)=1-F(\rho,\sigma). \tag{5.44}$$

For the latter, note that (5.23) shows that the relation holds with equality for pure states. For ϕ and ψ the optimal pure states in the fidelity $F(\rho,\sigma)$ we therefore have $\delta(\phi,\psi)^2+F(\rho,\sigma)^2=1$. Monotonicity of the distinguishability establishes the result. □

Using (5.41) we can define a distance measure between density operators which implicitly includes their purifications as follows.

Definition 5.2.2: Purification distance

The *purification distance* $P(\rho,\sigma)$ between two states ρ and σ is defined as

$$P(\rho,\sigma):=\sqrt{1-F(\rho,\sigma)^2}. \tag{5.45}$$

This will prove useful in defining entropy quantities with nice properties.

5.3 Distinguishing between many states

In §5.1.1 we motivated the definition of distinguishability of two states via the probabilty of correctly guessing the state (preparation procedure). The case of multiple states is also of interest. Now suppose the states are chosen according with probability distribution P_X. How well can we determine X by looking at the quantum state?

More concretely, we need to measure the quantum state with a POVM, the outcome of which corresponds to our guess. Suppose the ensemble of states is described by the CQ state $\psi_{XB}=\sum_x P_X(x)|x\rangle\langle x|_X\otimes(\varphi_x)_B$ for some states φ_x. Given a POVM $\{\Lambda_x\}$ on system B, the average guessing probability is $p_{\text{guess}}=\sum_x P_X(x)\text{Tr}[\Lambda_x\varphi_x]$. We can write this more compactly using the CQ operator $\Lambda_{XB}=\sum_x|x\rangle\langle x|_X\otimes(\Lambda_x)_B$ as $p_{\text{guess}}=\text{Tr}[\Lambda_{XB}\psi_{XB}]$. The conditions that the Λ_x form a POVM are now encoded by $\Lambda_{XB}\geq 0$ and $\Lambda_B=\mathbb{1}$. The optimal guessing probability is just

Definition 5.3.1: Optimal guessing probability

For an ensemble of quantum states described by the CQ state $\psi_{XB} = \sum_x P_X(x)|x\rangle\langle x|_X \otimes \varphi_x$, the guessing probability is defined as

$$p^{\text{opt}}_{\text{guess}}(X|B)_\psi := \max \quad \text{Tr}[\Lambda_{XB}\psi_{XB}] \qquad \text{s.t.} \quad \Lambda_{XB} \geq 0, \Lambda_B = \mathbb{1}. \tag{5.46}$$

Often it is not possible to analyze the behavior of the optimal guessing probability in a particular protocol. An excellent alternative is provided by the "pretty-good measurement":

Definition 5.3.2: Pretty good measurement

For an ensemble of quantum states $\{P_X(x), \varphi_x\}$ with the average state $\varphi = \sum_x P_X(x)\varphi_x$, the prettty good measurement is the POVM with elements

$$\Lambda_x = P_X(x)\varphi^{-1/2}\varphi_x\varphi^{-1/2}. \tag{5.47}$$

The pretty good measurement has an appealing interpretation in the classical setting, i.e. when all states φ_x commute. In this case we may write $\varphi_x = \sum_y P_{Y|X=x}(y)|y\rangle\langle y|$, where the $|y\rangle$ are the common eigenstates of the φ_x and $P_{Y|X=x}(y)$ are the associated eigenvalues. Since these are positive and sum to unity, we may regard the eigenvalues as forming a conditional distribution. The average state is simply $\varphi = \sum_y P_Y(y)|y\rangle\langle y|$, and therefore we find that the POVM elements of the pretty good measurement are given by

$$\Lambda_x = \sum_y P_{X|Y=y}(x)|y\rangle\langle y|. \tag{5.48}$$

Recalling the discussion at the beginning of §5.1.1, we may regard the entire POVM as consisting of two parts. The first part is a measurement in the common eigenbasis, and the second is the generation of a guess from the measurement result. The latter step is simply to generate a guess by picking an x according to the conditional distribution $P_{X|Y=y}$ for the observed value of y, i.e. to sample from the distribution $P_{X|Y=y}$. The optimal strategy, of course, is to pick the x which maximizes $P_{X|Y=y}$.

Importantly, the pretty good measurement is indeed pretty good, as seen in the following.

Proposition 5.3.1: Quality of the pretty good measurement

For any CQ state $\rho_{XB} = \sum_x P_X(x)|x\rangle\langle x|_X \otimes (\varphi_x)_B$,

$$p^{\text{PGM}}_{\text{guess}}(X|B)_\rho \geq (p^{\text{opt}}_{\text{guess}}(X|B)_\rho)^2. \tag{5.49}$$

Proof. The proof is an application of the Cauchy-Schwartz inequality. Suppose $\{\Lambda_x\}$ is the optimal guessing POVM. Defining $B_x = \varphi^{1/4}\Lambda_x\varphi^{1/4}$ we have

$$p_{\text{guess}}(X|B)_\rho = \sum_x \text{Tr}[B_x\varphi^{-1/4}P_X(x)\varphi_x\varphi^{-1/4}]. \tag{5.50}$$

We may regard this as the inner product between the sequences $u = (B_x)_x$ and $v = (\varphi^{-1/4} P_X(x) \varphi_x \varphi^{-1/4})_x$, where the inner product is $\langle u, v \rangle = \sum_x \mathrm{Tr}[u_x^* v_x]$. Applying the Cauchy-Schwartz inequality we have

$$(p_{\mathrm{guess}}(X|B)_\rho)^2 \leq \sum_x \mathrm{Tr}[B_x^2] \sum_x \mathrm{Tr}[(\varphi^{-1/4} P_X(x) \varphi_x \varphi^{-1/4})^2] \tag{5.51}$$

$$= \sum_x \mathrm{Tr}[B_x^2] p_{\mathrm{guess}}^{\mathrm{PGM}}(X|B)_\rho. \tag{5.52}$$

For the quantity $\mathrm{Tr}[B_x^2]$ we have

$$\mathrm{Tr}[B_x^2] = \mathrm{Tr}[\varphi^{1/2} \Lambda_x \varphi^{1/2} \Lambda_x] \tag{5.53}$$

$$\leq \|\varphi^{1/2} \Lambda_x \varphi^{1/2}\|_\infty \mathrm{Tr}[\Lambda_x] \tag{5.54}$$

$$\leq \|\varphi^{1/2}\|_\infty^2 \|\Lambda_x\|_\infty \mathrm{Tr}[\Lambda_x] \tag{5.55}$$

$$\leq \mathrm{Tr}[\Lambda_x], \tag{5.56}$$

where the first inequality follows from $\varphi^{1/2} \Lambda_x \varphi^{1/2} \leq \|\varphi^{1/2} \Lambda_x \varphi^{1/2}\|_\infty \mathbb{1}$, the second from the submultiplicativity of the infinity norm, and the last from the fact that $\varphi \leq \mathbb{1}$ and $\Lambda_x \leq \mathbb{1}$. Thus, the first factor in the bound is itself upper bounded by unity, completing the proof. □

5.4 Binary hypothesis testing

In the previous sections we considered the average guessing probability when distinguishing states and channels. It is also interesting and useful to consider the various errors separately. For two states, this is the setup of binary hypothesis testing.

Here we regard the two possibilities, that the device produces ρ or that it produces σ as different hypotheses. In the statistical setting one is called the null hypothesis, say in this case that the device produces ρ, and the other is called the alternate hypothesis. We would like to perform an experiment to confirm one of these two hypotheses (or perhaps more properly, reject one of them). The hypotheses can be anything that we would like to test; a common example is that the null hypothesis is that the defendant in a criminal trial is innocent, and the alternate hypothesis is that she is guilty.

Any such experiment can make two kinds of errors. The first, called the error of type-I, occurs when the test rejects the null hypothesis, even though it is true. That is, an innocent person is convicted. The other error, called the error of type-II, occurs when the test incorrectly rejects the alternate hypothesis. This time, a guilty person is set free.

Naturally, we are interested in tests whose errors are as small as possible. For instance, we may ask for the test with the minimal type-II error, given that its type-I error is no larger than some fixed value $1 - \varepsilon$. In the case of interest, testing quantum states, the test is given by a POVM. Call the POVM elements Λ and $\mathbb{1} - \Lambda$, where Λ indicates the null hypothesis that the device produces ρ. Then the type-I error probability is $\mathrm{Tr}[\rho(\mathbb{1} - \Lambda)] = 1 - \mathrm{Tr}[\Lambda \rho]$. The minimal type-II error for type-I error no larger than $1 - \varepsilon$ is given by the following optimization problem.

Definition 5.4.1: Minimal type-II error

For any quantum state ρ, an operator $\sigma \geq 0$, and $\varepsilon \in [0,1]$,

$$\begin{aligned} \beta_\varepsilon(\rho,\sigma) \coloneqq \min\ & \mathrm{Tr}[\Lambda\sigma] \\ \text{s.t. } & \mathrm{Tr}[\Lambda\rho] \geq \varepsilon \\ & 0 \leq \Lambda \leq \mathbb{1} \end{aligned} \tag{5.57}$$

Strictly speaking, β_ε only has the type-II error probability interpretation when σ is a normalized state, whereas the definition accepts any positive σ. This will be convenient later. Clearly $\beta_\varepsilon(\rho,\sigma) \geq 0$ and $\beta_\varepsilon(\mathcal{E}(\rho),\mathcal{E}(\sigma)) \geq \beta_\varepsilon(\rho,\sigma)$ for the same reasons as in Proposition 5.1.2.

5.5 Convex optimization

5.5.1 Definition

All the definitions involving optimization that we have seen so far, Equations (5.6), (5.21), (5.46), and (5.57), are examples of *convex optimization*. In convex optimization the goal is to minimize a convex function over a convex set (or equivalently, maximize a concave function). Indeed, in all these examples the function to be optimized, the *objective function*, is linear. Hence, these are called *linear programs* (in the mathematics literature) or *semidefinite programs* (in more applied settings). The latter name arises from the fact that the convex set over which we are to optimize, called the *feasible set*, are defined by positive semidefinite constraints.[2] For the state distinguishability, optimal guessing probability, and minimal type-II error it is easy to see that the optimization is to be carried out over a convex set. Later we will see that the channel distinguishability can be formulated as a linear program as well.

Formulating these quantities as semidefinite programs is useful for two reasons. First, the optimal solutions to semidefinite programs can be efficiently computed. Specifically, the solution can be found in a time that scales polynomially with the size of the variables and the logarithm of the desired accuracy. But more importantly for us is the well-developed theory of *duality* of semidefinite programs (and linear programs more generally). Often, duality allows us to infer important properties of the quantities in question, and we will see examples of this later.

Let us now define semidefinite programming more formally in our setting.

Definition 5.5.1: Semidefinite Program

A semidefinite program over $\mathcal{H}_1$ and $\mathcal{H}_2$ is specified by two Hermitian operators $A \in \mathrm{End}(\mathcal{H}_1)$ and $B \in \mathrm{End}(\mathcal{H}_2)$ and a Hermiticity-preserving map $\mathcal{E}_{1\to 2}$. The optimization task is to

$$\begin{aligned} \text{find}\quad & \alpha = \sup \mathrm{Tr}[AX] \\ \text{subject to}\quad & \mathcal{E}(X) \leq B, \\ & X \geq 0, \\ & X \in \mathrm{End}(\mathcal{H}_1). \end{aligned} \tag{5.58}$$

[2] In more applied settings, linear programming refers to the optimization of linear functions over convex sets defined by linear constraints. The same task is sometimes called polyhedral linear programming by mathematicians, since convex sets defined by linear constraints are polyhedra.

If the feasible set is empty, we take the value of the program to be $-\infty$.

5.5.2 Duality

Every feasible X yields a lower bound on the value of the semidefinite program as we have defined it. To find upper bounds, we may appeal to the dual form of the program. Consider a positive operator $Y \in \mathscr{H}_2$. Taking the inner product with the first constraint, we have

$$\mathrm{Tr}[Y\mathscr{E}(X)] \leq \mathrm{Tr}[YB]. \tag{5.59}$$

Now suppose we find a Y such that $\mathscr{E}^*(Y) \geq A$. Taking the inner product with any feasible X we find

$$\mathrm{Tr}[AX] \leq \mathrm{Tr}[\mathscr{E}^*(Y)X] \leq \mathrm{Tr}[Y\mathscr{E}(X)] \leq \mathrm{Tr}[YB]. \tag{5.60}$$

Thus, every such Y provides an upper bound to the optimal value of the program. The best bound is given by the dual program:

Definition 5.5.2: Semidefinite Program (Dual Formulation)

The dual formulation of the semidefinite program of (5.58) is to

$$\begin{aligned} \text{find} \quad & \gamma = \inf \mathrm{Tr}[YB] \\ \text{subject to} \quad & \mathscr{E}^*(Y) \geq A, \\ & Y \geq 0, \\ & Y \in \mathrm{End}(\mathscr{H}_2). \end{aligned} \tag{5.61}$$

In this case, should the feasible set be empty, we take the value of the program to be ∞.

We have obtained the dual formulation by considering upper bounds to the original formulation (usually called the primal form); the fact that the optimal value of the dual is never less than the optimal value of the primal is called *weak duality*. This statement is clear when the supremum and infimum are attained. In the general case when the feasible sets are nonempty, it holds by definition of supremum and infimum that for all ε_1 and ε_2 there exist feasible X and Y such that $\mathrm{Tr}[AX] \geq \alpha - \varepsilon_1$ and $\mathrm{Tr}[YB] \leq \gamma + \varepsilon_2$. Combining these with (5.60) and taking the limit $\varepsilon_1, \varepsilon_2 \to 0$ gives the formal statement of weak duality.

In cases of interest *strong duality* often holds, and the optimal values of primal and dual are equal. We state this as part of the following general proposition.

Proposition 5.5.1: Duality of semidefinite programs

The following hold in the semidefinite program specified by $(\mathscr{E}, A, B)$ in (5.58):

1) *Weak duality*: $\alpha \leq \gamma$.

2) *Slater's conditions for strong duality*: If the primal feasible set is nonempty and there exists a strictly feasible dual variable Y, meaning $\mathscr{E}^*(Y) > A$, then there exists a feasible X such that $\mathrm{Tr}[XA] = \alpha$ and $\alpha = \gamma$. The same holds when swapping primal and dual.

3) *Complementary slackness*: If there exist feasible X and Y such that $\mathrm{Tr}[AX] = \mathrm{Tr}[YB]$, then $\alpha = \gamma$ and $\mathscr{E}(X)Y = BY$ and $\mathscr{E}^*(Y)X = AX$.

5.5.3 Examples

Let us see that the optimization-based quantities really are semidefinite programs. Along the way, we will see that complementary slackness often allows us to infer several interesting properties of the various quantities.

Distinguishability We start with the simplest case, the distinguishability $\delta(\rho,\sigma)$. We can bring the definition in (5.6) into the standard primal form by choosing $\mathscr{H}_2 \simeq \mathscr{H}_1$ with $A = \rho - \sigma$, $B = \mathbb{1}$, and $\mathscr{E} = \mathscr{I}$. The dual program is

$$\begin{aligned} \text{find} \quad & \inf \operatorname{Tr}[Y] \\ \text{subject to} \quad & Y \geq \rho - \sigma,\ Y \geq 0. \end{aligned} \tag{5.62}$$

If we did not already know the form of the optimal measurement, we could infer from complementary slackness that it must act as a projection on the optimal dual variable, for we have $\Lambda Y = Y$. From the dual it is clear that a feasible Y is just $\{\rho-\sigma\}_+$, the positive part of $\rho-\sigma$. This leads to an upper bound on the distinguishability which is attained by $\Lambda = \{\rho - \sigma \geq 0\}$, the projection onto this subspace.

Minimal type-II error Under our conventions, it is more natural to express the minimal type-II error as a dual program. Choosing $\mathscr{H}_2 = \mathscr{H}$, $\mathscr{H}_1 = \mathscr{H} \oplus \mathbb{C}$, the space $\mathscr{H}$ plus one extra dimension, $B = \sigma$, $A = -\mathbb{1} \oplus \varepsilon$, and $\mathscr{E}^*(\Lambda) = -\Lambda \oplus \operatorname{Tr}[\Lambda\rho]$ brings it into standard form. This example shows the importance of considering more general $\mathscr{H}_1$ and $\mathscr{H}_2$ than just whatever state spaces are involved in the problem.

Let the primal variable be of the form $X = \tau \oplus \mu$ for positive semidefinite τ and positive real μ. Note that $\operatorname{Tr}[\mathscr{E}^*(\Lambda)X] = \operatorname{Tr}[\Lambda(\mu\rho - \tau)]$ and thus $\mathscr{E}(X) = \mu\rho - \tau$. The primal formulation is then

$$\begin{aligned} \text{find} \quad & \sup \mu\epsilon - \operatorname{Tr}[\tau] \\ \text{subject to} \quad & \mu\rho - \tau \leq \sigma \\ & \mu, \tau \geq 0. \end{aligned} \tag{5.63}$$

Slackness yields the conditions $\Lambda\tau = \tau$, $\mu\operatorname{Tr}[\Lambda\rho] = \mu\epsilon$, and $(\mu\rho - \sigma)\Lambda = \tau\Lambda$, where the latter comes from $\mathscr{E}(X)\Lambda = B\Lambda$ and the former two from $\mathscr{E}^*(\Lambda)X = AX$.

Clearly the optimal τ in this program is $\tau = \{\mu\rho - \sigma\}_+$, and therefore by the first slackness condition Λ is a projector onto this subspace, plus an unspecified action on its complement. The complement consists of the "null" subspace in which $\mu\rho = \sigma$ and the subspace in which $\mu\rho < \sigma$. By the last slackness condition, the optimal Λ must annihilate the latter, but the condition places no constraints on the action of Λ on the former. However, the second slackness condition shows that Λ has precisely the required action on the null subspace so that $\operatorname{Tr}[\Lambda\rho] = \epsilon$ is achieved.

Using complementary slackness we have in fact inferred the *Neyman*[3]*-Pearson*[4] *lemma* of classical statistics. This result states that when distinguishing between probability distributions, the optimal test is always based on the likelihood ratio. That is, given a value x which is a sample from either P or Q, the optimal test decides for P or Q based on the ratio $P(x)/Q(x)$. If the ratio is above a threshold $1/\mu$, the test decides for P, and below this threshold for Q. Put differently, the test decides for P for all x such that $\mu P(x) - Q(x) > 0$. In the language of density operators, the test is based on the positivity of $\mu\rho - \sigma$, just as we have found.

[3] Jerzy Neyman, 1894 – 1981, Polish mathematician and statistician.
[4] Egon Sharpe Pearson, 1895 – 1980, British statistician.

Optimal guessing probability The optimal guessing probability is more complicated, but nearly in the desired form. Choose $\mathscr{H}_1 = \mathscr{H}_X \otimes \mathscr{H}_B$, $\mathscr{H}_2 = \mathscr{H}_B$, $A = \rho_{XB}$, $B = \mathbb{1}$ and $\mathscr{E}_{1\to 2} = \mathrm{Tr}_X$. Then we have

$$\begin{aligned} \text{maximize} \quad & \mathrm{Tr}[\rho_{XB}\Lambda_{XB}] \\ \text{subject to} \quad & \mathrm{Tr}_X[\Lambda_{XB}] \leq \mathbb{1}, \\ & \Lambda_{XB} \geq 0. \end{aligned} \tag{5.64}$$

We do not need to explicitly enforce the constraint that Λ_{XB} be a CQ operator, i.e. of the form $\Lambda_{XB} = \sum_x |x\rangle\langle x|_X \otimes \Lambda_x$. Instead, since ρ_{XB} is a CQ state, the objective function automatically takes the form $\sum_x P_X(x)\mathrm{Tr}_B[\Lambda_x \varphi_x]$, where $\Lambda_x = \mathrm{Tr}_X[|x\rangle\langle x|_X \Lambda_{XB}]$. The first constraint is nothing other than $\sum_x \Lambda_x \leq \mathbb{1}$. We also need not enforce equality in this constraint, since the optimal value of the objective function will be attained on the boundary of the feasible set, i.e. when Λ_{XB} is such that $\sum_x \Lambda_x = \mathbb{1}$. Finally, positivity of Λ_{XB} implies positivity of all the Λ_x, since $\langle x| \otimes \langle \phi| \Lambda_{XB} |x\rangle \otimes |\phi\rangle \geq 0$ for all $|\phi\rangle$.

Let Y be the dual variable, an operator on $\mathscr{H}_2$. We can also express this as $Y = \lambda\sigma$ for some positive λ and normalized state σ. In these terms the dual program is given by

$$\begin{aligned} \text{find} \quad & \inf \lambda \\ \text{subject to} \quad & \lambda \mathbb{1}_X \otimes \sigma_B \geq \rho_{XB}, \\ & \lambda, \sigma \geq 0. \end{aligned} \tag{5.65}$$

The slackness conditions yield $\lambda(\sum_x \Lambda_x)\sigma = \lambda\sigma$ and $\lambda(\mathbb{1}_X \otimes \sigma_B)\sum_x |x\rangle\langle x|_X \otimes \Lambda_x = \rho_{XB}\Lambda_{XB}$. The first one is uninteresting, but the second implies that

$$\lambda\sigma\Lambda_x = P_X(x)\varphi_x\Lambda_x \qquad \forall x. \tag{5.66}$$

Summing over x gives $Y = \sum_x P_X(x)\varphi_x\Lambda_x$. This must be feasible in the dual program, and therefore the optimal measurement has the property that $Y \geq P_X(x)\varphi_x$ for all x. This may not look like a great help in finding the optimal measurement, but it turns out to be useful.

Channel distinguishability

Fidelity

5.6 Notes & Further reading

The distinguishability of quantum states was first studied by Helstrom [48, 49]. The fidelity as defined here was first defined by Bures [50] and studied in more detail by Uhlmann [51]. A good overview of distinguishability measures of quantum states can be found in the PhD thesis of Fuchs [52]. The distinguishability of quantum channels was studied by Kitaev [53] and is often known as the *diamond norm*. It is well-known in operator theory, where it is called the *completely bounded norm*. Paulsen gives a nice overview of operator theory which, despite the more advanced mathematical setting, will be recognizable to readers of these notes [54].

The pretty good measurement was named by Hausladen and Wootters [55], though it was first considered by Belavkin [56]. In the field of frame theory the pretty good measurement is known as the *canonical tight frame* (though this only applies to the pure state case). Frame theory and the study

of wavelets are useful in quantum information theory, particularly for observables on an infinite-dimensional space, such as position and momentum of a free particle. The interested reader is invited to consult the books by Christensen [57, 58] for more on this topic.

In this chapter we only scratch the surface of the very important topic of convex optimization. For more, see Boyd and Vanderberghe [59] for an applied approach, as well as more mathematical treatments (in rough order of increasing sophistication) by Tiel [62], Rockafellar [60], and Barvinok [61].

5.7 Exercises

Exercise 5.1. Minimum-error state discrimination → solution

Suppose that Alice sends Bob a signal which is either ρ_1 with probability p_1 or ρ_2 with probability p_2. Bob would like to know which one was sent, with the smallest possible error. His measurement consists of operators E_1 and E_2 such that $E_1 + E_2 = \mathbb{1}$. If outcome E_1 occurs he guesses that Alice sent ρ_1; if E_2, ρ_2.

a) Shouldn't we consider the possibility that Bob's measurement has more than two outcomes?

b) Show that the probability of error is given by

$$p_{\text{error}} = p_1 + \sum_i \lambda_i \langle e_i | E_1 | e_i \rangle,$$

where $\{|e_i\rangle\}$ is the orthonormal basis of eigenstates of the operator $p_2\rho_2 - p_1\rho_1$, while λ_i are the corresponding eigenvalues.

c) Find the nonnegative operator E_1 that minimizes p_{error}.

d) Show that the corresponding error probability is

$$p^*_{\text{error}} = p_1 + \sum_{i:\lambda_i<0} \lambda_i = \tfrac{1}{2}(1 - \|p_2\rho_2 - p_1\rho_1\|_1).$$

Exercise 5.2. Unambiguous state discrimination → solution

Now Bob would like to decide between ρ_1 and ρ_2 as in the previous exercise, but wants to avoid the possibility of deciding incorrectly.

a) Bob's measurement surely has outcomes E_1 and E_2 corresponding to ρ_1 and ρ_2, respectively. Assuming the two states ρ_j are pure, $\rho_j = |\phi_j\rangle\langle\phi_j|$ for some $|\phi_j\rangle$, what is the general form of E_j such that $\Pr(E_j|\rho_k) = 0$ for $j \neq k$?

b) Can these two elements alone make up a POVM? Is there generally an inconclusive result $E_?$?

c) Assuming ρ_1 and ρ_2 are sent with equal probability, what is the optimal unambiguous measurement, i.e. the unambigous measurement with the smallest probability of an inconclusive result?

Exercise 5.3. Decoupling → solution

a) Show that any purification of the state $\rho_{AB} = \frac{\mathbb{1}_A}{d_A} \otimes \rho_B$ has the form

$$|\psi\rangle_{AA'BB'} = |\Phi\rangle_{AA'} \otimes |\psi\rangle_{BB'},$$

for $|\Phi\rangle_{AA'} = \frac{1}{\sqrt{d_A}} \sum_k |k\rangle_A |k\rangle_{A'}$ a maximally entangled state and $|\psi\rangle_{BB'}$ a purification of ρ_B.

b) Consider a state that is ε-close to ρ_{AB} according to the trace distance:

$$\delta\left(\sigma_{AB}, \frac{\mathbb{1}_A}{d_A} \otimes \rho_B\right) \leq \varepsilon.$$

Show that there exists a purification $|\phi\rangle_{ABA'B'}$ of σ_{AB} with purifying system $\mathscr{H}_{A'} \otimes \mathscr{H}_{B'}$ such that

$$\delta\left(|\phi\rangle_{ABA'B'}, |\Psi\rangle_{AA'} \otimes |\psi\rangle_{BB'}\right) \leq \sqrt{2\varepsilon}.$$

Exercise 5.4. Entanglement and channel distinguishability → solution

Let $\delta_{1-1}(\mathscr{E}, \mathscr{F})$ be the distinguishability of two channels $\mathscr{E}_{A'\to B}$, $\mathscr{F}_{A'\to B}$ when considering only input states on the input system A'.

a) Show that in general $\delta(\mathscr{E}, \mathscr{F}) \geq \delta_{1-1}(\mathscr{E}, \mathscr{F})$.

b) Consider the depolarizing channel on one qubit, $\mathscr{E}_p(\rho) = p\frac{\mathbb{1}}{2} + (1-p)\rho$. Compute (or at least bound) and compare $\delta(\mathscr{E}_p, \mathscr{I})$ and $\delta_{1-1}(\mathscr{E}_p, \mathscr{I})$.

6

Divergence Measures and Entropies

Similar to the previous chapter, here we consider *divergence* quantities which attempt to measure the difference between quantum states. Now the focus is on quantities which are useful in proving converse results, statements that a given set of resources cannot be used to construct some desired protocol.

The main important property of a distinguishability measure is monotonicity, the intuitive property that applying a channel to two states can only make them less distinguishable.

6.1 f-Divergence

In the classical setting, a general class of measures with this property are the f-divergences.

Definition 6.1.1: f-Divergence

For any convex function f on $\mathbb{R}_+$ with $f(1)=0$, we define the associated f-divergence

$$D_f(P,Q)=\sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right). \tag{6.1}$$

Examples include

- the *relative entropy* $D(P,Q)$ with $f : t \mapsto t\log t$
$$D(P,Q)=\sum_x P(x)(\log P(x)-\log Q(x)), \tag{6.2}$$
- the relative entropy $D(Q,P)$ with $f : t \mapsto -\log t$,
- the *chi-squared divergence* $\chi^2(P,Q)$ with $f : t \mapsto (t-1)^2$
$$\chi^2(P,Q)=\sum_x \frac{(P(x)-Q(x))^2}{Q(x)}, \tag{6.3}$$
- the variational distance (distinguishability) $\delta(P,Q)$ with $f : t \mapsto |t-1|$, and
- the *Hellinger divergences* $H_\alpha(P,Q)$ with $f_\alpha : t \mapsto (t^\alpha-1)/(\alpha-1)$ for $\alpha>0, \alpha\neq 1$
$$\mathsf{H}_\alpha(P,Q)=\frac{1}{\alpha-1}\left(\sum_x P(x)^\alpha Q(x)^{1-\alpha}-1\right). \tag{6.4}$$

The Hellinger divergences are essentially the same as the *Renyi divergences*

$$D_\alpha(P,Q) \coloneqq \frac{1}{\alpha-1}\log\sum_x P(x)^\alpha Q(x)^{1-\alpha} \qquad \alpha\geq 0, \alpha\neq 1. \tag{6.5}$$

Taking the limit $\alpha\to 1$ recovers the relative entropy. Of the Renyi divergences, the cases $\alpha = 0, \frac{1}{2}, 2, \infty$ are are especially useful. When $\alpha = 0$ we have

$$D_0(P,Q)=-\log\sum_{x:P(x)>0} Q(x), \tag{6.6}$$

i.e. minus the logarithm of the probability Q assigns to the support of P. When $\alpha = 1/2$ we essentially have the fidelity of the two distributions

$$D_{1/2}(P,Q) = -2\log\sum_x \sqrt{P(x)Q(x)} = -\log F(P,Q)^2. \tag{6.7}$$

The case $\alpha = 2$ yields the logarithm of the average value, under P, of the ratio P/Q:

$$D_2(P,Q) = \log\sum_x P(x)\frac{P(x)}{Q(x)} = \log\big(1+\chi^2(P,Q)\big). \tag{6.8}$$

where the second equality follows by comparing with the quantity $\chi^2(P,Q)$. Finally, in the limit $\alpha \to \infty$ only the x for which $P(x)/Q(x)$ is largest contributes, and we have

$$D_\infty(P,Q) = \log\max_x \frac{P(x)}{Q(x)}. \tag{6.9}$$

It can be shown that the Renyi divergences are monotonically increasing in α, but we will not use this fact here. We can easily prove the monotonicity for all f-divergences at once.

Proposition 6.1.1: Monotonicity of f-divergence

For any stochastic map (channel) $W : \mathcal{X} \to \mathcal{Y}$ and two distributions P and Q over $\mathcal{X}$,

$$D_f(WP, WQ) \le D_f(P,Q). \tag{6.10}$$

Proof. The proof is an application of Jensen's inequality, Equation (2.11), and the dilation property of stochastic maps, Proposition 2.5.2.

First, we consider monotonicity under deterministic transformations. If W implements the function g, every output y is associated with the set $g^{-1}(y) = \{x : g(x) = y\}$ of preimages. We can then write

$$D_f(P,Q) = \sum_y \sum_{x\in g^{-1}(y)} Q(x) f\left(\frac{P(x)}{Q(x)}\right). \tag{6.11}$$

Now consider the inner summation and let $P' = WP$ and $Q' = WQ$, i.e. $P'(y) = \sum_{x\in g^{-1}(y)} P(x)$ and similarly for Q'. For each of the inner summations we have

$$\sum_{x\in g^{-1}(y)} Q(x) f\left(\frac{P(x)}{Q(x)}\right) = Q'(y) \sum_{x\in g^{-1}(y)} \frac{Q(x)}{Q'(y)} f\left(\frac{P(x)}{Q(x)}\right) \tag{6.12}$$

$$\ge Q'(y) f\left(\sum_{x\in g^{-1}(y)} \frac{Q(x)}{Q'(y)}\frac{P(x)}{Q(x)}\right) \tag{6.13}$$

$$\ge Q'(y) f\left(\frac{P'(y)}{Q'(y)}\right). \tag{6.14}$$

The inequality is Jensen's, since f is convex and the quantities $Q(x)/Q'(y)$ form a probability distribution. When f is strictly convex, equality can only hold when $P(x) = Q(x)$ for all $x \in g^{-1}(y)$. Using this bound in the outer summation over y gives $D_f(P,Q) \ge D_f(WP, WQ)$ for deterministic W.

For an arbitrary channel we appeal to Proposition 2.5.2. First, observe that

$$D_f(P,Q) = D_f(P\times R, Q\times R) \tag{6.15}$$

for any distribution R. Then, using the dilation W' of W, where the auxiliary random variable Z has distribution R, we find

$$\begin{aligned} D_f(P,Q) &= D_f(P\times R, Q\times R) && (6.16)\\ &\geq D_f(W'(P\times R), W'(Q\times R)) && (6.17)\\ &\geq D_f(\mathrm{Tr}_Z[W'(P\times R)], \mathrm{Tr}_Z[W'(Q\times R)]) && (6.18)\\ &= D_f(WP, WQ). && (6.19)\end{aligned}$$

In the penultimate step, the channel Tr_Z denotes marginalizing the distribution over the Z random variable. This is a deterministic map, since we can think of the marginalization as sending every $z\in\mathscr{Z}$ to 1. The two inequalities follow from the established monotonicity under deterministic channels. □

Corollary 6.1.1: Nonnegativity and joint convexity of f-divergence

Every f-divergence satsifies the following properties:

- Nonnegativity: $D_f(P,Q)\geq 0$ with equality iff $P=Q$ for strictly convex f, and
- Joint convexity: For $0\leq\lambda\leq 1$, distributions P_1, P_2, Q_1 and Q_2 with $P=\lambda P_1+(1-\lambda)P_2$ and similarly for Q,

$$D_f(P,Q)\leq \lambda D_f(P_1,Q_1)+(1-\lambda)D_f(P_2,Q_2). \tag{6.20}$$

Proof. For the former, applying the marginalization map Tr gives $D_f(P,Q)\geq D_f(\mathrm{Tr}P,\mathrm{Tr}Q) = f(1)=0$. The equality condition follows from the comments following (6.14).

For the latter, let Y be an auxiliary random variable and define

$$P'_{X,Y}(x,y) = \begin{cases} \lambda P_1(x) & y=0\\ (1-\lambda)P_2(x) & y=1\end{cases}, \tag{6.21}$$

and similarly for $Q'_{X,Y}$. Then joint convexity follows from monotonicity, since simple calculation reveals that

$$D_f(P',Q') = \lambda D_f(P_1,Q_1)+(1-\lambda)D_f(P_2,Q_2). \tag{6.22}$$

□

Interestingly, although the minimal type-II error is not an f-divergence, it can generate any of them according to the following formula. Letting $\beta'_\varepsilon(P,Q)=\frac{\mathrm{d}\beta_\varepsilon(P,Q)}{\mathrm{d}\epsilon}$,

$$D_f(P,Q) = \int_0^1 \mathrm{d}\varepsilon\, \beta'_\varepsilon(P,Q)\, f\left(\frac{1}{\beta'_\varepsilon(P,Q)}\right). \tag{6.23}$$

We will not make any use of this formula here, other than to highlight the importance of the function $\varepsilon \mapsto \beta_\varepsilon(P,Q)$.

Quantum f-divergences exist and one can prove monotonicity for them, but to do so would take us too far into matrix analysis. Instead, we will examine a few cases that have found application in quantum information processing.

6.2 Quantum divergences

Several useful quantum divergence measures can be constructed by adapting the Renyi divergences. A common and very useful choice is $\alpha = 1$, which leads to the quantum relative entropy. First we mention a few other options, to get a feel for the set of different possibilities:

$$D_0(\rho,\sigma) \coloneqq -\log \mathrm{Tr}[\Pi_\rho \sigma], \tag{6.24}$$

$$D_{1/2}(\rho,\sigma) \coloneqq -\log F(\rho,\sigma)^2, \tag{6.25}$$

$$D_2(\rho,\sigma) \coloneqq \log \mathrm{Tr}[(\sigma^{-1/4}\rho\sigma^{-1/4})^2], \tag{6.26}$$

$$D_\infty(\rho,\sigma) \coloneqq \log\{\min \lambda : \rho \leq \lambda\sigma\}. \tag{6.27}$$

Here Π_ρ is the projector onto the support of ρ, which we could also express as ρ^0. Note that D_2 makes a specific choice in the order of operators, which is necessary since ρ and σ do not commute. However, the particular choice here is by no means the only possibility.

Now we discuss the relative entropy in more detail, beginning with the formal definition.

Definition 6.2.1: Relative entropy

The relative entropy $D(\rho,\sigma)$ of two states ρ and σ is defined as

$$D(\rho,\sigma) \coloneqq \mathrm{Tr}[\rho(\log\rho - \log\sigma)]. \tag{6.28}$$

In the event that the support of σ is strictly contained in the support of ρ, we set $D(\rho,\sigma) = \infty$. Additionally, the definition also applies when σ is not a normalized state.

As with the distinguishability and fidelity, the relative entropy of two noncommuting states ρ and σ is equal to the relative entropy of two particular classical distributions P and Q. However, in contrast to those cases, there is no quantum operation mapping ρ to P and σ to Q. Expressing ρ and σ in their respective eigenbases as $\rho = \sum_k r_k |u\rangle\langle u|_k$ and $\sigma = \sum_j s_j |v\rangle\langle v|_j$, the distributions in question are

$$P_{XY}(x,y) = r_x |\langle u_x | v_y\rangle|^2 \tag{6.29}$$

$$Q_{XY}(x,y) = s_y |\langle u_x | v_y\rangle|^2. \tag{6.30}$$

Since $D(\rho,\sigma) = D(P,Q)$, it immediately follows that $D(\rho,\sigma) \geq 0$. In fact, equality holds only if $\rho = \sigma$.

The key property of the relative entropy is monotonicity under quantum operations. In this course we shall not attempt a proof; again, it would take us too far into matrix analysis.

Proposition 6.2.1: Monotonicity of relative entropy

For any states ρ and σ and quantum operation $\mathcal{E}$,

$$D(\mathcal{E}(\rho),\mathcal{E}(\sigma)) \leq D(\rho,\sigma). \tag{6.31}$$

The relative entropy is actually closely related to the hypothesis testing scenario. The following proposition states the fact that the relative entropy governs the exponential decay rate of the type-II error, when the type-I error is held fixed. We shall also take the statement in the quantum case without proof. The classical case is Exercise...

Proposition 6.2.2: "Quantum Stein's Lemma"

For any quantum states ρ and σ and $\varepsilon \in (0,1)$,

$$\lim_{n\to\infty} \tfrac{1}{n} \log \beta_\varepsilon(\rho^{\otimes n}, \sigma^{\otimes n}) = -D(\rho,\sigma). \tag{6.32}$$

It is also possible to find a relation between $\beta_\varepsilon(\rho,\sigma)$ and any monotonic divergence, by taking the quantum operation to be the optimal test. That is, suppose that Q is the optimal test in $\beta_\varepsilon(\rho,\sigma)$ and define the quantum operation $\mathcal{E}(\tau) = \mathrm{Tr}[Q\tau]|0\rangle\langle 0| + (1-\mathrm{Tr}[Q\tau])|1\rangle\langle 1|$. This gives $\mathcal{E}(\rho) = \varepsilon|0\rangle\langle 0| + (1-\varepsilon)|1\rangle\langle 1|$ and $\mathcal{E}(\sigma) = \beta_\varepsilon(\rho,\sigma)|0\rangle\langle 0| + (1-\beta_\varepsilon(\rho,\sigma))|1\rangle\langle 1|$. The relative entropy of two binary distributions such as $\mathcal{E}(\rho)$ and $\mathcal{E}(\sigma)$ is easy to compute. Supposing the two distributions are $(p,1-p)$ and $(q,1-q)$, then

$$d(p,q) \coloneqq D((p,1-p),(q,1-q)) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}. \tag{6.33}$$

Therefore, monotonicity implies that $d(\varepsilon,\beta_\varepsilon(\rho,\sigma)) \leq D(\rho,\sigma)$. It can be shown that $d(p,q) \geq -h(p) + a\log\frac{1}{b}$ for $h(p) = -p\log p - (1-p)\log(1-p)$ (often called the binary entropy; see the following section), and so we have the relation

$$\varepsilon \log \frac{1}{\beta_\varepsilon(\rho,\sigma)} \leq D(\rho,\sigma) + h(\varepsilon). \tag{6.34}$$

This will prove useful in bounding the performance of protocols for communication over noisy channels.

6.3 von Neumann & Shannon entropies

The most useful and well-studied entropy is the von Neumann entropy, or equivalently the Shannon entropy in the classical setting.

Definition 6.3.1: von Neumann / Shannon entropies

The von Neumann entropies (Shannon entropies in the classical case) are defined as follows.

- The entropy $H(A)_\rho$ of a single system A of dimension d in state ρ:

$$H(A)_\rho \coloneqq -D(\rho,\mathbb{1}), \tag{6.35}$$

- the conditional entropy $H(A|B)_\rho$ of system A conditioned on system B:

$$H(A|B)_\rho := -D(\rho_{AB}, \mathbb{1}_A \otimes \rho_B), \qquad \text{and} \tag{6.36}$$

- the mutual information $I(A:B)_\rho$ of A and B:

$$I(A:B)_\rho := D(\rho_{AB}, \rho_A \otimes \rho_B). \tag{6.37}$$

Note that we can write $H(A)_\rho = \log d - D(\rho, \frac{1}{d}\mathbb{1})$ and $H(A|B)_\rho = \log d - D(\rho_{AB}, \frac{1}{d}\mathbb{1}_A \otimes \rho_B)$.

Proposition 6.3.1: Properties of the entropy of a single system

1) Positivity: $H(A)_\rho \geq 0$ for all ρ, with equality iff ρ is a pure state,
2) Unitary invariance: $H(A)_{U\rho U^*} = H(A)_\rho$ for unitary U,
3) Upper bound: $H(A)_\rho \leq \log|\operatorname{supp}\rho|$,
4) Concavity: $H(A)_\rho \geq \sum_k p_k H(A)_{\rho_k}$ for $\rho = \sum_k p_k \rho_k$, and
5) Increase under projective measurement: $H(A)_{\rho'} \geq H(A)_\rho$ where $\rho' = \sum_k \Pi_k \rho \Pi_k$ for any complete set of projectors Π_k.

Proof. These are proven in Exercise 6.1. □

Proposition 6.3.2: Properties of the entropy of several systems

1) Duality: $H(A)_\rho = H(B)_\rho$ for ρ_{AB} pure,
2) Subadditivity: $H(AB)_\rho \leq H(A)_\rho + H(B)_\rho$ with equality if $\rho_{AB} = \rho_A \otimes \rho_B$,
3) Triangle inequality: $H(AB)_\rho \geq |H(A)_\rho - H(B)_\rho|$.

Proof. Since the entropy is a function only of the eigenvalues of the reduced state, duality follows from the form of the purification, Proposition 3.4.1, and the comments following (4.11).

For subadditivity, a simple computation shows that $D(\rho_{AB}, \rho_A \otimes \rho_B) = H(A)_\rho + H(B)_\rho - H(AB)_\rho$, where as usual $\rho_A = \operatorname{Tr}_B[\rho_{AB}]$ and similarly for ρ_B. Thus, positivity of the relative entropy implies subadditivity of the entropy.

The triangle equality follows from subadditivity by making use of duality. Let R be a third purifying reference system, so that $|\psi\rangle_{RAB}$ is a purification of ρ_{AB}. Then

$$H(B)_\psi = H(RA)_\psi \leq H(A)_\psi + H(R)_\psi = H(A)_\psi + H(AB)_\psi, \tag{6.38}$$

which implies that $H(AB)_\rho \geq H(B)_\rho - H(A)_\rho$. Swapping A and B in the proof gives the absolute value. It is also easy to see that $H(AB)_\rho = H(A)_\rho + H(B)_\rho$ for states of the form $\rho_{AB} = \rho_A \otimes \rho_B$. □

Proposition 6.3.3: Chain rules of von Neumann entropy

We have the following *chain rules* relating the mutual information and conditional entropy to unconditional entropies:

$$I(A:B)_\rho = H(A)_\rho + H(B)_\rho - H(AB)_\rho, \qquad \text{and} \tag{6.39}$$
$$H(A|B)_\rho = H(AB)_\rho - H(B)_\rho. \tag{6.40}$$

Proof. Straightforward calculation. □

Proposition 6.3.4: Properties of the conditional entropy

1) Duality: $H(A|B)_\rho = -H(A|C)_\rho$ for ρ_{ABC} pure,
2) Bounds: $-\log d \leq H(A|B)_\rho \leq \log d$ for $d = \dim(A)$,
3) Conditioning on a classical system: $H(B|X)_\rho = \sum_x P_X(x) H(B)_{\varphi_x}$ for $\rho_{XB} = \sum_x P_X(x)|x\rangle\langle x|_X \otimes (\varphi_x)_B$, and
4) Conditional entropy of a classical system: $H(X|B)_\rho \geq 0$ for ρ_{XB} a CQ state.

Proof. By the chain rule, the first statement is equivalent to $H(AB)_\rho - H(B)_\rho = -H(AC)_\rho + H(C)_\rho$, whence we can see that the previously established duality is equivalent to the conditional version.

The upper bound on the conditional entropy follows from positivity of the relative entropy and the expression $H(A|B)_\rho = \log d - D(\rho_{AB}, \frac{1}{d}\mathbb{1}_A \otimes \rho_B)$. The lower bound follows by using the chain rule.

The third property follows by straightforward calculation.

To establish the positivity of the conditional entropy for CQ states, consider the purification of a generic CQ state $\rho_{XB} = \sum_x P_X(x)|x\rangle\langle x|_X \otimes (\varphi_x)_B$:

$$|\psi\rangle_{XABR} = \sum_x \sqrt{P_X(x)}|x\rangle_X |x\rangle_A |\varphi_x\rangle_{BR}. \tag{6.41}$$

Here A is an additional system which purifies X, while R purifies B. We can regard this state as the result of measuring A in the standard basis, i.e. applying the Stinespring isometry $V_{A\to XA}|x\rangle_A = |x\rangle_X|x\rangle_A$ to

$$|\psi'\rangle_{ABR} = \sum_x \sqrt{P_X(x)}|x\rangle_A |\varphi_x\rangle_{BR}. \tag{6.42}$$

Since projective measurement increases entropy, it follows that $H(AR)_\psi \geq H(AR)_{\psi'}$. Moreover, since entropy is invariant under unitaries and, by the same reasoning, isometries, it follows that $H(XAR)_\psi = H(AR)_{\psi'}$. Then we have

$$\begin{aligned} H(X|B)_\rho &= H(X|B)_\psi && (6.43)\\ &= H(XB)_\psi - H(B)_\psi && (6.44)\\ &= H(AR)_\psi - H(XAR)_\psi && (6.45)\\ &\geq H(AR)_{\psi'} - H(AR)_{\psi'} = 0. && (6.46) \end{aligned}$$

□

Proposition 6.3.5: Monotonicity of mutual information and conditional entropy

For two quantum operations $\mathcal{E}_{A\to A'}$ and $\mathcal{F}_{B\to B'}$ and a quantum state ρ_{AB}, let $\rho'_{A'B'} = \mathcal{E}_{A\to A'} \otimes \mathcal{F}_{B\to B'}(\rho_{AB})$. Then

$$I(A':B')_{\rho'} \leq I(A:B)_{\rho}. \tag{6.47}$$

Furthermore, if $\mathcal{E}_{A\to A'}$ is unital,

$$H(A'|B')_{\rho'} \geq H(A|B)_{\rho}. \tag{6.48}$$

These inequalities are often called *data processing* inequalities, since processing data (random variables or quantum states) only increases entropy or decreases mutual information. They also equivalent to the *strong subadditivity* of the entropy, subadditivity for conditional entropy:

$$H(AB|C)_{\rho} \leq H(A|C)_{\rho} + H(B|C)_{\rho}. \tag{6.49}$$

When ρ_{ABC} is a CQ state with C classical, strong subadditivity follows from usual subadditivity using Property 3 of Proposition 6.3.4. In the general case, it is easy to work out the statement of strong subadditivity is equivalent to $H(B|AC)_{\rho} \leq H(B|C)_{\rho}$ (or $H(A|BC)_{\rho} \leq H(A|C)_{\rho}$), which is just monotonicity under the partial trace map.We can express this even more compactly by defining the *conditional mutual information*

$$I(A:C|B)_{\rho} \coloneqq H(A|B)_{\rho} - H(A|BC)_{\rho} \tag{6.50}$$

$$= H(B|A)_{\rho} - H(B|AC)_{\rho}. \tag{6.51}$$

Strong subadditivity is then just the statement $I(A:B|C)_{\rho} \geq 0$.

6.4 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 (not simultaneously!) of the other to their shared entanglement.

Proposition 6.4.1: 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 $H(X|B)_{\rho}$ the entropy of the result of measuring X on A conditional on system B, and similarly

for $H(Z|C)$, we have

$$H(X|B)_\rho + H(Z|C)_\rho \geq \log \frac{1}{c(X,Z)}, \quad \text{and} \tag{6.52}$$

$$H(X|B)_\rho + H(Z|B)_\rho \geq \log \frac{1}{c(X,Z)} + H(A|B)_\rho. \tag{6.53}$$

For the proof we shall need the following lemma.

Lemma 6.4.1. *For a quantum state ρ and positive operators σ, σ' such that $\sigma' \geq \sigma$,*

$$D(\rho,\sigma') \leq D(\rho,\sigma). \tag{6.54}$$

Proof. Let $\mathscr{H}$ be the state space on which ρ, σ, σ' are defined and suppose that $\mathscr{H}' \simeq \mathscr{H}$. We can then extend the space $\mathscr{H}$ by embedding it into $\mathscr{H} \oplus \mathscr{H}'$. For states on the bigger space, the relative entropy obeys

$$D(\rho,\sigma) = D(\rho \oplus 0, \sigma \oplus \tau) \tag{6.55}$$

for any $\tau \geq 0$, since only that part of the second state in the support of the first is relevant to the relative entropy. Now let $\{|b_k\rangle\}$ and $\{|b'_k\rangle\}$ be bases of $\mathscr{H}$ and $\mathscr{H}'$, respectively, so that their union is a basis of $\mathscr{H} \oplus \mathscr{H}'$. Defining $\Pi = \sum_k |b_k\rangle\langle b_k|$ and $V = \sum_k |b_k\rangle\langle b'_k|$, the map $\mathscr{E}(\eta) = \Pi\eta\Pi + V\eta V^*$ is a quantum operation since it is defined by a Kraus representation with $\Pi^*\Pi + V^*V = \mathbb{1}_{\mathscr{H}\oplus\mathscr{H}'}$. Finally, letting $\tau = \sigma' - \sigma$, the desired result follows by monotonicity and $\mathscr{E}(\sigma \oplus \tau) = \sigma'$. □

Proof of Proposition 6.4.1. The proof proceeds by showing the first statement and then deriving the second as a simple consequence. To prove the first, observe that by data processing it is sufficient to establish the statement for pure ρ_{ABC}. Then consider the states

$$|\psi\rangle_{XX'BC} \coloneqq V_{A\to XX'}|\rho\rangle_{ABC} = \sum_x |x\rangle_X |x\rangle_{X'A} \langle\varphi_x|\rho\rangle_{ABC} \tag{6.56}$$

$$|\xi\rangle_{ZZ'BC} \coloneqq U_{A\to ZZ'}|\rho\rangle_{ABC} = \sum_z |z\rangle_Z |z\rangle_{Z'A} \langle\vartheta_z|\rho\rangle_{ABC}, \tag{6.57}$$

where $V_{A\to XX'}$ ($U_{A\to ZZ'}$) is a Stinespring dilation of the X (Z) measurement process. By definition, $H(X|B)_\rho = H(X|B)_\psi$ and $H(Z|C)_\rho = H(Z|C)_\xi$.

Entropy duality implies $H(X|B)_\psi + H(X|X'C)_\psi = 0$, and thus

$$\begin{aligned} H(X|B)_\rho &= -H(X|X'C)_\psi && (6.58)\\ &= D(\psi_{XX'C}, \mathbb{1}_X \otimes \psi_{X'C}) && (6.59)\\ &= D(\rho_{AC}, V^*(\mathbb{1}_X \otimes \psi_{X'C})V) && (6.60)\\ &= D(\xi_{ZZ'C}, UV^*(\mathbb{1}_X \otimes \psi_{X'C})VU^*) && (6.61)\\ &\geq D(\xi_{ZC}, \mathrm{Tr}_{Z'}[UV^*(\mathbb{1}_X \otimes \psi_{X'C})VU^*]). && (6.62) \end{aligned}$$

Here we have used monotonicity in the last step and invariance of the relative entropy under isometries in the previous steps. The second argument in the final expression is just

$$\begin{aligned}
\mathrm{Tr}_{Z'}[UV^*(\mathbb{1}_X \otimes \psi_{X'C})VU^*] &= \sum_z |z\rangle\langle z|_Z \otimes \langle \vartheta_z|V^*(\mathbb{1}_X \otimes \psi_{X'C})V|\vartheta_z\rangle_A && (6.63)\\
&= \sum_z |z\rangle\langle z|_Z \otimes \langle \vartheta_z|\left(\sum_x |\varphi_x\rangle\langle\varphi_x|_Z \otimes \langle x|\psi_{X'C}|x\rangle_{X'}\right)|\vartheta_z\rangle_A && (6.64)\\
&= \sum_{xz} |\langle\varphi_x|\vartheta_z\rangle|^2 |z\rangle\langle z|_Z \otimes \langle x|\psi_{X'C}|x\rangle_{X'} && (6.65)\\
&\leq c(X,Z)\sum_{xz} |z\rangle\langle z|_Z \otimes \langle x|\psi_{X'C}|x\rangle_{X'} && (6.66)\\
&= c(X,Z)\mathbb{1}_Z \otimes \psi_C && (6.67)\\
&= c(X,Z)\mathbb{1}_Z \otimes \xi_C && (6.68)
\end{aligned}$$

By Lemma 6.4.1, we therefore have

$$\begin{aligned}
H(X|B)_\rho &\geq D(\xi_{ZC}, c(X,Z)\mathbb{1}_Z \otimes \xi_C) && (6.69)\\
&= D(\xi_{ZC}, \mathbb{1}_Z \otimes \xi_C) - \log c(X,Z) && (6.70)\\
&= -H(Z|C)_\xi + \log\frac{1}{c(X,Z)}, && (6.71)
\end{aligned}$$

completing the proof of the first statement.

For the second statement, it is a simple calculation to verify that $H(Z^A B)_\rho = H(Z^A C)_\rho$ when C is the purification of AB so that ρ_{ABC} is pure. This leads immediately to $H(Z^A|C)_\rho = H(Z^A|B)_\rho - H(A|B)_\rho$. Using this expression to replace $H(Z^A|C)_\rho$ in the first statement leads to the second. □

6.5 Min and max entropies

Applying the definition of the conditional von Neumann entropy in (6.36) to the $\alpha = \infty$ and $\alpha = 1/2$ quantum divergences leads to other useful entropic quantities, the min and max entropy, respectively. Actually, it proves convenient to add an additional optimization, as follows.

Definition 6.5.1: Min and max entropies

For a bipartite quantum state ρ_{AB} the conditional min and max entropies $H_{\min}(A|B)_\rho$ and $H_{\max}(A|B)_\rho$ are defined as follows.

$$\begin{aligned}
H_{\min}(A|B)_\rho &:= \max_\sigma -D_\infty(\rho_{AB}, \mathbb{1}_A \otimes \sigma_B) && (6.72)\\
&= \max_\sigma \log\{\min\lambda : \rho_{AB} \leq \mathbb{1}_A \otimes \sigma_B\} && (6.73)\\
H_{\max}(A|B)_\rho &:= \max_\sigma -D_{1/2}(\rho_{AB}, \mathbb{1}_A \otimes \sigma_B) && (6.74)\\
&= \max_\sigma \log F(\rho_{AB}, \mathbb{1}_A \otimes \sigma_B)^2. && (6.75)
\end{aligned}$$

Indeed, we could have included the optimization over σ_B in the defnition of the conditional von Neumann entropy, as Exercise 6.2 shows that doing so makes no difference to the resulting quantity.

To get a feeling for what these entropies quantify, consider the case that B is trivial. Looking back at (6.9), we see that $H_{\min}(A)_\rho = 1/\lambda_{\max}$ for $\lambda_{\max}$ the largest eigenvalue of ρ. This is a measure of how deterministic the distribution corresponding to the eigenvalues of ρ is, how close it is to a deterministic distribution. Meanwhile, from (6.7) we see that $H_{\max}(A)_\rho = \log F(\rho, \mathbb{1})^2 = d_A \log F(\rho, \frac{1}{d_A}\mathbb{1})^2$, a measure of how close the distribution is to uniform.

In fact, we have met the conditional min entropy before, at least for classical-quantum states ρ_{XB}, in (5.65). Thus, the conditional min entropy is related to the optmial probability of guessing the classical random variable X by measuring system B. Formally,

$$p^{\text{opt}}_{\text{guess}}(X|B)_\rho = 2^{-H_{\min}(X|B)_\rho}. \tag{6.76}$$

Later, it will be convenient to use *smoothed* versions of the min and max entropies. These are defined as optimizations over nearby states ρ', using the purification distance from (5.45) to quantify "nearby".

Definition 6.5.2: Smooth min and max entropies

For a bipartite quantum state ρ_{AB} the smooth conditional min and max entropies $H^\varepsilon_{\min}(A|B)_\rho$ and $H^\varepsilon_{\max}(A|B)_\rho$ are defined as follows.

$$H^\varepsilon_{\min}(A|B)_\rho := \max_{\rho' : P(\rho,\rho') \leq \varepsilon} H_{\min}(A|B)_{\rho'} \tag{6.77}$$

$$H^\varepsilon_{\max}(A|B)_\rho := \min_{\rho' : P(\rho,\rho') \leq \varepsilon} H_{\max}(A|B)_{\rho'}. \tag{6.78}$$

6.6 Entropic measures of entanglement

6.6.1 Negative conditional entropy

A corollary of monotonicity is the concavity of the conditional entropy. Just as in Corollary 6.1.1, monotonicity implies joint convexity, and thus $H(A|B)_\rho$ must be a concave function of ρ, since it is the negative of a convex function.

Corollary 6.6.1: Concavity of conditional entropy

Suppose $\rho_{XB} = \sum_x P_X(x)(\rho_x)_{AB}$. Then

$$H(A|B)_\rho \geq \sum_x P_X(x) H(A|B)_{\rho_x} \tag{6.79}$$

This result is interesting because it tells us that negative $H(A|B)_\rho$ is a sign of entanglement of ρ_{AB} (which we might well have suspected already). If ρ_{AB} is pure, then $H(A|B)_\rho \leq 0$ implies $H(B)_\rho > 0$ (indeed $-H(A|B)_\rho = H(B)_\rho$) and therefore there is more than one Schmidt coefficient. For the case of mixed states, consider $H(A|B)_\rho$ for a separable state $\rho_{AB} = \sum_k \lambda_k \sigma_k \otimes \xi_k$:

$$H(A|B)_\rho \geq \sum_k \lambda_k H(A|B)_{\sigma_k \otimes \xi_k} = \sum_k \lambda_k H(A)_{\sigma_k} \geq 0. \tag{6.80}$$

Therefore $H(A|B)_\rho < 0$ implies ρ_{AB} is not separable, i.e. ρ_{AB} is entangled.

However, the converse is false: There exist entangled states for which $H(A|B)_\rho \geq 0$. Thus, the conditional entropy is not a *faithful* measure of entanglement. Nonetheless, the duality of conditional entropy translates into the *monogamy* property of entanglement: A system A cannot be entangled with both B and C at the same time.

6.6.2 Squashed entanglement

Separable states can be created by purely local operations on A and B and classical communication between them. For instance, to create the state $\rho_{AB} = \sum_k \lambda_k \sigma_k \otimes \xi_k$, one party can sample from the distribution $P(k) = \lambda_k$ to obtain a value of k, create the state σ_k and then communicate to the other party that the state ξ_k is to be created.

Therefore, it is desirable to have a measure of entanglement which cannot increase under local operations and classical communication (LOCC). One possibility is the *squashed entanglement*:

Definition 6.6.1: Squashed entanglement

The squashed entanglement of a state ρ_{AB} is defined by

$$E_{\mathrm{sq}}(A:B) := \tfrac{1}{2} \inf_E I(A:B|E), \tag{6.81}$$

where the infimum extends over all extensions ρ_{ABE} of ρ_{AB}.

Note that we do not impose a limit on the dimension of E, and hence write inf instead of min since it is not apparent that the minimum can actually be achieved. (In fact, it is.) For a maximally entangled state $|\Phi\rangle_{AB}$, all extensions are of the form $\rho_{ABE} = |\Phi\rangle\langle\Phi|_{AB} \otimes \sigma_E$ for some state σ_E. Therefore, $E_{\mathrm{sq}}(A:B)_\Phi = \log d$. This is the maximal value of the conditional mutual information, so it is satisfying that the maximally entangled state has the maximal value of the squashed entanglement. Conversely, it is known that the squashed entanglement is zero if and only if the state is separable, i.e. it is a faithful measure. Showing one direction of this statement is Exercise 6.6. We proceed to establish the monotonicity under LOCC operations.

Proposition 6.6.1: Monotonicity of squashed entanglement under LOCC

For a bipartite quantum state ρ_{AB} and any LOCC operation $\mathcal{E}_{AB\to A'B'}$,

$$E_{\mathrm{sq}}(A':B')_{\rho'} \leq E_{\mathrm{sq}}(A:B)_\rho. \tag{6.82}$$

Proof. By the definition of the conditional mutual information in (6.50), monotonicity of the conditional entropy implies the squashed entanglement cannot increase by purely local operations on system B. This also holds for local operations on A by symmetry of the definition.

It remains to show that the squashed entanglement does not increase under classical communication. Suppose that Alice sends a classical system C (e.g. a bit string) to Bob. We can model this by a tripartite system ρ_{ABC} where Alice has possession of C prior to the communication, and Bob afterwards. Monotonicity under classical communication then becomes the statement $E_{\mathrm{sq}}(AC:B)_\rho \geq E_{\mathrm{sq}}(A:BC)_\rho$.

To show this, start with any extension E and apply monotonicity like so:

$$I(B:AC|E)_\rho = H(B|E)_\rho - H(B|ACE)_\rho \tag{6.83}$$
$$\geq H(B|EC)_\rho - H(B|AEC)_\rho \tag{6.84}$$
$$= I(B:A|EC)_\rho. \tag{6.85}$$

Since C is classical, the state has the form $\rho_{ABCE} = \sum_k p_k(\rho_k)_{ABE} \otimes |k\rangle\langle k|_C$, which we may extend to $\rho'_{ABCC'E} = \sum_k p_k(\rho_k)_{ABE} \otimes |k\rangle\langle k|_C \otimes |k\rangle\langle k|_{C'}$. Equivalently, we can generate C' from C in ρ_{ABC}, which implies $H(A|BCC'E)_{\rho'} = H(A|BCE)_\rho$. It follows that $I(BC':A|EC)_{\rho'} = I(B:A|EC)_\rho$ by using the chain rule. Observe that in this step it is crucial that C be a classical system.

Defining $E' = EC'$ to be a new extension and using the fact that we can interchange C and C', we therefore have $I(B:AC|E)_\rho \geq I(BC:A|E')_{\rho'}$ and consequently $E_{\text{sq}}(AC:B) \geq E_{\text{sq}}(A:BC)$, which completes the proof. □

6.7 Notes & Further reading

Divergences have a long history in information theory and statistics. The most widely-used, the relative entropy, was introduced by Kullback and Leibler in the context of statistics [63]. Its precise connection to asymmetric hypothesis testing, known as Stein's lemma after the statistician Charles Stein, was shown by Chernoff [64]. General f-divergences were introduced independently by Csiszár [65], Morimoto [66], and Ali and Silvey [67]. The fact that the type-II error as a function of the type-I error is sufficient to reconstruct any f-divergence, Equation (6.22), is proven in [68, Theorem 11], a detailed overview of the properties of f-divergence measures. Our treatment of monotonicity follows the introductory text of Csiszár and Shields [69].

Divergences are widely used in the field of *information geometry* which applies the methods of differential geometry to probability theory. See [70] for more, including a discussion of information geometry for quantum systems. Petz extended the notion of f-divergence to the quantum setting and showed that many of the important results, such as monotonicity, still hold [71, 72].

Entropy as we have defined it was first defined by Gibbs in the context of classical statistical mechanics [73] and later extended to the quantum setting by von Neumann [29]. Shannon introduced entropy as a quantification of uncertainty or information [74]. He used the name entropy on the advice of von Neumann, who told him,

> 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.[75]

In the next chapter we will take the position that any quantity which appears in a converse statement for an information processing task is an "entropy", particularly when the information processing task is known to be achievable up to the bound given by the converse.

The triangle inequality for the von Neumann entropy was proven by Araki and Lieb [76], and was one of the first uses of the purification of a quantum state (as in the present proof). Strong subadditivity was first shown by Lieb and Ruskai [77].

6.8 Exercises

Exercise 6.1. Properties of the von Neumann entropy → solution

Show that the properties of the von Neumann entropy listed in Proposition 6.3.1 indeed hold. Moreover, show that equality holds in the concavity statement if the states ρ_k are all pairwise disjoint, i.e. $\rho_j \rho_k = 0$ for $j \neq k$.

Hint: The latter three properties follow from positivity of the relative entropy. For the last property, use the fact that ρ' commutes with the projectors Π_k.

Exercise 6.2. Optimization in the conditional von Neumann entropy → solution

Show that $H(A|B)_\rho = \max_\sigma -D(\rho_{AB}, \mathbb{1}_A \otimes \sigma_B)$ for any bipartite state ρ_{AB}.

Exercise 6.3. Quantum mutual information → solution

a) Prove that the mutual information of the Bell state $|\Phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ is maximal.

b) Show that $I(A:B)_\rho \leq 1$ for classically correlated states, $\rho_{AB} = p|0\rangle\langle 0|_A \otimes \sigma_B^0 + (1-p)|1\rangle\langle 1|_A \otimes \sigma_B^1$ (where $0 \leq p \leq 1$).

c) Consider the so-called *cat state* of four qubits, $A \otimes B \otimes C \otimes D$, defined as

$$|\text{cat}\rangle = \frac{1}{\sqrt{2}}(|0000\rangle + |1111\rangle). \tag{6.86}$$

Show that the mutual information between A and B changes with the knowledge of the remaining qubits, in that $I(A:B) = 1$, $I(A:B|C) = 0$, but $I(A:B|CD) = 1$. How should we interpret these results?

Exercise 6.4. Data processing for classical mutual information → solution

a) First show the chain rule for mutual information: $I(X:YZ) = I(X:Z) + I(X:Y|Z)$, which holds for arbitrary classical random variables X, Y, Z.

b) By expanding the mutual information $I(X:YZ)$ in two different ways, prove the data processing in equality.

Exercise 6.5. Fano's inequality → solution

Given random variables X and Y, how well can we predict X given Y? Fano's inequality bounds the probability of error in terms of the conditional entropy $H(X|Y)$. The goal of this exercise is to prove the inequality

$$P_{\text{error}} \geq \frac{H(X|Y) - 1}{\log|X|}.$$

a) Representing the guess of X by the random variable $\widehat{X}$, which is some function, possibly random, of Y, show that $H(X|\widehat{X}) \geq H(X|Y)$.

b) Consider the indicator random variable E which is 1 if $\widehat{X} \neq X$ and zero otherwise. Using the chain rule we can express the conditional entropy $H(E, X|\widehat{X})$ in two ways:

$$H(E,X|\widehat{X}) = H(E|X,\widehat{X}) + H(X|\widehat{X}) = H(X|E,\widehat{X}) + H(E|\widehat{X})$$

Calculate each of these four expressions and complete the proof of the Fano inequality.
Hints: For $H(E|\widehat{X})$ use the fact that conditioning reduces entropy: $H(E|\widehat{X}) \leq H(E)$. For $H(X|E,\widehat{X})$ consider the cases $E = 0, 1$ individually.

Exercise 6.6. Squashed entanglement of separable states → solution
Show that the squashed entanglement of any separable state is zero.

Exercise 6.7. A sufficient entanglement criterion → solution
In general it is very difficult to determine if a state is entangled or not. In this exercise we will construct a simple entanglement criterion that correctly identifies all entangled states in low dimensions.

a) Let $\mathscr{F}_A : \mathrm{End}(\mathscr{H}_A) \to \mathrm{End}(\mathscr{H}_A)$ be a positive superoperator. Show that $\mathscr{F}_A \otimes \mathscr{I}_B$ maps separable states as defined in (4.9) to positive operators.

This means that if we apply $\mathscr{F}_A \otimes \mathscr{I}_B$ to a bipartite state ρ_{AB} and obtain a non-positive operator, we know that ρ_{AB} is entangled. In other words, this is a sufficient criterion for entanglement.

b) Apply the partial transpose, $\mathscr{T}_A \otimes \mathscr{I}_B$, to the ε-noisy Bell state

$$\rho_{AB}^{\varepsilon} = (1-\varepsilon)|\Phi\rangle\langle\Phi|_{AB} + \varepsilon \, \tfrac{1}{4}\mathbb{1}_{AB}.$$

For what values of ε can we be sure that ρ^{ε} is entangled?

Remark: Indeed, it can be shown that the PPT criterion (positive partial transpose) is necessary and sufficient for bipartite systems of dimension 2×2 and 2×3.

7 Information Processing Protocols

In this chapter we present a framework for understanding information processing protocols and discuss several particularly important cases.

7.1 Background: The problem of reliable communication & storage

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.[74]

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 communication 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, setting up Shannon's approach,

> 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.[78]

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 the intended message can be correctly inferred by the receiver.

7.2 The resource simulation approach

Once we have phrased the problem of communication in terms of reproducing message symbols, it is clear that the *ideal* communication channel simply reproduces the input symbol at the output. This is a particular physical operation when, following the theme of this course, the messages we intend to send are instantiated in some physical degrees of freedom. The actual channel at our disposal will invariably fall short of this ideal due to noise and imperfections. Nonetheless, it is still a physical transformation, so we may model it by a quantum operation, or a stochastic map if the inputs and outputs are classical.

[1] Claude Elwood Shannon, 1916 – 2001, American mathematician and electrical engineer.
[2] Ralph Vinton Lyon Hartley, 1888 – 1970, American electrical engineer.

The goal in building a communication system is to *simulate* the ideal channel by using the actual noisy channel and whatever other *resources* the sender and receiver have at their disposal. In particular, the sender is able to apply any physical operation mapping the message degrees of freedom to the degrees of freedom input to the channel. And the receiver may apply any physical operation to the degrees of freedom at the output of the channel to whatever message degrees of freedom he or she desires. Call the input (output) message degrees of freedom M (M') and A (B) the channel input (output). Thus, for a given noisy channel $\mathcal{N}_{A\to B}$, the sender and receiver would like to find "encoding" and "decoding" operations $\mathcal{E}_{M\to A}$ and $\mathcal{D}_{B\to M'}$ such that $\mathcal{D}\circ\mathcal{N}\circ\mathcal{E}$ is essentially $\mathcal{I}_{M\to M'}$, as depicted in Figure 7.1.

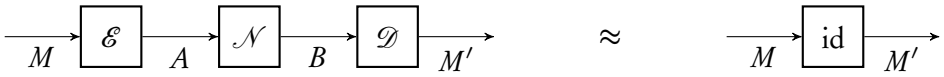


Figure 7.1: Noisy channel coding

This example is the prototype of the "resource simulation" approach to information processing. We begin by selecting an ideal resource that we would like to have (the ideal channel), and we then attempt to construct a *protocol* (in this case the encoder and decoder) such that in the actual resource (the noisy channel) can be made to simulate the ideal resource. The protocol consists of a set of physical operations which connect the input (output) degrees of freedom of the ideal resource to those of the actual resources. A generic ideal resource simulation is depicted in Figure 7.2.

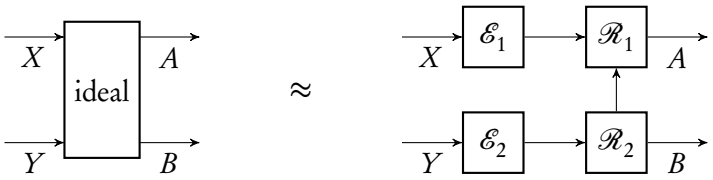


Figure 7.2: A generic simulation of an ideal resource by two given resources $\mathcal{R}_1$ and $\mathcal{R}_2$ with a protocol consisting of operations $\mathcal{E}_1$ and $\mathcal{E}_2$.

The important aspect of this approach is that *the quality of a protocol is measured by its ability to allow the real resources to simulate the ideal resource.* Although this seems like the most naive approach, in fact it is common in the study of information theory and especially cryptography that *ad hoc* measures of the quality of a protocol are introduced for each particular task. The great advantage of focusing on simulation is that constructed resources can be *composed* to make new resources. Given an ideal resource consisting of several parts, we can build a protocol to simulate the whole by constructing protocols to simulate each part.

For instance, suppose we would like to transmit a given source of data (e.g. an audio file) from a sender to a receiver who are connected by a noisy channel. The original is still retained by the sender, so we are really interested in making a copy of the source available to the receiver. Let us call the output of the source, a random variable, X, and the copy X'. The ideal behavior is depicted on the righthand side of Figure 7.3. We can simulate the ideal behavior with the noisy channel coding protocol of Figure 7.1, choosing $|M| = |X|$. Since the identity channel faithfully transmits any input message, not just those we expect from the source, $X' = X$.

However, we can reduce the necessary number of messages that that the noiseless channel is required to transmit by *compressing* the output of the source. Figure 7.3 depicts the goal of the data

compression task. By finding appropriate compression and decompression operations $\mathscr{C}$ and $\mathscr{D}$ we can potentially find a C with $|C| \leq |X|$. The combined protocol is then formed by replacing the wire transmitting C from $\mathscr{C}$ to $\mathscr{D}$ in Figure 7.3 with the noisy channel coding protocol of Figure 7.1.

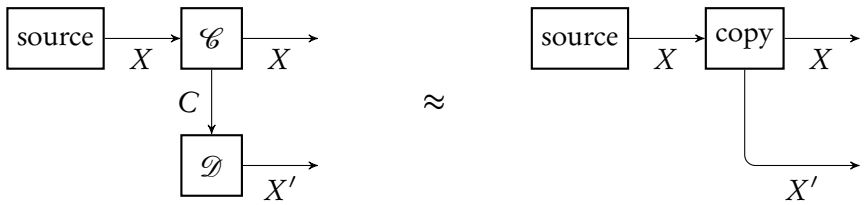


Figure 7.3: Data compression. The random variable C is the compressed version of X, since the decompressor $\mathscr{D}$ can recreate the particular output of the source (stored in X) from it.

For composability of resources to be defined formally, we must choose a suitable measure of "simulatability", the ability of one channel to simulate another. The distinguishability introduced in Chapter 5 exactly satisfies this requirement, because it satisfies the triangle inequality and monotonicity. We say a channel $\mathscr{N}$ *ϵ-approximates* another $\mathscr{N}'$ if $\delta(\mathscr{N}, \mathscr{N}') \leq \epsilon$. This notion of distinguishability has an explicitly operational interpretation since it governs the probability with which any experiment can tell the difference between $\mathscr{N}$ and $\mathscr{N}'$. Loosely speaking, we can interpret ϵ as the probability that $\mathscr{N}$ fails to precisely simulate $\mathscr{N}'$. Then, if we combine two resources that each simulate ideal versions with parameters ϵ_1 and ϵ_2, respectively, the overall approximation parameter will be no worse than $\epsilon_1 + \epsilon_2$.

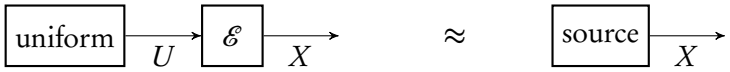
In analyzing any given information processing task in the resource framework, there are two main questions.

- First, what resources are absolutely necessary for performing the task?
  This is the question of establishing a *converse*, a statement that quantifies the properties needed by any collection of actual resources that can simulate a chosen ideal resource.
- Second, can we actually construct protocols that meet the bounds of the converse statements?
  This is the question of *achievability*.

The quantities involved in converses are typically entropies of some kind. In fact, we can regard this as the definition and say that *any quantity that appears in a converse is an "entropy", particularly when the achievability statement matches the converse bound*. This is the situation in classical thermodynamics, where the name "entropy" was coined by Clausius[3] in 1850. There entropy arises from the most famous converse statement, the second law of thermodynamics. In many textbooks the focus is first on defining a quantitative measure of information and then analyzing information processing protocols. However, this seems logically backwards, for a given quantity can only be regarded as a useful measure of information if it arises in an operational context.

In the remainder of the chapter we examine these two questions for several important information processing tasks in detail. Figure 7.4 depicts several information processing tasks in the resource simulation approach.

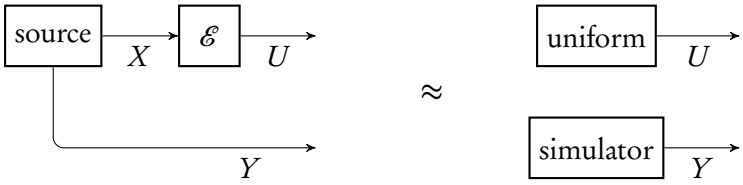
[3] Rudolf Julius Emanuel Clausius, 1822 – 1888, German physicist and mathematician.



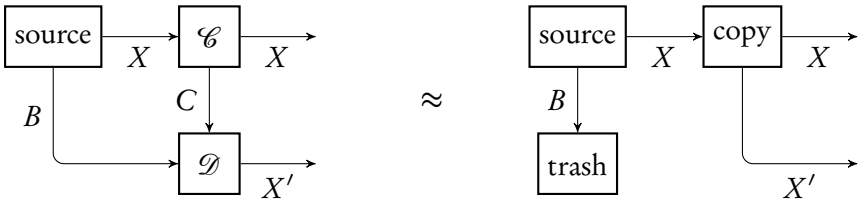
(a) Resolvability, or generating a random variable X from the uniform distribution. The map $\mathcal{E}$ must be deterministic for the resource simulation task to be nontrivial, since otherwise it could just ignore U and prepare X itself.

(b) Randomness extraction. A uniformly-distributed random variable is constructed from the source output X. Again, the map $\mathcal{E}$ must be deterministic for the simulation task to be nontrivial.

(c) "Deletion" coding. For all inputs, the output of the deletion channel is zero. The map $\mathcal{E}$ must be reversible (on its image) for the resource simulation task to be nontrivial, since otherwise it could just map all m to a fixed input to A. The uniform randomness is used to aid in masking the input from the output. The simulator operation prepares a state close to the output of $\mathcal{N} \circ \mathcal{E}(m, U)$ in system B.



(d) Privacy amplification. Randomness uncorrelated with Y is extracted from X. As with randomness extraction, $\mathcal{E}$ must be deterministic. The simulator prepares a state close to the Y output of the source.



(e) Information reconciliation, or data compression with side information. The decompressor now has access to system B output by the source, and which is in general correlated with X.

Figure 7.4: More simple information processing tasks.

7.3 Optimality of superdense coding and teleportation

In this section we use properties of the von Neumann entropy to show that the superdense coding and teleportation protocols of §3.6 are optimal. First, it is useful to express the resource statements of the two protocols as inequalities. For instance, the protocol of superdense coding allows one to simulate two ideal classical single-bit channels from Alice to Bob with one maximally-entangled state

of two qubits and one ideal qubit channel. Thus, we may write

$$[qq] + [q \to q] \geq 2[c \to c], \tag{7.1}$$

where $[qq]$ denotes the entangled state, $[q \to q]$ the ideal quantum channel and $[c \to c]$ the ideal classical channel. Similarly, the resource inequality for teleportation reads

$$2[c \to c] + [qq] \geq [q \to q]. \tag{7.2}$$

The inequality expresses the fact that there exists a protocol which transforms the resources listed on the lefthand side into the resource(s) on the righthand side. Superdense coding and teleportation are *exact* protocols, meaning the ideal resources are simulated perfectly. We can indicate that only simulation with distinguishability ϵ is achieved by writing

$$2[c \to c] + [qq] \gtrsim_\epsilon [q \to q]. \tag{7.3}$$

7.3.1 Optimality of superdense coding

The resource inequality for any scheme for simulating $2[c \to c]$ using resources $[qq]$ and $[q \to q]$ can be written

$$a[qq] + b[q \to q] \geq 2[c \to c], \tag{7.4}$$

for some $a, b \geq 0$. Since a and b ought to be integers (half of an ideal quantum channel makes no sense), this is not much of a generalization. However, we can interpret the case of rational a and b by multiplying the resource inequality through with a large enough integer n such that the resulting inequality has integer coefficients. In fact, we must have $a + b \geq 2$ and $b \geq 1$, as we now show.

The first constraint, $a + b \geq 2$, follows from the *Holevo*[4] *bound*, a corollary of monotonicity of the relative entropy for CQ states (though first established independently of monotonicity). It reads

Corollary 7.3.1: Holevo bound

For any CQ state ρ_{XB} and POVM $\{\Lambda_y\}$ on system B producing the random variable Y and joint state ρ'_{XY},

$$I(X:B)_\rho \geq I(X:Y)_{\rho'}. \tag{7.5}$$

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. The precise argument is left Exercise 7.1.

Now we show that $b \geq 1$. If we concatenate our hypothetical superdense coding scheme (7.4) with the standard teleportation protocol (7.2), i.e. by replacing the classical communication in teleportation by superdense coding, we arrive at the resource inequality

$$(1+a)[qq] + b[q \to q] \geq [q \to q]. \tag{7.6}$$

Therefore, to infer that $b \geq 1$, it suffices to show that shared entanglement does not help with quantum communication.

[4] Alexander Semenovich Holevo, born 1943, Russian mathematician.

An ideal quantum channel can transmit entanglement just as well as any particular pure state, so let us consider the a scenario in which Alice shares n entangled qubits with a third party, Charlie, that she would like to transfer to Bob. She already shares k entangled qubits with Bob and is willing to send him m qubits more through a noiseless channel. Let A and C be the entangled systems shared by Alice and Charlie, A' and B' the entangled systems shared by Alice and Bob. To tranfer the Charlie entanglement to Bob, Alice performs some quantum operation $\mathcal{E}_{AA'\to Q}$ which outputs an m-qubit system Q that she transmits to Bob. He then performs some quantum operation $\mathcal{F}_{B'Q\to B}$ which should result a n qubit system B which is entangled with C.

Assuming the protocol does work as intended, the quantum mutual information between B and C at the end of the protocol is $2n$. Using the properties of mutual information we further have

$$\begin{aligned} 2n &= I(C:B) && (7.7)\\ &\leq I(C:B'Q) && (7.8)\\ &= I(C:B') + I(C:Q|B') && (7.9)\\ &= H(Q|B') - H(Q|CB') && (7.10)\\ &\leq \log d_Q - (-\log d_Q) && (7.11)\\ &= 2m. && (7.12) \end{aligned}$$

The first inequality is monotonicity, while the second equality uses the chain rules and definition of conditional mutual information. Since C is independent of B', $I(C:B') = 0$. The second inequality comes from the upper and lower bounds on the conditional entropy in Proposition 6.3.4.

Thus, we have shown that the mutual information between Bob and Charlie cannot increase by more than two times the number of qubits Bob receives, regardless of the amount of shared entanglement. Returning to (7.6), this implies $b \geq 1$. Quantities with the opposite behavior, that can increase sharply when only a few qubits are communicated, are known as *lockable*, in the sense that absent the extra few qubits the information (as measured by the quantity in question) is locked in the system and cannot be extracted. An example is the *accessible information* of a CQ state, $I_{\text{acc}}(X:B) = \max_\Lambda I(X:Y)$, the classical mutual information of the optimal measurement. We have just shown that the quantum mutual information is *nonlockable*.

7.3.2 Optimality of teleportation

As with superdense coding, we consider the resource inequality for any protocol which simulates an ideal qubit channel using shared entanglement and classical communication,

$$c[qq] + d[c\to c] \geq [q\to q]. \tag{7.13}$$

In this case we can show that $c \geq 1$ and $d \geq 2$ must hold.

The latter is similar to the argument for $b \geq 1$ in superdense coding. Consider concatenating this hypothetical teleportation protocol with the standard superdense coding protocol, i.e. replacing the quantum channel in superdense coding with the teleportation scheme. This leads to the resource inequality

$$(1+c)[qq] + d[c\to c] \geq 2[c\to c]. \tag{7.14}$$

Thus, to show that $d \geq 2$, it suffices to show that entanglement does not help with transmitting classical information. The argument is similar to that for quantum communication used in the analysis of superdense coding and is included in Exercise 7.2.

To show that $c \geq 1$, consider using the ideal quantum channel simulated by the hypothetical teleportation protocol (7.13) to transmit halves of entangled states. Doing so yields the resource inequality

$$c[qq] + d[c \to c] \geq [qq]. \tag{7.15}$$

Therefore, to show $c \geq 1$ it suffices to show that the use of classical communication cannot by itself lead to an increase in the number of maximally entangled qubit pairs. This certainly sounds plausible, but we are interested in a rigorous statement. We can make such a statement by appealing to the properties of the squashed entanglement, defined in (6.81).

Indeed, the desired result is immediate. Since the squashed entanglement is monotonic under local operations and classical communication, as shown in Proposition 6.6.1, the squashed entanglement of the lefthand side of (7.15) must be larger than the righthand side. But the squashed entanglement of a maximally entangled qubit pair is unity, and therefore $c \geq 1$ is necessary for the resource inequality to hold. In fact, the statement also holds if one requires the transformation to only work approximately. The proof is more technical and requires the continuity of squashed entanglement.

7.4 Compression of classical data

Compression of classical data, possibly with side information, is depicted in Figure 7.4(e). This task is also known as information reconciliation, because in the case of classical B we can view the task as reconciling the classical value B with that of X (i.e. making the former equal to the latter). The idea is for Alice, who has X, to send Bob, who has B, enough "information" C about X to enable him to reconstruct X from B and C. Clearly, Alice could just send X, but the goal is complete the task with as small a C as possible.

The source produces a CQ state

$$\psi^{XB} = \sum_{x \in \mathscr{X}} P_X(x) |x\rangle\langle x|^X \otimes \varphi_x^B, \tag{7.16}$$

for some fixed distribution P_X and set of states φ_x. The protocol consists of a compression operation $\mathscr{C}$ which generates the classical random variable C from X, and a decompression operation $\mathscr{D}$ which acts on C and B to produce X'. The input and output of the former are classical, so we can model it as a classical channel with conditional probability distribution $P_{C|X}$. As the output X' is classical, we can model the decompressor as a measurement of BC. However, as C is also classical, without loss of generality the measurement has elements $\Gamma_{x,c}^{BC} = \Lambda_{x;c}^B \otimes |c\rangle\langle c|^C$. That is, we can view the measurement of BC as consisting of two steps: The value of C is first determined, and conditioned on this value the POVM with elements $\Lambda_{x;c}$ is performed on B. The outcomes of the latter POVM must be labelled with elements of $\mathscr{X}$.

It is perhaps surprising that the such a scheme is possible at all, since Alice does not know exactly what it is that Bob needs to know. For example, suppose that the source produces two classical bit strings of length n; Alice's string is random and Bob's differs in at most t positions. If Alice knew in *which* t positions Bob's string differed from hers, then the protocol would be simple. However, even by sending a sufficient amount of essentially random information about her string (in the form of the output of a randomly-chosen function), Bob can combine this information with his string to determine Alice's string.

We can gain a better understanding by considering the protocol from Bob's point of view. His system is in one of the states φ_x, but he is unsure which. Furthermore, the states are generally not

distinguishable, so he cannot just measure the system to determine x with high reliability. The information he receives from Alice narrows the set of possible states, making the distinguishing task simpler. Since both parties know which state is produced by the source, for each x Alice also knows how likely Bob is to correctly determine that x. She just needs to sufficiently narrow the set possible states at his end to make the guessing probability close to unity.

The distinguishability of the actual resource from the ideal is easily computed to indeed be the probability that $X' \neq X$. The ideal resource is described by the distribution $P_{XX'}(x,x') = P_X(x)\delta_{x,x'}$. Let us denote by $Q_{XX'}$ the distribution produced by the actual resource. Using the state and the description of the compressor and decompressor maps, we have

$$Q_{XX'}(x,x') = \sum_{x,c,x'} P_X(x) P_{C|X=x}(c) \mathrm{Tr}[\Lambda_{x';c}\varphi_x] \tag{7.17}$$

$$= \sum_{x,x'} P_X(x) \mathrm{Tr}[\hat{\Lambda}_{x'}\varphi_x]. \tag{7.18}$$

Here, the conditional distribution $P_{C|X}$ describes the compressor, and we have implicitly defined the POVM with elements $\hat{\Lambda}_{x'} = \sum_c P_{C|X=x}(c)\Lambda_{x';c}$. The distinguishability between $P_{XX'}$ and $Q_{XX'}$ is then

$$\delta(P_{XX'}, Q_{XX'}) = \tfrac{1}{2}\sum_{x,x'} |P_{XX'}(x,x') - Q_{XX'}(x,x')| \tag{7.19}$$

$$= \tfrac{1}{2}\sum_x P_X(x) \sum_{x'} |\delta_{x,x'} - \mathrm{Tr}[\hat{\Lambda}_{x'}\varphi_x]| \tag{7.20}$$

$$= \tfrac{1}{2}\sum_x P_X(x) \Big((1 - \mathrm{Tr}[\hat{\Lambda}_x\varphi_x]) + \sum_{x'\neq x} \mathrm{Tr}[\hat{\Lambda}_{x'}\varphi_x] \Big) \tag{7.21}$$

$$= \sum_{x,x'\neq x} P_X(x) \mathrm{Tr}[\hat{\Lambda}_{x'}\varphi_x], \tag{7.22}$$

which is the probability that $X' \neq X$ under $Q_{XX'}$.

7.4.1 Converse to compression of classical data

For a given CQ source ψ^{XB} we are interested in the smallest $|C|$ such that it is possible to construct a compression map $\mathscr{C} : X \to C$ and decompression map $\mathscr{D} : BC \to X'$ such that the probability of $X' \neq X$ is less than ϵ. What tradeoffs of $|C|$ in terms of ϵ do we face for a given source ψ^{XB}? Put differently, given ϵ and ψ^{XB} is a desired value of $|C|$ even possible?

We can find a constraint based on using any possible compression scheme to construct a hypothesis testing measurement for two particular states associated with the source.

Proposition 7.4.1: Converse to compression of classical data

Any compression scheme for X in the CQ state ψ^{XB} with average error ϵ obeys

$$|C| \geq \max_\sigma \beta_{1-\epsilon}(\psi^{XB}, \mathbb{1}^X \otimes \sigma^B). \tag{7.23}$$

Note that none of the quantities involve the compression and decompression operations, i.e. the protocol. They are purely properties of the real resource ψ^{XB} as well as the error probability ϵ.

Proof. Consider the task of distinguishing between ψ^{XB} and any operator of the form $\mathbb{1}^X \otimes \sigma^B$. From monotonicity, it follows that

$$\beta_{1-\epsilon}(\psi^{XB}, \mathbb{1}^X \otimes \sigma^B) \leq \beta_{1-\epsilon}(\psi^{XCB}, \mathscr{C}(\mathbb{1}^X) \otimes \sigma^B), \tag{7.24}$$

where the state ψ^{XCB} is the state produced by the compressor $\mathscr{C}_{X\to XC}$ applied to ψ^{XB}. In particular,

$$\psi^{XCB} = \sum_{x,c} P_X(x) P_{C|X=x}(c) |x\rangle\langle x|^X \otimes |c\rangle\langle c|^C \otimes \varphi_x^B. \tag{7.25}$$

Now define the following test for the setup involving XCB:

$$Q^{XCB} = \sum_{x,c} |x\rangle\langle x|^X \otimes |c\rangle\langle c|^C \otimes \Lambda_{x;c}^B. \tag{7.26}$$

It is straightforward to see that $\mathrm{Tr}[Q^{XCB}\psi^{XCB}] = \sum_x Q_{XX'}(x,x)$, which is just $1-\epsilon$ by assumption. Therefore, Q^{XCB} is feasible for finding the minimum type-II error in $\beta_{1-\epsilon}(\psi^{XCB}, \mathscr{C}(\mathbb{1}^X) \otimes \sigma^B)$. To bound this quantity, first observe that

$$\mathscr{C}(\mathbb{1}^X) = \sum_{xc} P_{C|X=x}(c) |x\rangle\langle x|^X \otimes |c\rangle\langle c|^C \leq \mathbb{1}^{XC}. \tag{7.27}$$

Then we have

$$\begin{aligned}
\beta_{1-\epsilon}(\psi^{XCB}, \mathscr{C}(\mathbb{1}^X) \otimes \sigma^B) &\leq \mathrm{Tr}[Q^{XCB} \mathscr{C}(\mathbb{1}^X) \otimes \sigma^B] && (7.28)\\
&\leq \mathrm{Tr}[Q^{XCB} \mathbb{1}^{XC} \otimes \sigma^B] && (7.29)\\
&= \sum_{x,c} \mathrm{Tr}[\Lambda_{x;c}\sigma] && (7.30)\\
&= \sum_c \mathrm{Tr}[\sigma] && (7.31)\\
&= |C|. && (7.32)
\end{aligned}$$

Since this holds for arbitrary σ, the desired statement holds. □

7.4.2 Achievability of compressing classical data

To show that protocols exist which essentially meet the converse bound in Proposition 7.4.1, we follow the method of *random coding* originally used by Shannon. The basic idea is to show that if the compression function is chosen at random, a suitable decompression scheme leads to low error probability on average, and therefore there must exist at least one compressor whose error probability is no larger than the average value. This is a commonly-used trick in analyzing algorithms in computer science, where it is called the *probabilistic method*.

Observe that the compressor can be chosen to be a deterministic function without loss of generality. This follows from the representation of classical channels as convex combinations of deterministic channels, Proposition 2.5.2. Consider the dilation of $\mathscr{C}$ with auxiliary random variable Z. The overall distinguishability of the scheme will just be an average, over P_Z, of the error probability for the operation of the compression scheme given the value of $Z = z$. But at least one value z must have an error probability less than the average value, and so the encoder might as well fix Z to this optimal value. The result is a deterministic encoder. In the random coding argument, we may therefore choose deterministic compression functions according to whatever distribution we like.

For the decompressor we will use a variant of the pretty good measurement of Definition 5.3.2 based on the optimal test in a certain hypothesis testing scenario.

Proposition 7.4.2: Achievability of compressing classical data

For any CQ state ψ^{XB}, desired error probability ϵ, and any $\eta \leq \epsilon$, there exists a deterministic compressor $\mathscr{C}_{X\to C}$ and decompressor $\mathscr{D}_{CB\to X'}$ with

$$|C| = \left\lceil \frac{4\epsilon}{\eta^2} \beta_{1-\epsilon+\eta}(\psi^{XB}, \mathbb{1}^X \otimes \psi^B) \right\rceil. \tag{7.33}$$

Comparing with the converse from Proposition 7.4.1, the expression contains nearly the same hypothesis testing type-II error, but with the additional contribution η. One can show that the function $\varepsilon \mapsto \beta_\varepsilon$ increases monotonically, meaning that absent the prefactor, the expression is larger than that in the converse for $\eta > 0$. But the dependence of the prefactor on η prohibits us from taking $\eta \to 0$ without increasing the size of the compressor output.

Proof. The proof proceeds in three steps. First, we analyze the error probability using a fixed deterministic compressor and decompressor constructed from the pretty good measurement for an arbitrary set of operators. Next, we simplify the expression by averaging over the compression functions. Finally, we make a specific choice of the operators forming the decompressor to make the connection to the type-II error.

Consider an arbitrary positive CQ operator $Q^{XB} = \sum_x |x\rangle\langle x|^X \otimes Q_x^B$. Given a compression function $f : X \to C$, we define the decompressor to be the POVM with elements

$$\Gamma_x^{BC} = \sum_c \delta_{f(x),c} (\hat{Q}_c^{-1/2} Q_x \hat{Q}_c^{-1/2})^B \otimes |c\rangle\langle c|^C, \tag{7.34}$$

where $\hat{Q}_c = \sum_{x:f(x)=c} Q_x$. The error probability using this compressor and decompressor is given by

$$p_{\text{error}} = \sum_x P_X(x) \text{Tr}[(\mathbb{1} - \Gamma_x)^{BC} (\varphi_x^B \otimes |f(x)\rangle\langle f(x)|^C)]. \tag{7.35}$$

Now apply Lemma A.10.1 to $\mathbb{1} - \Gamma_x$ in each term of the sum. Setting $S = Q_x$ and $T = \hat{Q}_c - Q_x$ gives

$$\mathbb{1} - \Gamma_x \leq \sum_c \delta_{f(x),c} \left((1+a) Q_x^B + (2 + a + a^{-1}) \sum_{x' \neq x : f(x') = c} Q_{x'}^B \right) \otimes |c\rangle\langle c|^C, \tag{7.36}$$

and leads to an upper bound on the error probability

$$p_{\text{error}} \leq \sum_x P_X(x) \left((1+a) \text{Tr}[Q_x \varphi_x] + (2 + a + a^{-1}) \sum_{x' \neq x : f(x) = f(x')} \text{Tr}[Q_{x'} \varphi_x] \right) \tag{7.37}$$

$$= (1+a) \text{Tr}[Q^{XB} \psi^{XB}] + (2 + a + a^{-1}) \sum_x P_X(x) \sum_{x' \neq x} \delta_{f(x)=f(x')} \text{Tr}[Q_{x'} \varphi_x]. \tag{7.38}$$

Next, we average over a uniformly random choice of f, denoting the averaged error probability by angle brackets. Only the quantity $\delta_{f(x)=f(x')}$ for any distinct x and x' is affected. It is unity if f takes them to the same value and zero otherwise; averaged uniformly over all functions its expected value is just $1/|C|$. To see this, observe that fixing the output value of all inputs (including x) not

equal to x' specifies $|C|$ possible functions, one for each value that x' can take. The chance x' takes the same value as x is thus $1/|C|$. Therefore we have

$$\langle p_{\text{error}} \rangle \leq (1+a)\text{Tr}[Q^{XB}\psi^{XB}] + (2+a+a^{-1})\frac{1}{|C|}\sum_{x,x'\neq x} P_X(x)\text{Tr}[Q_{x'}\varphi_x] \tag{7.39}$$

$$\leq (1+a)\text{Tr}[Q^{XB}\psi^{XB}] + (2+a+a^{-1})\frac{1}{|C|}\sum_{x,x'} P_X(x)\text{Tr}[Q_{x'}\varphi_x] \tag{7.40}$$

$$= (1+a)\text{Tr}[Q^{XB}\psi^{XB}] + (2+a+a^{-1})\frac{1}{|C|}\text{Tr}[Q^{XB}(\mathbb{1}^X \otimes \psi^B)]. \tag{7.41}$$

Finally, we make a specific choice for Q^{XB}: the optimal test in $\beta_{1-\xi}(\psi^{XB}, \mathbb{1}^X \otimes \psi^B)$. In the first term we can substitute $\text{Tr}[Q^{XB}\psi^{XB}] = 1-\xi$ and in the second $\text{Tr}[Q^{XB}(\mathbb{1}^X \otimes \psi^B)] = \beta_{1-\xi}(\psi^{XB}, \mathbb{1}^X \otimes \psi^B)$. Altogether, we can infer the existence of at least one encoding function f such that

$$p_{\text{error}} \leq (1+a)\xi + (2+a+a^{-1})\frac{1}{|C|}\beta_{1-\xi}(\psi^{XB}, \mathbb{1}^X \otimes \psi^B). \tag{7.42}$$

Now we choose a, ξ and $|C|$ such that the righthand side equals ϵ. A nice choice is $\xi = \epsilon - \eta$, $a = \eta/(2\epsilon - \eta)$ and $|C|$ as in (7.33). □

A few comments on the decompression operation are in order. First, when system B is trivial (i.e. all φ_x are identical to ψ^B), the minimal type-II error reduces to $\beta_{1-\xi}(\psi^X, \mathbb{1}^X)$. Using the form of the optimal test derived in §5.5.3, we find that the optimal Q_x which generate the pretty good measurement all have the form $Q_x^B = q_x \mathbb{1}^B$ for $q_x = 1$ when $P_X(x) > 1/\mu$ for the optimal cutoff value μ. For $P_X(x) < 1/\mu$, $q_x = 0$, and otherwise q_x takes a value necessary to attain $\text{Tr}[Q^{XB}\psi^{XB}] = 1-\xi$. Only those Q_x with $f(x) = c$ go into the pretty good measurement for fixed value of c. Thus, apart from x values that are considered too unlikely ($P_X(x) \leq 1/\mu$), the decompression map simply guesses randomly between all possible x which are consistent with the given value of c.

For nontrivial B, we can get an intuition for why the optimal Q_x in the hypothesis test generate a useful measurement by considering an example involving classical states, depicted in Figure 7.5. Suppose that there are three equally-likely states φ_x, each a normalized Gaussian of width σ separated from its neighbor by a distance 4σ. (Continuous distributions are outside the scope of this course, but in this example no difficulties arise.) The optimal test Q^{XB} for a given μ will have Q_x which project onto the region $\{y : \mu\varphi_x(y) - 3\overline{\varphi} \geq 0\}$; that is, $Q_x(y)$ is an indicator function onto this region. Informally, the region consists of those y for which φ_x takes on a value that substantially exceeds the value taken by the average of the states. These regions can overlap, as indicated in the figure, and for small ξ will include most of the support of each state φ_x. The resulting pretty good measurement simply guesses between the two states in the overlapping regions, so a balance has to be found between ξ and the resulting error probability of the measurement. In the example shown in the figure, $\mu = 3$ which leads to $\xi \approx 0.008$, $\beta_{1-\xi}(\psi^{XB}, \mathbb{1} \otimes \psi^B) \approx 3$, and a guessing probability of approximately 95% under the pretty good measurement formed from the Q_x.

7.4.3 Compression of i.i.d. sources

For a source which is N independent and identical instances of a fixed source ψ^{XB}, i.e. $(\psi^{XB})^{\otimes N}$, we can show that the converse and achievability bounds meet in the limit $N \to \infty$. Let $\ell(\psi^{XB}, \epsilon, N)$ be the smallest $|C|$ such that there exists an ϵ-good compression scheme for $\psi_{XB}^{\otimes N}$. It is more convenient

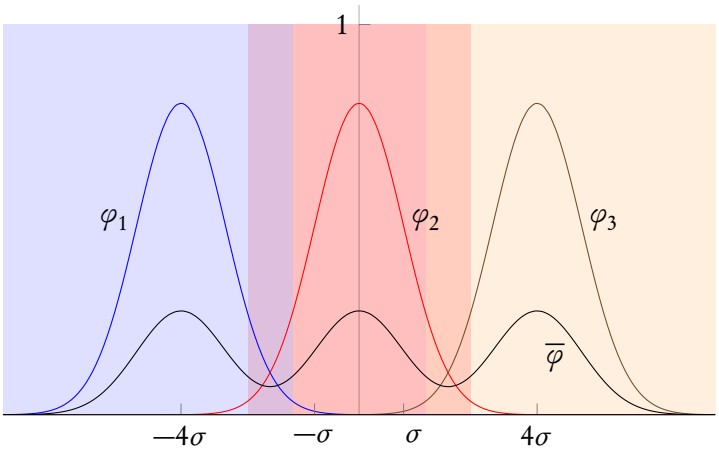


Figure 7.5: Hypothesis testing scenario for $\psi^{XB} = \frac{1}{3}\sum_x |x\rangle\langle x|^X \otimes \varphi_x^B$ versus $\mathbb{1}^X \otimes \psi^B$ with the φ_x Gaussian functions of width σ separated by 4σ. The average state (function) ψ^B is also denoted $\overline{\varphi}$. For a fixed value of μ, a feasible test of the form $Q^{XB} = \sum_x |x\rangle\langle x|^X \otimes Q_x^B$ can be defined by setting Q_x to be the projector onto the region $\{y : \mu\varphi_x(y) - 3\overline{\varphi}(y) \geq 0\}$. The resulting regions for $\mu = 3$ are shaded blue, red, and orange, respectively. Note that Q_2 overlaps with Q_1 and Q_3.

to work with the *rate* of the compressor, $\log|C|$. The optimal rate is given by

$$R(\psi^{XB}, \epsilon, N) := \frac{\log \ell(\psi^{XB}, \epsilon, N)}{N}. \tag{7.43}$$

In the limit $N \to \infty$, the optimal rate tends to the *ϵ-compressibility* $C(\psi^{XB}, \epsilon)$. Using Stein's lemma, Proposition 6.2.2, we can show the following.

Proposition 7.4.3: Compressibility of classical information

For any CQ state ψ^{XB} and $\epsilon \in (0,1)$,

$$C(\psi^{XB}, \epsilon) = H(X|B)_\psi. \tag{7.44}$$

Whereas we might have expected that by allowing larger error we could achieve higher rates, this result states that there is no tradeoff between ϵ and the optimal rate. Any compression protocol operating at a rate below the conditional entropy will have essentially unit error. And for rates above the conditional entropy, we can always achieve a vanishing error rate. So $H(X|B)_\psi$ represents a sharp transition from essentially ideal to completely incorrect in the achievability of the protocol.

Proof. Start with the converse, (7.23). Choosing $\sigma = \psi_B^{\otimes N}$ gives, for any $\epsilon \in (0,1)$,

$$\lim_{N\to\infty} \frac{\log|C|}{N} \geq \lim_{N\to\infty} \frac{1}{N} \log \beta_{1-\epsilon}(\psi_{XB}^{\otimes N}, (\mathbb{1}_X \otimes \psi_B)^{\otimes N}) \tag{7.45}$$

$$= -D(\psi^{XB}, \mathbb{1}^X \otimes \psi^B) = H(X|B)_\psi. \tag{7.46}$$

The achievability, (7.33), directly gives

$$\lim_{N\to\infty}\frac{\log|C|}{N} \leq \lim_{N\to\infty}\frac{1}{N}\left(\log\frac{\eta^2}{4\epsilon}+\log\beta_{1-\epsilon+\eta}(\psi_{XB}^{\otimes N},(\mathbb{1}_X\otimes\psi_B)^{\otimes N})\right) \tag{7.47}$$

$$= -D(\psi^{XB},\mathbb{1}^X\otimes\psi^B) = H(X|B)_\psi. \tag{7.48}$$

□

7.5 Classical communication over noisy channels

The resource simulation task of noisy channel coding is depicted in Figure 7.1. Here we are interested in transmitting classical information, the messages $m\in\mathscr{M}$, not quantum information. The encoding operation $\mathscr{E}$ then translates any message m into a quantum state ρ_m^A which can be input to the quantum channel $\mathscr{N}$. The decoding operation $\mathscr{D}$ outputs a classical value m' from system B, so it can be modeled as a general measurement.

Often, the reason reliable communication over noisy channels is possible is attributed to the fact that the information to be transmitted is somehow redundantly encoded. This is an appropriate description when considering a message alphabet of fixed-size. For instance, a simple code for a noisy channel with binary inputs and outputs (i.e. taking bits to bits) is to just send the information three times and take the majority vote of the outcomes. However, from the channel point of view, since we are using the channel three times there were eight possible inputs, six of which are not used. Thus, from a different viewpoint the reason noisy channel coding works is by only using a subset of channel inputs which are reliably distinguishable at the output.

Ultimately, the compound channel $\mathscr{D}\circ\mathscr{N}\circ\mathscr{E}$ takes classical inputs to classical outputs. In this case the distinguishability can be shown to be just the worst-case error probability:

$$\delta(\mathscr{I},\mathscr{D}\circ\mathscr{N}\circ\mathscr{E}) = \max_{m\in\mathscr{M}}\left(1-\mathrm{Tr}[\Lambda_m^B\mathscr{N}\circ\mathscr{E}(|m\rangle\langle m|)]\right). \tag{7.49}$$

Here we have represented the messages m as basis states $|m\rangle$, which we can regard as part of the definition of $\mathscr{E}$.

However, a better approach is to split $\mathscr{E}$ into two parts, a *quantizer* $\mathscr{Q}$ which maps classical letters $x\in\mathscr{X}$ to A, and a *classical* encoding map $\mathscr{E}$ from M to X. The name quantizer comes from the classical setting, where symbols from a discrete alphabet are mapped to a continuous alphabet, as for instance in an electromagnetic channel which takes real-valued inputs. The quantizer makes the continuous channel into a discrete one, so perhaps "discretizer" would be a better name. But the name also works in our setting, since we can think of it as translating classical symbols into quantum states, "quantizing" them.

We can now simplify the problem by fixing the quantizer and combining it with the channel to create a *CQ channel* $W=\mathscr{N}\circ\mathscr{Q}$. This channel maps any input letter x to a quantum state of B, call it φ_x^B. Then, at the encoder end, we only need to look for classical maps $\mathscr{E}$ from M to X to build a reliable communication channel. The distinguishability now takes the form

$$\delta(\mathscr{I},\mathscr{D}\circ W\circ\mathscr{E}) = \max_{m\in\mathscr{M}}\left(1-\sum_{x\in\mathscr{X}}P_{X|M=m}(x)\mathrm{Tr}[\Lambda_m^B\varphi_x^B]\right), \tag{7.50}$$

where $P_{X|M}$ is the conditional distribution associated with the encoding map $\mathscr{E}$.

7.5.1 Converse to channel coding

As with data compression, we can find a constraint based on using any possible coding scheme to construct a hypothesis testing measurement for two particular states associated with the coding scheme. For any given distribution P_X of X, define $\psi^{XB} = \sum_{x\in\mathcal{X}} P_X(x)|x\rangle\langle x|^X \otimes \varphi_x^B$. Then we have the following converse bound for coding schemes with fixed error probability averaged over a uniformly random choice of the message. Since any encoder and decoder with worst-case error probability ϵ also have an average error no greater than ϵ, the converse applies to worst-case schemes as well.

Proposition 7.5.1: Converse for classical communication over noisy channels

Any coding scheme for the CQ channel W with average error probability ϵ obeys

$$|M| \leq \max_{P_X} \min_{\sigma} \frac{1}{\beta_{1-\epsilon}(\psi^{XB}, \psi^X \otimes \sigma^B)}, \tag{7.51}$$

where σ is any normalized state.

Note that none of the quantities involve the encoding and decoding operations, i.e. the protocol. They are purely properties of the real resource W used to simulate the ideal noiseless channel, as well as the approximation parameter ϵ.

Proof. The idea is to construct a hypothesis test between ψ^{XB} and $\psi^X \otimes \sigma^B$, for a certain P_X and any σ^B, using the encoder and decoder of the coding scheme.

Suppose that the messages are chosen at random. We can describe the state of the message M, the input to the channel X, and the output B by the following tripartite state

$$\psi^{MXB} = \tfrac{1}{|M|} \sum_{m\in\mathcal{M}} \sum_{x\in\mathcal{X}} P_{X|M=m}(x)|m\rangle\langle m|^M \otimes x^X \otimes \varphi_x^B, \tag{7.52}$$

where the conditional distribution $P_{X|M}$ describes the action of the encoding map $\mathcal{E}$. Tracing out M gives the state

$$\psi^{XB} = \sum_{x\in\mathcal{X}} P_X(x)|x\rangle\langle x|^X \otimes \varphi_x^B, \tag{7.53}$$

where P_X is the marginal of the joint distribution $P_{MX}(m,x) = \frac{1}{|M|} P_{X|M=m}(x)$.

Now consider task of distinguishing between ψ^{XB} (the null hypothesis) and $\psi^X \otimes \sigma^B$ (the alternate hypothesis) for any state σ^B. We can use the encoder and decoder to design a test Q^{XB} with type-I error less than ϵ. In particular, define

$$Q^{XB} = \sum_{m\in\mathcal{M}} \sum_{x\in\mathcal{X}} P_{M|X=x}(m)|x\rangle\langle x|^X \otimes \Lambda_m^B, \tag{7.54}$$

where the conditional distribution $P_{M|X}$ is formed from the conditional probability describing the encoder and the uniform input distribution P_M via the usual rules of conditional probability. Note that $Q^{XB} \geq 0$, since both $P_{M|X=x}(m)$ and Λ_m are. Moreover, $Q^{XB} \leq \mathbb{1}^{XB}$ since $\Lambda_m \leq \mathbb{1}$ and $P_{M|X=x}(m) \leq 1$. Therefore, Q^{XB} is a valid POVM element. It detects ψ^{XB} with type-I error less

than ϵ, for

$$\begin{aligned}
\mathrm{Tr}[Q^{XB}\psi^{XB}] &= \sum_{m\in\mathscr{M}}\sum_{x\in\mathscr{X}} P_{M|X=x}(m)P_X(x)\mathrm{Tr}[\Lambda_m\varphi_x] && (7.55)\\
&= \sum_{m\in\mathscr{M}}\sum_{x\in\mathscr{X}} P_{MX}(m,x)\mathrm{Tr}[\Lambda_m\varphi_x] && (7.56)\\
&= \tfrac{1}{|M|}\sum_{m\in\mathscr{M}}\sum_{x\in\mathscr{X}} P_{X|M=m}(x)\mathrm{Tr}[\Lambda_m\varphi_x] && (7.57)\\
&\geq 1-\epsilon. && (7.58)
\end{aligned}$$

In the last step we have used the assumption that the coding scheme has an average error probability less than ϵ. The minimal type-II error must therefore satisfy, for any σ,

$$\begin{aligned}
\beta_{1-\epsilon}(\psi^{XB},\psi^X\otimes\sigma^B) &\leq \mathrm{Tr}[Q^{XB}\psi^X\otimes\sigma^B] && (7.59)\\
&= \sum_{m\in\mathscr{M}}\sum_{x\in\mathscr{X}} P_{MX}(m,x)\mathrm{Tr}[\Lambda_m\sigma] && (7.60)\\
&= \frac{1}{|M|}\sum_{m\in\mathscr{M}}\mathrm{Tr}[\Lambda_m\sigma] && (7.61)\\
&= \frac{1}{|M|}. && (7.62)
\end{aligned}$$

Taking the σ which maximizes the lefthand side gives

$$|M| \leq \min_{\sigma}\frac{1}{\beta_{1-\epsilon}(\psi^{XB},\psi^X\otimes\sigma^B)}. \tag{7.63}$$

Finally, weakening the bound by taking the maximum over P_X gives the desired statement. □

7.5.2 Achievability of channel coding

The achievability argument for channel coding is very similar to that of data compression. Again we can choose the encoding map to be deterministic without loss of generality and we use the random coding argument. And again we construct the decoder map from the pretty good measurement based on the optimal test in a certain hypothesis testing scenario.

Although we are interested in the worst-case error performance of a coding scheme, achievability is simpler to show when considering the error probability averaged over a uniformly random choice of the message (this is distinct from the average over encoding functions). For the simulation argument to carry through, we need to design an encoding and decoding scheme with low worst-case error, so by itself the average-case analysis is insufficient. However, a slight change to the encoder is sufficient to transform a scheme with low average error to one with low worst-case error. Let us first see how this is done.

Suppose that $\mathscr{E}$ and $\mathscr{D}$ are an encoder and decoder pair which achieve an average error probability of ϵ. That is, for each message, the error is given by

$$p_{\mathrm{error}}(m) = \mathrm{Tr}[(\mathbb{1}-\Lambda_m)\varphi_{f(m)}], \tag{7.64}$$

where $f : M \to X$ is the encoding function. Averaging over m gives, by assumption,

$$p_{\mathrm{error}} \coloneqq \tfrac{1}{|M|}\sum_m p_{\mathrm{error}}(m) \leq \epsilon \tag{7.65}$$

Now throw away the worst half of the messages; the result will be a code with worst-case error less than 2ϵ. More specifically, split the message set into two equally-sized subsets, where the elements of the first set have lower error probability than those of the second (i.e. we divide the set ordered by error probability at its median). Suppose the average error for the first (second) set is ϵ_1 (ϵ_2), and observe that $\epsilon_1 + \epsilon_2 = 2\epsilon$. But the largest error from the first set must be less than the average error of the second set, ϵ_2, which in turn is less than twice the original average error ϵ. Modifying the encoder to only accept input from the low-error set leads to a coding scheme with worst-case error at most 2ϵ.

Proposition 7.5.2: Achievability for classical communication over noisy channels

For a given CQ channel $W_{X\to B}$, any desired average error ϵ, and any $\eta \leq \epsilon$, there exists a deterministic encoder $\mathcal{E}_{M\to X}$ and decoder $\mathcal{D}_{B\to M}$ with

$$|M| = \left\lfloor \max_{P_X} \frac{\eta^2}{4\epsilon} \frac{1}{\beta_{1-\epsilon+\eta}(\psi^{XB}, \psi^X \otimes \psi^B)} \right\rfloor. \tag{7.66}$$

Proof. The proof proceeds in three steps. First, we analyze the error probability when using a fixed deterministic encoder and a decoder constructed from the pretty good measurement for an arbitrary set of operators. Next, we simplify the expression by averaging over the choice of encoding function. Finally, we make a specific choice of the operators forming the decoder to make the connection to the hypothesis testing type-II error.

For a given distribution P_X, define the state $\psi^{XB} = \sum_x P_X(x)|x\rangle\langle x|^X \otimes \varphi_x^B$. Then consider an arbitrary positive CQ operator $Q^{XB} = \sum_x |x\rangle\langle x|^X \otimes Q_x^B$. Given an encoding function $f : M \to X$, we define the decoder to be the POVM with elements

$$\Lambda_{m;f} = \hat{Q}_f^{-1/2} Q_{f(m)} \hat{Q}_f^{-1/2}, \tag{7.67}$$

where $\hat{Q}_f = \sum_m Q_{f(m)}$. The average error probability using this encoder and decoder is then

$$p_{\text{error}} = \frac{1}{|M|} \sum_m \text{Tr}[(\mathbb{1} - \Lambda_{m;f})\varphi_{f(m)}]. \tag{7.68}$$

For each term in the sum we can apply Lemma A.10.1 with $S = Q_{f(m)}$ and $T = \hat{Q}_f - Q_{f(m)} = \sum_{m'\neq m} Q_{f(m')}$ to bound the error probability from above. For any $a > 0$ we have

$$p_{\text{error}} \leq \frac{1}{|M|} \sum_m (1+a)\text{Tr}[(\mathbb{1} - Q_{f(m)})\varphi_{f(m)}] - \frac{1}{|M|} \sum_m (2 + a + a^{-1}) \sum_{m'\neq m} \text{Tr}[Q_{f(m')}\varphi_{f(m)}]. \tag{7.69}$$

Next we average over the choice of f. Consider the stochastic map which maps each m to an $x \in \mathcal{X}$ with probability P_X. This induces a distribution on deterministic functions, again by Proposition 2.5.2. Averaging over the choice of deterministic function is the same as using the stochastic map in the expression for the error probability. Denoting this average by angle brackets, we have

$$\langle p_{\text{error}} \rangle \leq \frac{1}{|M|} \sum_{m,x} P_X(x) \left((1+a)\text{Tr}[(\mathbb{1} - Q_x)\varphi_x] + (2 + a + a^{-1}) \sum_{m'\neq m, x'} P_X(x')\text{Tr}[Q_{x'}\varphi_x] \right) \tag{7.70}$$

$$= \frac{1}{|M|} \sum_m \left((1+a)(1 - \text{Tr}[Q^{XB}\psi^{XB}]) + (2 + a + a^{-1}) \sum_{m'\neq m} \text{Tr}[Q^{XB}(\psi^X \otimes \psi^B)] \right) \tag{7.71}$$

$$\leq (1+a)(1 - \text{Tr}[Q^{XB}\psi^{XB}]) + (2 + a + a^{-1})|M|\text{Tr}[Q^{XB}(\psi^X \otimes \psi^B)]. \tag{7.72}$$

Here ψ^X and ψ^B are the marginal states of ψ^{XB}.

Finally, we make a specific choice for Q^{XB}: the optimal test in $\beta_{1-\xi}(\psi^{XB}, \psi^X \otimes \psi^B)$. Then in the first term we can substitute $\mathrm{Tr}[Q^{XB}\psi^{XB}] = 1-\xi$ and in the second $\mathrm{Tr}[Q^{XB}(\psi^X \otimes \psi^B)] = \beta_{1-\xi}(\psi^{XB}, \psi^X \otimes \psi^B)$. Altogether, we can infer the existence of at least one encoding function f such that

$$p_{\text{error}} \leq (1+a)\xi + (2+a+a^{-1})|M|\beta_{1-\xi}(\psi^{XB}, \psi^X \otimes \psi^B). \tag{7.73}$$

We can now choose a, ξ, and $|M|$ in such a way that the righthand side equals ϵ. A nice choice is $\xi = \epsilon - \eta$, $a = \eta/(2\epsilon - \eta)$ and

$$|M| = \frac{\eta^2}{4\epsilon} \frac{1}{\beta_{1-\epsilon+\eta}(\psi^{XB}, \psi^X \otimes \psi^B)}. \tag{7.74}$$

Maximizing over the distribution P_X gives the desired statement. □

7.5.3 Coding for i.i.d. channels

When the channel in question is actually N independent instances of a fixed channel W, i.e. the channel $W^{\otimes N}$, then we can show that the converse and achievability bounds meet in the limit $N \to \infty$. In this setting it is more convenient to work with the logarithm of the number of messages and even more the *rate* of the coding scheme. Let $m(W, \epsilon, N)$ be the largest $|M|$ such that there exists an ϵ-good coding scheme for $W^{\otimes N}$. Then the optimal rate $R(W, \epsilon, N)$ is defined as

$$R(W, \epsilon, N) \coloneqq \frac{\log m(W, \epsilon, N)}{N}. \tag{7.75}$$

In the limit of $N \to \infty$, the optimal rate tends to the *ϵ-capacity* $C(W, \epsilon)$. Often we are interested in the limit $\epsilon \to 0$ of the ϵ-capacity, which is just called the capacity $C(W)$.

Using Stein's lemma in the achievability and converse, we can show the following.

Proposition 7.5.3: Capacity of CQ channels

For any CQ channel W and corresponding state $\psi^{XB} = \sum_x P_X(x)|x\rangle\langle x|^X \otimes \varphi_x^B$,

$$C(W) = \max_{P_X} I(X:B)_{\psi^{XB}}. \tag{7.76}$$

Proof. In the achievability statement choose $P_{X^N} = P_X^{\times N}$ to get

$$\log|M| \geq \log\frac{\eta^2}{4\epsilon} - \log\beta_{1-\epsilon+\eta}(\psi_{XB}^{\otimes N}, (\psi_X \otimes \psi_B)^{\otimes N}), \tag{7.77}$$

where we have moved system labels from superscript to subscript to simplify the notation. Dividing by N and taking the limit using Stein's lemma, Proposition 6.2.2, gives

$$\lim_{N\to\infty} R(W, \epsilon, N) \geq \lim_{N\to\infty} \frac{1}{N}\log\frac{\eta^2}{4\epsilon} + \max_{P_X}\left(-\frac{1}{N}\log\beta_{1-\epsilon+\eta}(\psi_{XB}^{\otimes N}, (\psi_X \otimes \psi_B)^{\otimes N})\right) \tag{7.78}$$

$$= \max_{P_X} I(X:B)_{\psi^{XB}} \tag{7.79}$$

for all $\eta < \epsilon$. Thus, the capacity or indeed ϵ-capacity for any $\epsilon \in (0,1)$ is at least $\max_{P_X} I(X:B)_{\psi^{XB}}$.

To show that the capacity cannot be larger is more difficult, because we must show that input distributions of the form $P_{X^N} = P_X^{\times N}$ are optimal. Properties of the mutual information make this possible. First, we may use Equation 6.34 in the converse (7.51) to obtain

$$\log|M| \leq \max_{P_{X^N}} \min_{\sigma} \frac{1}{1-\epsilon} D(\psi^{X^N B^N}, \psi^{X^N} \otimes \sigma^{B^N}) + h(\epsilon) \tag{7.80}$$

$$= \max_{P_{X^N}} \frac{I(X^N:B^N)_{\psi^{X^N B^N}} + h(\epsilon)}{1-\epsilon}. \tag{7.81}$$

In the second step we have loosened the bound by setting $\sigma^{B^N} = \psi^{B^N}$, though in fact this choice is the minimum. Dividing by N and taking the limit $N \to \infty$ removes the second term, so we need only focus on the first. By the chain rule we have

$$I(X^N:B^N)_\psi = H(B^N)_\psi - H(B^N|X^N)_\psi = H(B^N)_\psi - \sum_{j=1}^N H(B_j|B_1,\ldots,B_{j-1}X^N)_\psi. \tag{7.82}$$

Since each channel use in independent of all others, the state output by the jth channel in system B_j depends only on the input X_j to the jth channel. Therefore, $H(B_j|B_1,\ldots,B_{j-1}X^N) = H(B_j|X_j)$. Using subadditivity for the first term, we then obtain

$$I(X^N:B^N)_\psi = H(B^N)_\psi - \sum_{j=1}^N H(B_j|X_j)_\psi \tag{7.83}$$

$$\leq \sum_{j=1}^N H(B_j)_\psi - H(B_j|X_j)_\psi = \sum_{j=1}^N I(X_j:B_j)_\psi. \tag{7.84}$$

Using this in the expression for the ϵ capacity gives

$$C(W,\epsilon) = \lim_{N\to\infty} \frac{\log|M|}{N} \tag{7.85}$$

$$\leq \frac{1}{1-\epsilon} \lim_{N\to\infty} \frac{1}{N} \max_{P_{X^N}} \sum_{j=1}^N I(X_j:B_j)_\psi \tag{7.86}$$

$$= \frac{1}{1-\epsilon} \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^N \max_{P_X} I(X_j:B_j)_\psi \tag{7.87}$$

$$= \frac{1}{1-\epsilon} \max_{P_X} I(X:B)_{\psi^{XB}}. \tag{7.88}$$

The bound blows up for $\epsilon \to 1$, but for $\epsilon \to 0$ we recover the desired result. □

Here we have established the *weak converse*, which governs the behavior of coding schemes with $\epsilon \to 0$ as N grows. However, the maximum achievable rate is in fact independent of ϵ, a statement known as the *strong converse*:

$$C(W,\epsilon) = \max_{P_X} I(X:B)_{\psi^{XB}} \qquad \forall \epsilon \in (0,1). \tag{7.89}$$

The strong converse can in fact be obtained from the converse bound in (7.51), but the derviation is much lengthier than that of the weak converse.

7.6 Compression of quantum data

Quantum data compression is the quantum version of classical data compression. In the eigenbasis of the source output, its eigenvalues form a classical probability distribution, to which we can apply the classical data compression scheme of §7.4. Returning to Figure 7.3, this means that we set define the copy operation to be the quantum operation which copies the eigenbasis of the output of the source. An ϵ-good classical data compression scheme will therefore become an ϵ-good quantum data compression scheme, where the approximation parameter again refers to the distinguishability metric.

Recycling the classical protocol more or less ignores all the quantum aspects to the problem. In particular, we might also want the compression scheme to preserve the entanglement that the source output has with its purification, as depicted in Figure 7.6. However, a good approximation including

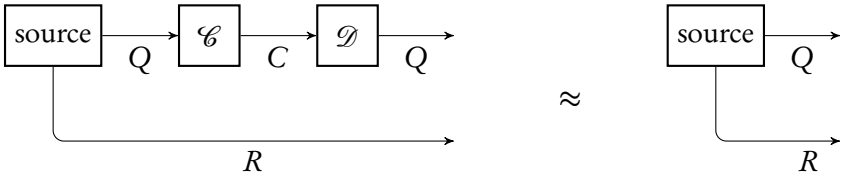


Figure 7.6: Compression of quantum data. In contrast to the classical case, here we also want to retain correlations or entanglement of the system Q with its purification R.

the purification can only be guaranteed when the source outputs an essentially pure state, as the following argument shows.

When including the purification into the distinguishability, it is equivalent to work with the fidelity, defined in (5.25). Suppose that the source outputs the state $\rho^Q = \sum_x P_X(x)|b_x\rangle\langle b_x|^Q$, with purification $|\psi\rangle^{QR} = \sum_x \sqrt{P_X(x)}|b_x\rangle^Q|\xi_x\rangle^R$ for some orthonormal states $|\xi_x\rangle$. Applying a perfect classical compression scheme in the basis $|b_x\rangle$ would lead to the state

$$|\psi'\rangle^{QQ'R} = \sum_x \sqrt{P_X(x)}|b_x\rangle^Q|b_x\rangle^{Q'}|\xi_x\rangle^R, \tag{7.90}$$

where Q and Q' are the X and X' of Figure 7.3. But the QR subsystem (which is now a CQ state) does not have high fidelity with $|\psi\rangle^{QR}$:

$$\langle\psi|\psi'^{QR}|\psi\rangle^{QR} = \sum_x P_X(x)^2. \tag{7.91}$$

This is only close to unity if the distribution P_X has essentially all weight on one particular value of x, i.e. if the original state ρ^Q is nearly a pure state.

The difficulty is the "copy" Q' of Q that is created, which we may think of as leftover at the compressor. To have a good approximation in the sense of Figure 7.6, we must remove all traces of the input state at the compressor. Let us be a little more clever about implementing the compressor. By a slight change to the classical compression protocol of §7.4, we can show

Proposition 7.6.1: Achievability of quantum data compression

For any source emitting a state ρ^Q and parameters $\epsilon \geq 0$ and $0 < \eta \leq \epsilon$, there exists compression and decompression maps $\mathscr{C}_{Q\to C}$ and $\mathscr{D}_{C\to Q}$ such that

$$F(\mathscr{D}\circ\mathscr{C}(\rho^Q),\rho^Q) \geq 1-\epsilon, \qquad \text{and} \tag{7.92}$$

$$|C| = \left\lceil \frac{4\epsilon}{\eta^2}\beta_{1-\epsilon+\eta}(\rho^Q, \mathbb{1}^Q)\right\rceil. \tag{7.93}$$

Proof. First consider the problem of classical data compression absent any side information B. For a given deterministic compression map f the optimal decoder is also deterministic and simply picks the most likely x which is compatible with the compressor output c, namely $\operatorname{argmax}_x P_{X|C=c}(x)$. Moreover, Proposition 7.4.2, shows that a suboptimal decoder can achieve error probability ϵ with a compressor output of size given by (7.93). Therefore, the optimal decoder will only have a lower probability of error.

The function f is deterministic, but not necessarily invertible (on its image). But we can extend it to an invertible function in the following way. 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 $\mathscr{X}_{f(x)}^{\downarrow}$, counting from 0. The map $x \to (f(x), g(x))$ is reversible, and $g(x)$ (or something like it) is the smallest additional output that makes this possible. For future use call the inverse map h:

$$x \xleftarrow{h} (f(x), g(x)). \tag{7.94}$$

As a quantum operation we can define the compressor to perform the unitary

$$U_{Q\to CT} = \sum_x |f(x)\rangle_C |g(x)\rangle_T \langle x|_Q, \tag{7.95}$$

followed by tracing out system T. The decompressor implements the isometry

$$V_{C\to Q} = \sum_y |h(y,0)\rangle_Q \langle y|_C, \tag{7.96}$$

which maps y to the most likely choice of X among those with $f(x) = y$, since it assumes $g(x) = 0$.

Now let $|\psi'\rangle^{QTR} = V_{C\to Q}U_{Q\to CT}|\psi\rangle^{QR}$, the output of the actual protocol. More explicitly, the state has the form

$$|\psi'\rangle^{QTR} = \sum_x \sqrt{P_X(x)}|h(f(x),0)\rangle^Q |g(x)\rangle^T |\xi_x\rangle^R. \tag{7.97}$$

The (squared) fidelity $F(\mathscr{D}\circ\mathscr{C}(\rho^Q),\rho^Q)^2$ with the ideal output is

$$F(\mathscr{D}\circ\mathscr{C}(\rho^Q),\rho^Q)^2 = \langle\psi|\psi'^{QR}|\psi\rangle^{QR} \tag{7.98}$$

$$= \operatorname{Tr}_T[\langle\psi|\psi'^{QTR}|\psi\rangle^{QR}], \tag{7.99}$$

where the second equation holds because the partial trace over T commutes with the expectation in QR. To evaluate this expression we need only compute the vector ${}^{QR}\langle\psi|\psi'\rangle^{QTR}$. We have

$$
\begin{aligned}
{}^{QR}\langle\psi|\psi'\rangle^{QTR} &= \sum_x P_X(x)\langle x|h(f(x),0)\rangle|g(x)\rangle^T && (7.100)\\
&= \sum_x P_X(x)\delta_{g(x),0}|g(x)\rangle^T && (7.101)\\
&= \sum_{x:g(x)=0} P_X(x)|0\rangle^T && (7.102)\\
&= (1-\epsilon)|0\rangle^T. && (7.103)
\end{aligned}
$$

In the last line we have used the fact that the x with $g(x)=0$ are correctly reconstructed at the output by the decompressor, and hence $\sum_{x:g(x)=0} P_X(x) = 1-\epsilon$. The trace over T is now immediate, and leads to $F(\mathscr{D}\circ\mathscr{C}(\rho^Q),\rho^Q) = 1-\epsilon$. □

Since the fidelity of the output with the ideal state is nearly one, the actual state produced by the protocol is essentially $|\psi\rangle^{QR}W_T|0\rangle^T$ for some isometry W on T. Because the compression map is an isometry from Q to CT, and T ends up in a pure state, we can regard the compressor as storing the important information about Q in C (since the decompressor reconstructs the state) and *erasing* the rest of Q, creating pure states in T. As we saw above, it must erase the unneeded part of Q in some manner. Put differently, we can regard quantum data compression not as the task of putting all the important "information" about Q into the smallest system C possible, but rather as erasing as much of Q as possible. When the erasure transformation is a reversible operation, as here, whatever part cannot be erased must contain all the useful information about the input.

Note that the two compression maps are different; $\mathrm{tr}_T[U(\cdot)U^*]$ is not the same *quantum* 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

$$\mathrm{tr}_T[U|x\rangle\langle x'|U^*] = |f(x)\rangle\langle f(x')|\delta(g(x),g(x')), \tag{7.104}$$

while the latter gives

$$\mathscr{C}(|x\rangle\langle x'|) = |f(x)\rangle\langle f(x')|\delta(x,x'). \tag{7.105}$$

Finally, observe that since a protocol for quantum compression as in Figure 7.6 can also be used to compress classical data (by copying C and running the decompressor on both copies), the quantum task inherits the converse from the classical task. As this is essentially tight, there is no (substantially) stronger converse for the quantum case.

7.7 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

$$|\Psi\rangle_{AB} = \sum_x \sqrt{P_X(x)}|\varphi_x\rangle_A \otimes |\xi_x\rangle_B \tag{7.106}$$

(expressed here in the Schmidt basis) into an approximate version of $|\Phi_M\rangle_{A'B'} = \frac{1}{\sqrt{M}}\sum_{y=0}^{M-1}|y\rangle_{A'}\otimes|y\rangle_{B'}$, for the largest M possible, using only local operations and classical communication.

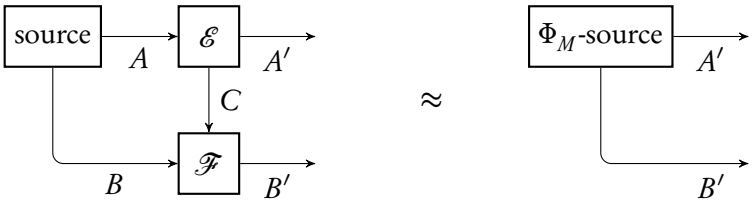


Figure 7.7: Entanglement purification. System C is classical.

As we saw in §7.6, data compression can be seen as a means of producing pure states from a given source, that is, an output (T) which has entropy zero. In entanglement purification, on the other hand, the goal is to make the marginal 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, depicted in Figure 7.4(b).

Suppose that f is a map fulfilling the requirements for randomness extraction from the distribution P_X (the eigenvalues of Ψ_{AB}). Furthermore, let $V_{A\to AA'}$ be the isometry which implements f in the basis $|\varphi_x\rangle$, that is

$$V_{A\to AA'}|\varphi_x\rangle_A = |\varphi_x\rangle_A|f(x)\rangle_{A'}. \tag{7.107}$$

Applied to $|\Psi\rangle_{AB}$ gives

$$|\Psi'\rangle_{AA'B} = V_{A\to AA'}|\Psi\rangle_{AB} \tag{7.108}$$

$$= \sum_x \sqrt{P_X(x)}|\varphi_x\rangle_A|f(x)\rangle_{A'}|\xi_x\rangle_B. \tag{7.109}$$

The marginal state $\Psi'_{A'}$ is essentially the mixed state, since f extracts randomness from X. In particular, $\Psi'_{A'} = \sum_x P_X(x)|f(x)\rangle\langle f(x)|_{A'}$, so the probability of any particular $|y\rangle_{A'}$ is just $\sum_{x:f(x)=y} P_X(x)$. This is precisely the probability of the yth output of f applied to X, and therefore,

$$\delta(\Psi'_{A'}, \tfrac{1}{M}\mathbb{1}_{A'}) \leq \epsilon. \tag{7.110}$$

By (5.40) this implies $F(\Psi'_{A'}, \frac{1}{M}\mathbb{1}_{A'}) \geq 1-\varepsilon$. The fidelity is an optimization over purifications, and one possible purification of $\Psi'_{A'}$ is simply $|\Psi'\rangle_{A'AB}$ itself. A possible purification of $\frac{1}{M}\mathbb{1}_{A'}$ is $|\Phi_M\rangle_{A'B'}$. Then, by the definition of fidelity, there must exist an isometry $W_{AB\to B'}$ such that

$${}_{A'B'}\langle\Phi_M|W_{AB\to B'}|\Psi'\rangle_{A'AB} \geq 1-\varepsilon. \tag{7.111}$$

Thus, knowing that the A' system is maximally mixed allows us to infer the existence of an operation which creates the desired state $|\Phi_M\rangle_{A'B'}$. This trick of inferring that entangled states can be created by showing that the marginal is completely mixed is quite widespread in quantum information theory.

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, we simply erase system A with the quantum eraser discussed after (3.34). To "erase" A, first define a conjugate basis to $|\varphi_x\rangle$ by

$$|\vartheta_z\rangle = \tfrac{1}{\sqrt{d}}\sum_{x=0}^{d-1}\omega^{xz}|\varphi_x\rangle, \tag{7.112}$$

where $\omega = e^{2\pi i/d}$ and d is the dimension of the system A. Measuring A in this basis gives the state

$$ {}_A\langle\vartheta_z|\Psi'\rangle_{A'AB} = \tfrac{1}{\sqrt{d}} \sum_x \sqrt{P_X(x)}\,\omega^{-xz} |f(x)\rangle_{A'} |\xi_x\rangle_B, \tag{7.113} $$

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

$$ R^z = \sum_x \omega^{xz} |\xi_x\rangle\langle\xi_x| \tag{7.114} $$

and this will produce the state (now normalized)

$$ |\Psi''\rangle_{A'B} = \sum_x \sqrt{P_X(x)} |f(x)\rangle_{A'} |\xi_x\rangle_B. \tag{7.115} $$

With A now out of the picture, applying the above decoupling argument to this state gives an LOCC entanglement purification protocol. In particular, $\mathscr{E}$ consists of the map $V_{A\to AA'}$ followed by measurement of A in the basis $|\vartheta_z\rangle$. The resulting outcome z is transmitted in the classical system C. Finally, $\mathscr{D}$ is just the operation $W_{B\to B'}$. 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.

7.8 Exercises

Exercise 7.1. Optimality of superdense coding → solution

Show that $a+b \geq 2$ in order for (7.4) to hold.

Exercise 7.2. Classical resource inequalities → solution

Consider a communication system of two partners Alice A and Bob B and an eavesdropper Eve E. The classical analog of an entangled bit is a secret bit shared between A and B, modeled by a probability distrubution P_{ABE}, such that

$$P_{ABE} = P_{AB} \cdot P_E, \qquad P_{AB}[A=i, B=j] = (Q)_{ij}, \qquad Q = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Furthermore, suppose that the classical communication between A and B is insecure in that everything broadcasted over the channel will be heard by E. Prove the following lemma:

Lemma 7.8.1. *Given A and B share l secure bits and unlimited classical communication, they cannot create more than l secure bits.*

a) Calculate the mutual information $I(A:B|E) = H(A|E) - H(A|B,E)$ when A and B share l secure bits.

b) Explain why the lemma follows after we show that the mutual information $I(A:B|E)$ is non-increasing under local operations and classical communication (LOCC).

c) Show that creating local randomness cannot increase mutual information:

$$I(A,X:B|E) \leq I(A:B|E).$$

d) Show that deterministic local operations $A \mapsto f(A)$ cannot increase mutual information:

$$I(f(A):B|E) \leq I(A:B|E).$$

e) Show that classical communication cannot increase conditional mutual information:

$$I(A,A':B,A'|E,A') \leq I(AA':B|E).$$

Exercise 7.3. Classical channel capacities → solution

Compute the classical capacity of the binary symmetric channel (BSC) and the binary erasure channel (BEC). These channels are defined by

$$W_{\text{BSC}}(y|x) = \begin{cases} 1-p & y=x \\ p & y \neq x \end{cases} \qquad \text{and} \qquad W_{\text{BEC}}(y|x) = \begin{cases} 1-p & y=x \\ p & y=? \\ 0 & y \neq x \end{cases}$$

for $x \in \{0,1\}$ and $y \in \{0,1\}$ or $y \in \{0,?,1\}$, respectively.

Exercise 7.4. Classical capacity of the depolarizing channel → solution

Recall that the qubit depolarizing channel is defined by the action $\mathcal{E}(\rho) = (1-p)\rho + p\frac{1}{2}\mathbb{1}$. Now we will see what happens when we use this quantum channel to send classical information. Starting from an arbitrary input probability distribution $P_X(0) = q$, $P_X(1) = 1-q$, encode X in the state $\rho = |0\rangle\langle 0| + (1-q)|1\rangle\langle 1|$. Suppose ρ is transmitted over the quantum channel and measured in the $|0\rangle, |1\rangle$ basis at the output, yielding the random variable Y.

a) Compute the conditional probability distributions $P_{Y|X=x}(y)$.

b) Maximize the mutual information over q to find the classical channel capacity of the depolarizing channel with this particular encoding and decoding.

c) What happens to the channel capacity if we measure the final state in a different basis?

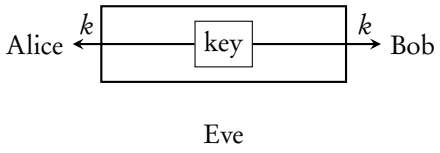
8

Quantum Key Distribution

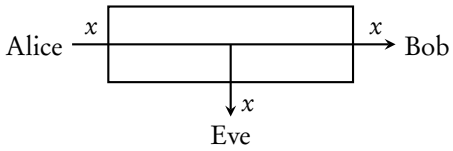
8.1 Introduction

In this chapter, we introduce the concept of quantum key distribution. Traditionally, cryptography is concerned with the problem of securely sending a secret message from A to B. Note however that secure message transmission is only one branch of cryptography. Another example for a problem studied in cryptography is coin tossing. There the problem is that two parties, Alice and Bob, which are physically separated and do not trust each other want to toss a coin over the telephone. Blum showed that this problem cannot be solved as long as one does not introduce additional asssumptions [79]. Note that coin tossing is possible using quantum communication.

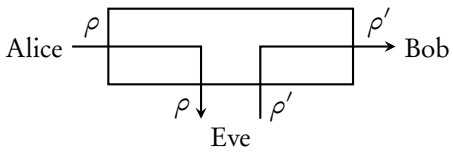
A central element in modeling security is that of *resources* – resources used in protocols and resources constructed by protocols. For example, a quantum key distribution (QKD) protocol constructs a functionality $\mathscr{K}_n$, which shares a secret key k of length n between two players:



This functionality is a resource, which can be used by other protocols, e.g., to encrypt a message. To construct this secret key resource, a QKD protocol typically uses two other resources, a multiple use authentic classical channel $\mathscr{A}_\infty$ – which guarantees that the message x received comes from the legitimate sender, but allows the adversary to get a copy:



and a multiple use insecure quantum channel $\mathscr{Q}_\infty$, completely under the control of the adversary:



We write this transformation of resources as

$$\{\mathscr{A}_\infty, \mathscr{Q}_\infty\} \xrightarrow{\text{QKD}} \mathscr{K}_n. \tag{8.1}$$

More generally, if a protocol π constructs $\mathscr{S}$ from $\mathscr{R}$,

$$\mathscr{R} \xrightarrow{\pi} \mathscr{S},$$

then for any resource $\mathscr{R}'$,

$$\{\mathscr{R}, \mathscr{R}'\} \xrightarrow{\pi} \{\mathscr{S}, \mathscr{R}'\}.$$

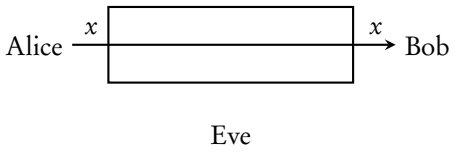
And if additionally π' constructs $\mathscr{T}$ from $\mathscr{S}$,

$$\mathscr{S} \xrightarrow{\pi'} \mathscr{T},$$

then the composition of the two protocols $\pi' \circ \pi$ constructs $\mathscr{T}$ from $\mathscr{R}$,

$$\mathscr{R} \xrightarrow{\pi' \circ \pi} \mathscr{T}.$$

In the following section we will introduce a classical encryption protocol, the *one-time pad* (OTP), which uses an n-bit secret key resource $\mathscr{K}_n$ and an authentic channel $\mathscr{A}_n$ to construct a secure channel resource for an n-bit message $\mathscr{S}_n$, i.e., a channel which does not allow an adversary to read the contents of the message x:



We thus have

$$\{\mathscr{A}_n, \mathscr{K}_n\} \xrightarrow{\text{OTP}} \mathscr{S}_n. \tag{8.2}$$

Composing these two protocols we can construct a secure channel from an insecure quantum channel and classical authentic channel. The corresponding resource transformation is obtained by combining (8.1) and (8.2):

$$\{\mathscr{A}_\infty, \mathscr{Q}_\infty\} \xrightarrow{\text{OTP}\circ\text{QKD}} \mathscr{S}_n.$$

In classical cryptography there exists a protocol [80], called *authentication protocol*, that constructs an authentic channel for an m-bit message, $\mathscr{A}_m$, from a secret key $\mathscr{K}_k$ for $m \gg k$:

$$\{\mathscr{K}_k, \mathscr{C}_\infty\} \xrightarrow{\text{AUTH}} \mathscr{A}_m, \tag{8.3}$$

where $\mathscr{C}_\infty$ is a classical insecure channel. Since the number of bits that need to be authenticated in a QKD protocol is typically of the order of the length of the key produced, combining authentication and QKD results in a key expansion protocol:

$$\{\mathscr{K}_k, \mathscr{Q}_\infty\} \xrightarrow{\text{QKD}\circ\text{AUTH}} \mathscr{K}_n,$$

for $n \gg k$.

8.2 Classical message encryption

The *one-time pad* protocol works as follows. Let m be an n-bit message and k an n-bit secret key. The operation $\oplus$ denotes the bitwise addition modulo 2. Alice first computes $c = m \oplus k$ and sends c over a classical authentic channel to Bob. Bob then computes $m' = c \oplus k$. This is illustrated in Figure 8.1.

The protocol is correct as

$$m' = c \oplus k = (m \oplus k) \oplus k = m \oplus (k \oplus k) = m.$$

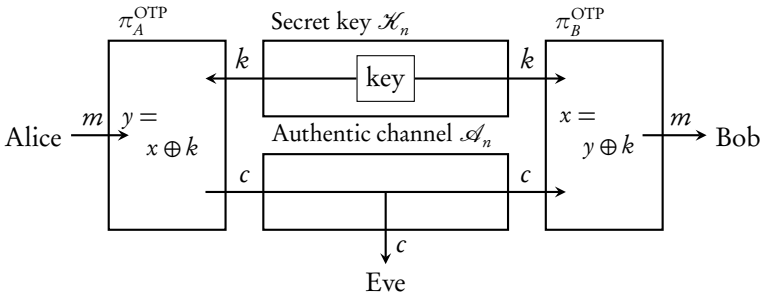


Figure 8.1: The real one-time pad system – Alice has access to the left interface, Bob to the right interface and Eve to the lower interface – consists of the one-time pad protocol $(\pi_A^{\text{OTP}}, \pi_B^{\text{OTP}})$, and the secret key and authentic channel resources. The combination of these resources and protocol constructs a new resource that takes a message m at Alice's interface, outputs a ciphertext c at Eve's interface and the original message m at Bob's interface.

To prove that the protocol does indeed construct a secure channel $\mathscr{S}_n$. We will argue that one cannot distinguisher between a system running the real one-time pad and a system using an ideal secure channel resource $\mathscr{S}_n$.

Let $\mathscr{K}_n||\mathscr{A}_n$ denote the parallel compositions of the two resources used by the OTP, which is depicted in Figure 8.1 — this is a new resource providing each player with an interface that is the composition of both interfaces of the individual resources, and let $\pi_A^{\text{OTP}}\pi_B^{\text{OTP}}(\mathscr{K}_n||\mathscr{A}_n)$ denote the resource constructed after running the OTP — the indexes A and B denote the interfaces of $\mathscr{K}_n||\mathscr{A}_n$ on which the systems π_A^{OTP} and π_B^{OTP} are plugged.

The argument involves as a thought experiment a simulator system σ_E^{OTP} which transforms the ideal adversarial interface of the resource $\mathscr{S}_n$ into the real interface provided by the system $\pi_A^{\text{OTP}}\pi_B^{\text{OTP}}(\mathscr{K}_n||\mathscr{A}_n)$: anything that can be done by a dishonest player Eve accessing the E interface of the real system may also be achieved in the ideal world by first running the simulator σ_E^{OTP}. In the case of the OTP, this simulator simply outputs a random string of length n, which, by construction, is independent from the message m. This is illustrated in Figure 8.2.

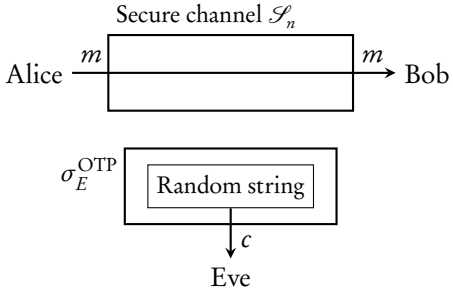


Figure 8.2: The ideal one-time pad system – Alice has access to the left interface, Bob to the right interface and Eve to the lower interface – consists of the ideal secure channel and a simulator σ_E^{OTP} that generates a random string c of length n.

One can easily verify that the two systems depicted in Figures 8.1 and 8.2 are identical, i.e.,

$$\pi_A^{\mathrm{OTP}} \pi_B^{\mathrm{OTP}} (\mathscr{K}_n \| \mathscr{A}_n) = \sigma_E^{\mathrm{OTP}} \mathscr{S}_n,$$

and one says that the OTP has perfect secrecy.

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 [81], i.e.

Proposition 8.2.1: Key requirements for information-theoretic security

$$\{\mathscr{A}_\infty, \mathscr{K}_\ell\} \xrightarrow{\mathrm{ENC}} \mathscr{S}_n \implies \ell \geq n. \tag{8.4}$$

The proof is left to Exercise 8.1. 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 believed[1] to 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 $M = g(C, S)$.

8.3 Quantum cryptography

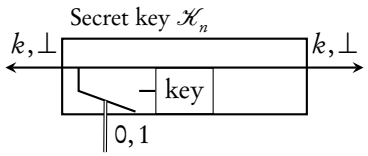
In this section, we will see that Proposition 8.2.1 does not hold in the quantum setup. Having a quantum channel we can achieve

$$\{\mathscr{A}_\infty, \mathscr{Q}_\infty\} \xrightarrow{\pi} \mathscr{S}_n, \tag{8.5}$$

i.e., no key — other than what could be used to generate the authentic channels $\mathscr{A}$, but this can be made arbitrarily small in n — is used to construct the secure channel $\mathscr{S}_n$. (8.5) is obtained by composing a QKD protocol that satisfies (8.1) and the OTP satisfying (8.2). Note that this does not contradict Shannon's proof of Proposition 8.2.1, since in the quantum regime the no-cloning theorem (cf. Proposition 3.5.1) 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.

Ideally, we would like a QKD protocol to construct a secret key resource such as that used by the OTP in Figure 8.1. This is however not possible, since an eavesdropper introducing noise on the quantum channel can always prevent the players from generating a key. Instead, a QKD protocol constructs a weaker resource, which gives the adversary a switch (symbolised by a 1 bit input) that she can activate to prevent a key from being generated, in which case the resource $\mathscr{K}_n$ outputs an error $\perp$ instead of a key k:

[1] In classical cryptography one usually makes statements of the following form. If f was easy to compute then some other function F is also easy to compute. For example F could be the decomposition of a number into its prime factors.



A QKD protocol generally has two parts. In the first, the two players, Alice and Bob, exchange and measure quantum states on an insecure quantum channel. In the second part, the *post-processing*, they use a two-way authenticated classical channel to discuss the results of their measurements and distill a secure key. This is illustrated in Figure 8.3.

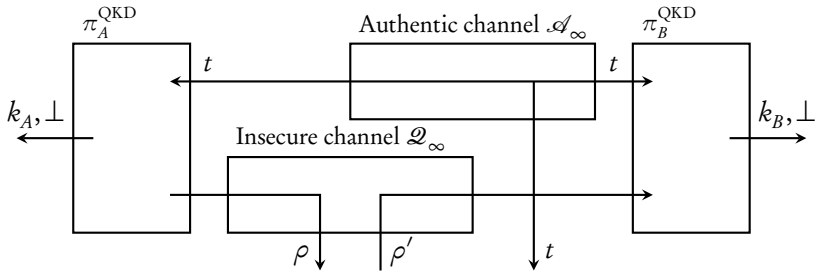


Figure 8.3: The real QKD system — Alice has access to the left interface, Bob to the right interface and Eve to the lower interface — consists of the protocol $(\pi_A^{\mathrm{QKD}}, \pi_B^{\mathrm{QKD}})$, the insecure quantum channel $\mathscr{Q}_\infty$ and two-way authentic classical channel $\mathscr{A}_\infty$. Alice and Bob abort if the insecure channel is too noisy, i.e., if ρ' is not similar enough to ρ to obtain a secret key of the desired length. They run the classical post-processing over the authentic channel, obtaining keys k_A and k_B. The message t depicted on the two-way authentic channel represents the entire classical transcript of the classical post-processing.

To prove that a QKD protocol constructs the resource $\mathscr{K}_n$ depicted above, we need to find a simulator σ_E^{QKD} such that the real and ideal systems are indistinguishable (up to some ε), i.e.,

$$\pi_A^{\mathrm{QKD}} \pi_B^{\mathrm{QKD}} (\mathscr{A}_\infty \| \mathscr{Q}_\infty) \approx_\varepsilon \sigma_E^{\mathrm{QKD}} \mathscr{K}_n.$$

In the real setting (Figure 8.3), Eve has full control over the quantum channel and obtains the entire classical transcript of the protocol. So for the real and ideal settings to be indistinguishable, a simulator σ_E^{QKD} must generate the same communication as in the real setting. This can be done by internally running Alice's and Bob's protocol $(\pi_A^{\mathrm{QKD}}, \pi_B^{\mathrm{QKD}})$, producing the same messages at Eve's interface as the real system. However, instead of letting this (simulated) protocol decide the value of the key as in the real setting, the simulator only checks whether they actually produce a key or an error message, and presses the switch on the secret key resource accordingly. We illustrate this in Figure 8.4.

Let ρ_{ABE} denote the quantum state gathered by a distinguisher interacting with the real system of Figure 8.3 and $\tilde{\rho}_{ABE}$ the stated obtained when interacting with the ideal system from Figure 8.4. We use the subscripts A, B and E to denote the information gathered at Alice's, Bob's and Eve's interface of the systems, i.e., E contains all the information that an eavesdropper may obtain, whereas the A and B registers only hold the final (classical) output of the system. The real and ideal systems are

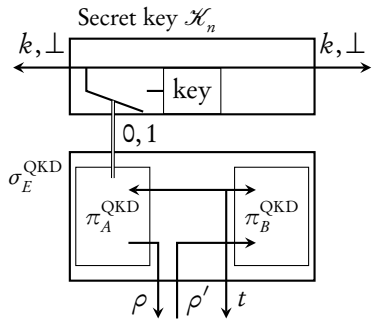


Figure 8.4: The ideal QKD system — Alice has access to the left interface, Bob to the right interface and Eve to the lower interface — consists of the ideal secret key resource $\mathscr{K}_n$ and a simulator σ_E^{QKD}.

indistinguishable (except with probability ε), if (for all possible distinguisher behaviours) the states ρ_{ABE} and $\tilde{\rho}_{ABE}$ are ε-close with respect to the trace distance, i.e.,

$$\delta(\rho_{ABE}, \tilde{\rho}_{ABE}) \leq \varepsilon. \tag{8.6}$$

Note that the validity of quantum theory is essential for this reasoning to hold. Assuming that quantum theory is correct, anything that a distinguisher could possibly do is described within our framework.

(8.6) can be simplified by noting that with the simulator from Figure 8.4, the states of the ideal and real systems are identical when no key is produced. The outputs at Alice's and Bob's interfaces are classical, elements of the set $\{\perp\} \cup \mathscr{K}$, where $\perp$ symbolizes an error and $\mathscr{K}$ is the set of possible keys. The states of the real and ideal systems can be written as

$$\rho_{ABE} = p^{\perp} |\perp_A, \perp_B\rangle\langle\perp_A, \perp_B| \otimes \rho_E^{\perp} + \sum_{k_A, k_B \in \mathscr{K}} p_{k_A, k_B} |k_A, k_B\rangle\langle k_A, k_B| \otimes \rho_E^{k_A, k_B},$$

$$\tilde{\rho}_{ABE} = p^{\perp} |\perp_A, \perp_B\rangle\langle\perp_A, \perp_B| \otimes \rho_E^{\perp} + \frac{1}{|\mathscr{K}|} \sum_{k \in \mathscr{K}} |k, k\rangle\langle k, k| \otimes \sum_{k_A, k_B \in \mathscr{K}} p_{k_A, k_B} \rho_E^{k_A, k_B}.$$

Plugging these in (8.6) we get

$$\delta\left(\rho_{ABE}, \tilde{\rho}_{ABE}\right) = (1 - p^{\perp}) \delta\left(\rho_{ABE}^{\top}, \tau_{AB} \otimes \rho_E^{\top}\right) \leq \varepsilon, \tag{8.7}$$

where

$$\rho_{ABE}^{\top} := \frac{1}{1 - p^{\perp}} \sum_{k_A, k_B \in \mathscr{K}} p_{k_A, k_B} |k_A, k_B\rangle\langle k_A, k_B| \otimes \rho_E^{k_A, k_B} \tag{8.8}$$

is the renormalized state of the system conditioned on not aborting and

$$\tau_{AB} := \frac{1}{|\mathscr{K}|} \sum_{k \in \mathscr{K}} |k, k\rangle\langle k, k|$$

is a perfectly uniform shared key.

We now break (8.7) down into two components, often referred to as *correctness* and *secrecy*. The correctness of a QKD protocol refers to the probability that Alice and Bob end up holding different keys. We say that a protocol is ε_{cor}-*correct* if for all adversarial strategies,

$$\Pr[K_A \neq K_B] \leq \varepsilon_{\text{cor}}, \tag{8.9}$$

where K_A and K_B are random variables over the alphabet $\mathcal{K} \cup \{\perp\}$ describing Alice's and Bob's outputs. The secrecy of a QKD protocol measures how close the final key is to a distribution that is uniform and independent of the adversary's system. Let $p^{\perp}$ be the probability that the protocol aborts, and $\rho_{AE}^{\top}$ be the resulting state of the AE subsystems conditioned on not aborting. A protocol is ε_{sec}-*secret* if for all adversarial strategies,

$$(1-p^{\perp})\delta\left(\rho_{AE}^{\top}, \tau_A \otimes \rho_E^{\top}\right) \leq \varepsilon_{\text{sec}}, \tag{8.10}$$

where the distance $\delta(\cdot,\cdot)$ is the trace distance and τ_A is the fully mixed state.

Proposition 8.3.1: Secrecy and correctness of QKD

If a QKD protocol is ε_{cor}-correct and ε_{sec}-secret, then (8.6) is satisfied for $\varepsilon = \varepsilon_{\text{cor}} + \varepsilon_{\text{sec}}$.

Proof. Let us define γ_{ABE} to be a state obtained from $\rho_{ABE}^{\top}$ (see (8.8)) by throwing away the B system and replacing it with a copy of A, i.e.,

$$\gamma_{ABE} = \frac{1}{1-p^{\perp}} \sum_{k_A,k_B \in \mathcal{K}} p_{k_A,k_B} |k_A,k_A\rangle\langle k_A,k_A| \otimes \rho_E^{k_A,k_B}.$$

From the triangle inequality we get

$$\delta(\rho_{ABE}^{\top}, \tau_{AB} \otimes \rho_E^{\top}) \leq \delta(\rho_{ABE}^{\top}, \gamma_{ABE}) + \delta(\gamma_{ABE}, \tau_{AB} \otimes \rho_E^{\top}).$$

Since in the states γ_{ABE} and $\tau_{AB} \otimes \rho_E^{\top}$ the B system is a copy of the A system, it does not modify the distance. Furthermore, $\mathrm{Tr}_B(\gamma_{ABE}) = \mathrm{Tr}_B(\rho_{ABE}^{\top})$. Hence

$$\delta(\gamma_{ABE}, \tau_{AB} \otimes \rho_E^{\top}) = \delta(\gamma_{AE}, \tau_A \otimes \rho_E^{\top}) = \delta(\rho_{AE}^{\top}, \tau_A \otimes \rho_E^{\top}).$$

For the other term note that

$$\begin{aligned}
&\delta(\rho_{ABE}^{\top}, \gamma_{ABE}) \\
&\quad \leq \sum_{k_A,k_B} \frac{p_{k_A,k_B}}{1-p^{\perp}} \delta\left(|k_A,k_B\rangle\langle k_A,k_B| \otimes \rho_E^{k_A,k_B}, |k_A,k_A\rangle\langle k_A,k_A| \otimes \rho_E^{k_A,k_B}\right) \\
&\quad = \sum_{k_A \neq k_B} \frac{p_{k_A,k_B}}{1-p^{\perp}} = \frac{1}{1-p^{\perp}} \Pr[K_A \neq K_B].
\end{aligned}$$

Putting the above together with (8.7), we get

$$\begin{aligned}
\delta(\rho_{ABE}, \tilde{\rho}_{ABE}) &= (1-p^{\perp})\delta(\rho_{ABE}^{\top}, \tau_{AB} \otimes \rho_E^{\perp}) \\
&\leq \Pr[K_A \neq K_B] + (1-p^{\perp})\delta(\rho_{AE}^{\top}, \tau_A \otimes \rho_E^{\top}),
\end{aligned}$$

which concludes the proof. □

Secrecy and correctness as defined above guarantee that if Alice and Bob obtain keys, then they are identical and unknown to any adversary. But for a QKD protocol to be useful, it must satisfy one more requirement: if no eavesdropper is present, then with high probability the players get a key, i.e., $\Pr[K_A = \perp] \leq \varepsilon$. This is called *robustness*.

8.4 QKD protocol

In the seventies, Wiesner had the idea to construct unforgeable money based on the fact that quantum states cannot be cloned [82]. 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 [83] which is based on Wiesner's ideas and will be explained next.

The main idea of the BB84 protocol is for Alice to send (random) quantum states to Bob, who measures them upon reception. They then publicly compare results to estimate the noise on the channel. Since an eavesdropper necessarily introduces noise, this measure upper bounds how much information an eavesdropper could have about their measurement results. If this is too high, they abort the protocol. If it is low, they then proceed to the classical post-processing: they extract a secret key from their noisy insecure measurement results.

In the following we consider an asymmetric version of BB84, in which Alice generates states in the computational or diagonal basis with different probabilities. Let the computational basis be denoted by $\{|0\rangle,|1\rangle\}$ and the diagonal basis by$\{|\bar{0}\rangle,|\bar{1}\rangle\}$, where $|\bar{0}\rangle := \frac{1}{\sqrt{2}}(|0\rangle+|1\rangle)$ and $|\bar{1}\rangle := \frac{1}{\sqrt{2}}(|0\rangle-|1\rangle)$.

Distribution step Alice and Bob perform the following task N times and let $i = 1,\dots,N$. Alice first chooses a basis $B_i \in \{0,1\}$ with probability p and $1-p$, and a bit $X_i \in \{0,1\}$ with uniform probability. She then prepares a state of a qubit Q_i (with basis $\{|0\rangle,|1\rangle\}$) according to

B	X	Q
0	0	$\|0\rangle$
0	1	$\|1\rangle$
1	0	$\|\bar{0}\rangle$
1	1	$\|\bar{1}\rangle$.

Alice sends Q_i to Bob.

Bob next chooses $B'_i \in \{0,1\}$ also with probability p and $1-p$, 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	$\|1\rangle$	$\|1\rangle$	$\|1\rangle$	$\|\bar{0}\rangle$	$\|0\rangle$	$\|\bar{1}\rangle$	$\|0\rangle$	$\|1\rangle$	$\|\bar{1}\rangle$
B	0	0	0	1	0	1	0	0	1
X	1	1	1	0	0	1	0	1	1
B'	0	0	0	1	1	1	0	1	0
X'	1	1	0	0	1	0	0	0	1
no.	①	②	③	④	⑤	⑥	⑦	⑧	⑨

Hence, Alice and Bob discard columns ⑤, ⑧ and ⑨.

Let z and x denote the substrings of $X_1,\dots,X_N$ that Alice prepared in the computational and diagonal bases, respectively, and kept after the sifting step. And let z' and x' denote the substrings of $X'_1,\dots,X'_N$ that Bob measured in the computational and diagonal bases, respectively, and kept after

the sifting step. In the example given in the table above, z and z' consist of the bits in positions ①, ②, ③ and ⑦, and x and x' consist of the bits in positions ④ and ⑥.

Alice and Bob use the strings x and x' (obtained from the diagonal basis) to estimate the noise on the channel, and they use the strings z and z' to obtain a secret key. Note that it is sufficient to measure the noise in the diagonal basis to estimate the adversary's information about the string z, since a measurement in the computational basis which must be done by Eve to get a bit of z generates noise on qubits that are one of the diagonal basis states.

Checking step Alice sends Bob the string x over the classical *authentic* channel. If the average number of bit flips between x and x' is larger than a predefined tolerance, $\frac{1}{|x|}w(x \oplus x') > Q_{\text{tol}}$, where $w(\cdot)$ is the Hamming weight and $|x|$ the length of x, Bob notifies Alice of this and they abort. Otherwise they continue the protocol.

Alice and Bob now hold two strings z and z' and have an estimate of Eve's information about z (this is computed from Q_{tol}, see Section 8.5.3). In order to extract a secret key from z, they first run an information reconciliation procedure, in which Bob corrects z' to obtain z. Then they run a privacy amplification procedure, which extracts a secret key from z.

Information reconciliation Let $g : \mathscr{Z} \to \mathscr{S}$ and $g' : \mathscr{Z} \times \mathscr{S} \to \mathscr{Z}$ be some predefined functions (where g could be a randomized function). Alice computes a string $s = g(z)$, which she sends to Bob. Bob computes $\tilde{z} = g'(z', s)$, which should be equal to z, if there are not too many bit flips between z and z'. The details for possible choices for the functions g and g' are given in Section 8.5.2.

Error verification The information reconciliation procedure outlined above (only) works if the number of bit flips is not too high. But since Alice and Bob do not know the number of bit flips in the computational basis, they need to check if indeed $z = \tilde{z}$. Let $\mathscr{H} = \{h : \{0,1\}^n \to \{0,1\}^m\}$ be a universal family of hash functions, i.e., for any $z_1 \neq z_2$, $\Pr_h[h(z_1) = h(z_2)] \leq \frac{1}{2^m}$. Alice picks $h \in \mathscr{H}$ at random, and sends both h and $h(z)$ to Bob. If $h(z) \neq h(\tilde{z})$, they abort.

Privacy amplification At this point of the protocol, Alice and Bob hold two strings z and $\tilde{z}$, which are equal with high probability and partially known to Eve. They now need to extract a secret key from these strings. This is done by (randomly) hashing the strings to a shorter string. With high probability, the result is unknown to the adversary. Let $\mathscr{F} = \{f : \{0,1\}^n \to \{0,1\}^\ell\}$ be a family of functions known as an extractor (see Section 8.5.5 for an exact defintion). Alice chooses $f \in \mathscr{F}$ uniformly at random, and sends f to Bob. They set $k_A := f(z)$ and $k_B := f(\tilde{z})$, which they use as secret keys.

8.5 Security proof of BB84

8.5.1 Overview

It took almost 20 years until the security of BB84 could be proven [84–87]. Here we follow the *finite-key* proof of [88], i.e., the proof does not only hold asymptotically — when the number of signals exchanged between the parties goes to infinity — but also for finite values.

To prove the security of the BB84 protocol, i.e., to prove that the real system constructed is indistinguishable from an ideal key and simulator (except with probability ε), one has to show that with high probability Bob holds the same key as Alice at the end of the protocol (correctness), and that Eve

has approximately no information about the key (secrecy). The correctness of the protocol depends only on the information reconciliation step, and is shown in Section 8.5.2.

The main idea of the proof of secrecy is to use an uncertainty relation to bound the information that an eavesdropper has about Alice's string z. This is done in Section 8.5.3. For measurements performed in complementary bases on n qubits, we have

$$H^{\varepsilon}_{\min}(Z|E) + H^{\varepsilon}_{\max}(X|B) \geq n.$$

The secret key is obtained from the string measured in the computational basis, Z, and we need to lower bound Eve's information about Z, namely $H^{\varepsilon}_{\min}(Z|E)$. From the uncertainty relation, this reduces to upper bounding $H^{\varepsilon}_{\max}(X|B)$, information that Bob has about the string Alice would have obtained had she created her qubits in the diagonal basis.

Note that $H^{\varepsilon}_{\max}(X|B)$ does not depend on Eve's system, but only on Alice and Bob. This can be bounded by the number of bit flips that Alice and Bob count in the diagonal basis in the checking step of the protocol. The accuracy of this bound depends on the length of the strings z and x (varying the probability of choosing the computational or diagonal basis in the first step of the protocol changes the expected lengths of these strings). The exact bound is derived in Section 8.5.4.

Now that Alice and Bob have a bound on $H^{\varepsilon}_{\min}(Z|E)$, they can extract a key from Z, which is (approximately) unknown to Eve, using a procedure known as privacy amplification. Roughly speaking, they hash the string Z to a new string that is shorter than Eve's entropy. This is done in Section 8.5.5.

8.5.2 Information reconciliation

The information reconciliation step is a purely classical step, which does not involve Eve and her quantum side information. Alice and Bob each hold a string, z and z' respectively, and Bob wishes to correct z' to obtain z. Alice could simply send him z, but in this case Eve would get this string as well, and they could not obtain any secret key. Instead, Alice sends bob a string $s = g(z)$ which is shorter than z, and allows him to correctly recover $z = g'(z', s)$ if there are not too many differences between z and z'.

Here we will show how to do this using *error correcting codes (ECC)*. ECC are well understood procedures, that are widespread in all the communication means we use on a daily basis, e.g., information on CDs is encoded in an ECC so that the data can be correctly read if the CD has minor scratches. The basic principle of an ECC is to encode a string in a higher dimension by adding redundancy, e.g., a 3 bit string x_1, x_2, x_3 can be encoded into 7 bits as $x_1, x_2, x_3, x_1 \oplus x_2, x_1 \oplus x_3, x_2 \oplus x_3, x_1 \oplus x_2 \oplus x_3$. Upon receiving a 7 bit string $y_1, \ldots, y_7$, the decoding procedure searches for the valid code word that is the closest (measured in the number of bit flips) to the received string, and decodes to the corresponding message x_1, x_2, x_3. One can easily verify that with the code given in this paragraph, any two valid code words have distance at least 3 from each other. Thus, if only 1 bit flip occurred, the decoding procedure always works. In practical ECC the same is done for strings of thousands or millions of bits and a constant rate of errors can be corrected.

The same principle can be used in the information reconciliation step of QKD. Alice and Bob have an idea of how much noise they will find on the quantum channel under normal conditions, and choose an ECC that can correct the corresponding number of bit flips. But instead of just one ECC, they choose a family of codes that cover the entire code word space, and which is chosen before

running the protocol. For example, consider the following $2^4 = 32$ ECC:

$$
\begin{array}{c}
x_1, x_2, x_3, x_1 \oplus x_2, x_1 \oplus x_3, x_2 \oplus x_3, x_1 \oplus x_2 \oplus x_3 \\
x_1, x_2, x_3, \overline{x_1 \oplus x_2}, x_1 \oplus x_3, x_2 \oplus x_3, x_1 \oplus x_2 \oplus x_3 \\
x_1, x_2, x_3, x_1 \oplus x_2, \overline{x_1 \oplus x_3}, x_2 \oplus x_3, x_1 \oplus x_2 \oplus x_3 \\
x_1, x_2, x_3, x_1 \oplus x_2, x_1 \oplus x_3, \overline{x_2 \oplus x_3}, x_1 \oplus x_2 \oplus x_3 \\
x_1, x_2, x_3, x_1 \oplus x_2, x_1 \oplus x_3, x_2 \oplus x_3, \overline{x_1 \oplus x_2 \oplus x_3} \\
x_1, x_2, x_3, \overline{x_1 \oplus x_2}, \overline{x_1 \oplus x_3}, x_2 \oplus x_3, x_1 \oplus x_2 \oplus x_3 \\
\dots \\
x_1, x_2, x_3, \overline{x_1 \oplus x_2}, \overline{x_1 \oplus x_3}, \overline{x_2 \oplus x_3}, \overline{x_1 \oplus x_2 \oplus x_3},
\end{array}
$$

where $\bar{x}$ is the negation of the bit x. Each line corresponds to one ECC which encodes 2^3 messages into a space of dimension 2^7. The 2^4 codes in this table cover the entire code space with their $2^4 2^3 = 2^7$ code words.

The string z that Alice holds corresponds to a code word. The function $g(z)$ assigns to every code word the corresponding code from the predefined list, i.e., $s = g(z)$ is an ECC (e.g., the line number of the corresponding ECC in the list) in which z is a valid word. Once Bob has received this information s, he runs the corresponding decoding procedure which searches for the nearest neighbour to z' in the code described by s. Let this nearest neighbour be $\tilde{z}$. If the number of bit flips between z and z' is less than the number of errors that the code can correct, then we have $z = \tilde{z}$.

Of course, Alice and Bob do not know how many bit flips really occurred. In particular if an eavesdropper was active, this might be very different from what they expect under normal conditions. So to check if the information reconciliation procedure worked, they run the error verification step. Alice picks a function from a universal hash family, i.e., $\mathscr{H} = \{h : \{0,1\}^n \to \{0,1\}^m\}$ such that for any $z_1 \neq z_2$, $\Pr_h[h(z_1) = h(z_2)] \leq \frac{1}{2^m}$. And Alice and Bob publicly compare their value of the hash of z and $\tilde{z}$. If $z \neq \tilde{z}$, then with probability $p \geq \frac{1}{2^m}$ the hashes will not match and the protocol aborts. Hence, if the players want the protocol to have a correctness parameter ε_{cor} they need to choose a family of hash functions with $m \geq \log \frac{1}{\varepsilon_{\text{cor}}}$.

This entire procedure provides Eve with some information about z, which she obtains from s and $h(z)$. But this information can be bounded by the number of bits of these two strings. This is done in Section 8.5.5.

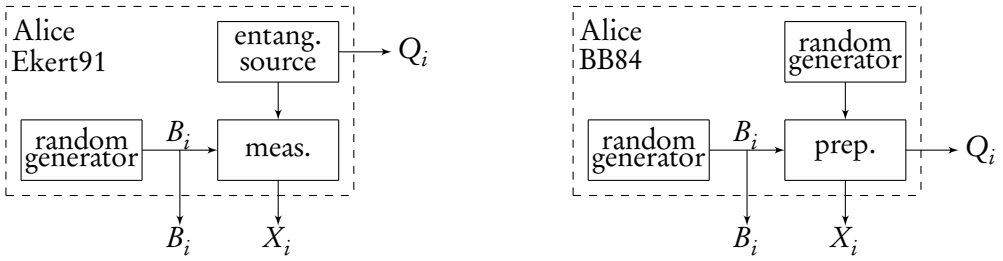
8.5.3 Uncertainty relation

The uncertainty relation makes a statement about the entropy of two different measurements performed on the same system, one in the computational basis and the other in diagonal basis. However, in the BB84 protocol, Alice *prepares* her qubits in two different bases, she does not *measure*. To get around this, we consider an alternative entanglement-based protocol (called Ekert91 [89]), which only differs from BB84 in the first step:

Distribution step of Ekert91 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) chosen with probabilities p and $1-p$, respectively. They repeat this step N times.

[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.

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.



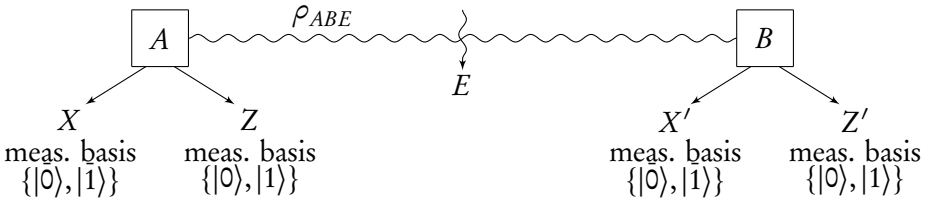
In the Ekert91 protocol Alice's task is described by a state

$$\rho^{\text{Ekert91}}_{B_i X_i Q_i} = \sum_{b\in\{0,1\}} p_b |b\rangle\langle b|_{B_i} \otimes \frac{1}{2} \sum_{x\in\{0,1\}} |x\rangle\langle x|_{X_i} \otimes |\varphi^{b,x}\rangle\langle\varphi^{b,x}|_{Q_i}, \tag{8.11}$$

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

$$\rho^{\text{BB84}}_{B_i X_i Q_i} = \sum_{b\in\{0,1\}} p_b |b\rangle\langle b|_{B_i} \otimes \frac{1}{2} \sum_{x\in\{0,1\}} |x\rangle\langle x|_{X_i} \otimes |\varphi^{b,x}\rangle\langle\varphi^{b,x}|_{Q_i}. \tag{8.12}$$

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 securtiy, the BB84 protocol is technologically simpler to implement.



Now, neither player needs to perform the measurement in the computational basis until after the checking step, since the resulting string is only needed for the classical post-processing (information reconciliation, error verification, and privacy amplification). Let ρ^{pass}_{ABE} denote the state shared by Alice, Bob and Eve after having measured the corresponding qubits in the diagonal basis, after having shared the basis choices and having performed the sifting step, after having publicly shared x and conditioned on passing the test $\frac{1}{|x|} w(x \oplus x') \leq Q_{\text{tol}}$, but before measuring the qubits of A and B in the computational basis. We apply the following uncertainty relation to this state ρ^{pass}_{ABE},

$$H^{\varepsilon}_{\min}(Z|E)_{\rho} + H^{\varepsilon}_{\max}(X|B)_{\rho} \geq n,$$

where the right-hand side is n because the system A has n-qubits and the two possible measurements are complimentary.

In the privacy amplification step (Section 8.5.5) we need a bound on the adversary's (smooth) min-entropy of the raw key Z. This is directly obtained from the uncertainty relation, since

$$H_{\min}^{\varepsilon}(Z|E)_{\rho} \geq n - H_{\max}^{\varepsilon}(X|B)_{\rho}.$$

So the problem reduces to upper bounding $H_{\max}^{\varepsilon}(X|B)_{\rho}$. Note that Alice never actually measures these n qubits in the diagonal basis. Bounding this is a *gedankenexperiment*: how much entropy would Bob have about X were Alice to perform a measurement of the A system in the diagonal basis?

8.5.4 Finite key statistics

Even though Alice does not perform a measurement on these qubits in the diagonal basis to obtain X, we can estimate Bob's information about X from the number of bit flips on the other qubits that they did measure in the diagonal basis. Think of a bucket filled with red and white balls (bits with flips and without flips). If we pick k balls at random from the bucket, and a proportion of μ of these balls are red, then we can be fairly certain that roughly the same proportion μ of the balls remaining in the bucket must be red. A bound known as the Serfling bound makes this precise.

Let $\Lambda_{\text{key}} = \frac{1}{n} w(X \oplus X')$ be the random variable denoting the proportion of bit flips that Alice and Bob would find were they to perform measurements on these n qubits in the diagonal basis. Let $\Lambda_{\text{PE}} = \frac{1}{k} w(\bar{X} \oplus \bar{X}')$ denote the proportion of bit flips that Alice and Bob find on the qubits that they did measure in the diagonal basis (the Parameter Estimation qubits). Let $\Lambda_{\text{tot}} = \frac{n}{n+k}\Lambda_{\text{key}} + \frac{k}{n+k}\Lambda_{\text{PE}}$ denote the total proportion of bit flips on all $n+k$ qubits. The Serfling bound tells us that

$$\Pr\Big[\Lambda_{\text{key}} \geq \lambda_{\text{tot}} + \delta | \Lambda_{\text{tot}} = \lambda_{\text{tot}}\Big] \leq e^{\frac{-2n(n+k)\delta^2}{k+1}}.$$

We now use this to bound the probability that the proportion of bit flips in the n key qubits is different from the proportion of bit flips in the Parameter Estimation qubits, conditioned on passing the test $\frac{1}{|x|} w(x \oplus x') \leq Q_{\text{tol}}$.

$$\begin{aligned}
\Pr\Big[\Lambda_{\text{key}} \geq \Lambda_{\text{PE}} + \mu | \text{“pass”}\Big] &\leq \frac{1}{p_{\text{pass}}} \Pr\Big[\Lambda_{\text{key}} \geq \Lambda_{\text{PE}} + \mu\Big] \\
&\leq \frac{1}{p_{\text{pass}}} \Pr\left[\Lambda_{\text{key}} \geq \Lambda_{\text{tot}} + \frac{k}{n+k}\mu\right] \\
&= \frac{1}{p_{\text{pass}}} \sum_{\lambda} \Pr[\Lambda_{\text{tot}} = \lambda] \Pr\left[\Lambda_{\text{key}} \geq \Lambda_{\text{tot}} + \frac{k}{n+k}\mu \middle| \Lambda_{\text{tot}} = \lambda\right] \\
&\leq \frac{1}{p_{\text{pass}}} e^{\frac{-2nk^2\mu^2}{(n+k)(k+1)}}.
\end{aligned}$$

For convenience we set $\varepsilon_{\text{PE}} = e^{\frac{-nk^2\mu^2}{(n+k)(k+1)}}$ from which we get

$$\Pr\Big[\Lambda_{\text{key}} \geq \Lambda_{\text{PE}} + \mu | \text{“pass”}\Big] \leq \frac{\varepsilon_{\text{PE}}^2}{p_{\text{pass}}}$$

for $\mu = \sqrt{\frac{(n+k)(k+1)}{nk^2} \ln \frac{1}{\varepsilon_{\text{PE}}}}$.

We now have a bound on the bit flips on X and use it to bound the smooth max-entropy of X given B. Let X' denote Bob's random variable were he to measure his n qubits in the diagonal basis, let $P_{XX'}$ be the corresponding joint probability distribution of X and X' and let $Q_{XX'}$ be a distribution very similar to $P_{XX'}$, except that all values of X and X' for which $\Lambda_{\text{key}} \geq Q_{\text{tol}} + \mu$ have probability 0. A quick calculation shows that the purified distance between these probability distributions is bounded by

$$P(P_{XX'}, Q_{XX'}) \leq \frac{\varepsilon_{\text{PE}}}{\sqrt{p_{\text{pass}}}}.$$

For $\varepsilon = \frac{\varepsilon_{\text{PE}}}{\sqrt{p_{\text{pass}}}}$ we thus have

$$H^{\varepsilon}_{\max}(X|B)_{\rho} \leq H^{\varepsilon}_{\max}(X|X')_{P} \leq H_{\max}(X|X')_{Q} \leq nh(Q_{\text{tol}} + \mu),$$

where $h(\cdot)$ is the binary entropy, $h(p) = -p\log p - (1-p)\log(1-p)$.

8.5.5 Privacy amplification

In Section 8.5.4 we derived the bound

$$H^{\varepsilon}_{\min}(Z|E) \geq n(1 - h(Q_{\text{tol}} + \mu))$$

for $\varepsilon = \frac{\varepsilon_{\text{PE}}}{\sqrt{p_{\text{pass}}}}$ and $\mu = \sqrt{\frac{(n+k)(k+1)}{nk^2} \ln \frac{1}{\varepsilon_{\text{PE}}}}$. In this equation, the system E contains all the information gathered by an eavesdropper by interacting with the quantum channel. The information reconciliation step then leaks more information about the random variable Z. Let $S = g(Z)$ be the random variable sent to Bob for information reconciliation and $V = h(Z)$ be the random variable sent for error verification. The complete system $E' = ESV$ which the adversary (or distinguisher) obtains also contains S and V.

From an entropy chain rule, we have

$$H^{\varepsilon}_{\min}(Z|ESV) \geq H^{\varepsilon}_{\min}(Z|E) - \log|\mathcal{S}| - \log|\mathcal{V}|.$$

Note that $\log|\mathcal{S}| = \text{leak}_{\text{ECC}}$ and $\log|\mathcal{V}| = \log\frac{1}{\varepsilon_{\text{cor}}}$ are parameters of the protocol which are fixed in advance. Hence,

$$H^{\varepsilon}_{\min}(Z|E') \geq n(1 - h(Q_{\text{tol}} + \mu)) - \text{leak}_{\text{ECC}} - \log\frac{1}{\varepsilon_{\text{cor}}}.$$

Now that we have a bound on the smooth min-entropy, we can extract a secret key from Z. To do this, we use an extractor.

Definition 8.5.1

A function $F : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{\ell}$ is a (strong, quantum-proof) (k, ε)-extractor if for any cq state $\rho_{ZE} = \sum_z p_z |z\rangle\langle z| \otimes \rho_E^z$ with $H_{\min}(Z|E)_{\rho} \geq k$ and a uniform seed Y on $\{0,1\}^d$,

$$\delta(\rho_{F(Z,Y)YE}, \tau_U \otimes \rho_Y \otimes \rho_E) \leq \varepsilon,$$

where τ_U is the fully mixed state in dimension 2^ℓ and

$$\rho_{F(Z,Y)YE} = \sum_{z,y} \frac{p_z}{2^d} |F(z,y)\rangle\langle F(z,y)| \otimes |y\rangle\langle y| \otimes \rho_E^z.$$

Note that picking a seed y uniformly at random and applying $F(\cdot, y)$ to a string z is the same as picking a function $f \in \mathscr{F}$ uniformly at random and applying it to z for $\mathscr{F} = \{F(\cdot, y)\}_y$.

Intuitively, the system E will contain information about some of the bits of Z. Some of the seeds y might be "bad" in the sense that the output of the corresponding function $F(\cdot, y)$ might depend strongly on the bits of Z known by E. However, the definition of an extractor guarantees that on average over the choice of y, this is very unlikely.

Proving that a specific function F satisfies the definition of an extractor is beyond the scope of this lecture. It is sufficient to know that there are many constructions. One particularly practical construction is obtained by choosing at random a function from a family of universal hash functions. This construction satisfies Definition 8.5.1 with $\varepsilon = \frac{1}{2}\sqrt{2^{\ell-k}}$ for any input size n [87]. This means that the error is exponentially small in the number of bits "sacrificed", namely $k - \ell$, the difference between the entropy of the input and the output length.

Our bound on the entropy of Z is given by the smooth min-entropy, not the min-entropy. In Exercise 8.4 it is shown that if F satisfies Definition 8.5.1, then for any state ρ_{ZE} with smooth min-entropy $H_{\min}^{\bar{\varepsilon}}(Z|E)_\rho \geq k$,

$$\delta(\rho_{F(Z,Y)YE}, \tau_U \otimes \rho_Y \otimes \rho_E) \leq \varepsilon + 2\bar{\varepsilon}.$$

The last step of the QKD protocol thus consists in applying an extractor to Z for an output chosen small enough to get a satisfying error $\varepsilon + 2\bar{\varepsilon}$. The notion of secrecy introduced in Section 8.3 requires that

$$p_{\text{pass}} \delta\left(\rho_{AE}^{\top}, \tau_A \otimes \rho_E^{\top}\right) \leq \varepsilon_{\text{sec}},$$

where the choice of function f used in privacy amplification (the seed y) is included in the E register. Plugging in the error for an extractor from universal hash functions in this, we get

$$\begin{aligned} p_{\text{pass}} \delta\left(\rho_{AE}^{\top}, \tau_A \otimes \rho_E^{\top}\right) &\leq p_{\text{pass}} \left(\frac{1}{2}\sqrt{2^{\ell-k}} + 2\frac{\varepsilon_{\text{PE}}}{\sqrt{p_{\text{pass}}}} \right) \\ &\leq \frac{1}{2}\sqrt{2^{\ell-k}} + 2\varepsilon_{\text{PE}}. \end{aligned}$$

8.6 Exercises

Exercise 8.1. One-time pad → solution

Consider three random variables: a message M, a secret key K and a ciphertext C. We want to encode M as a ciphertext C using K with perfect secrecy, so that no one can guess the message from the cipher: $I(M:C)=0$. After the transmission, we want to be able to decode the ciphertext: someone that knows the key and the cipher should be able to obtain the message perfectly, i.e. $H(M|C,K)=0$. Show that this is only possible if the key contains at least as much randomness as the message, namely $H(K)\geq H(M)$.

Exercise 8.2. Secrecy and correctness → solution

Let ρ_{ABE} be the tripartite ccq-state held by Alice, Bob and Eve after a run of a QKD protocol. We showed in the lecture that if the protocol is ε_1-secret,

$$p_{\text{key}}\delta\left(\rho_{AE}^{\text{key}},\tau_A\otimes\rho_E^{\text{key}}\right)\leq\varepsilon_1,$$

and ε_2-correct,

$$\Pr[A\neq B]\leq\varepsilon_2,$$

then the real and ideal systems are $\varepsilon=\varepsilon_1+\varepsilon_2$ indistinguishable, i.e.,

$$\exists\sigma_E \text{ such that } \pi_A\pi_B(\mathscr{A}\|\mathscr{Q})\approx_\varepsilon \sigma_E\mathscr{K}. \tag{8.13}$$

Show that if (8.13) holds for some ε, then the protocol must be ε-correct and 2ε-secret.

Hint: you cannot assume that (8.13) is necessarily satisfied by the same simulator used to prove the converse.

Exercise 8.3. Min-entropy chain rule → solution

Let ρ_{XZE} be a ccq-state. Show that the following holds:

$$H_{\min}^{\varepsilon}(X|ZE)_\rho\geq H_{\min}^{\varepsilon}(X|E)_\rho-\log|\mathscr{Z}|.$$

Exercise 8.4. Privacy amplification with smooth min-entropy → solution

A function $F:\{0,1\}^n\times\{0,1\}^d\to\{0,1\}^m$ is a (quantum-proof, strong) (k,ε)-extractor if for all cq states ρ_{XE} with $H_{\min}(X|E)\geq k$ and a uniform Y,

$$\delta\left(\rho_{F(X,Y)YE},\tau_U\otimes\tau_Y\otimes\rho_E\right)\leq\varepsilon.$$

Show that for any (k,ε)-extractor F, if a cq state ρ_{XE} has smooth min-entropy $H_{\min}^{\bar{\varepsilon}}(X|E)\geq k$, then

$$\delta\left(\rho_{F(X,Y)YE},\tau_U\otimes\tau_Y\otimes\rho_E\right)\leq\varepsilon+2\bar{\varepsilon}.$$

Exercise 8.5. Quantum one-time pad → solution

The quantum one-time pad encrypts a one qubit message ρ with two bits of key k_1,k_2 as

$$\mathscr{E}_{k_1,k_2}(\rho)=X^{k_1}Z^{k_2}\rho Z^{k_2}X^{k_1}.$$

Because $\mathscr{E}_{k_1,k_2}$ is unitary, it guarantees that the encryption is reversible (given the secret key k_1,k_2) and the receiver can read the message.

For any purification $|\psi\rangle_{AB}$ of ρ_A, the mixture over all possible keys is then

$$\frac{1}{4}\sum_{k_1,k_2} \mathcal{E}^A_{k_1,k_2} \otimes \mathcal{I}^B(|\psi\rangle\langle\psi|_{AB}) = \tau_A \otimes \rho_B, \tag{8.14}$$

where τ_A is the fully mixed state and $\rho_B = \mathrm{Tr}_A(|\psi\rangle\langle\psi|_{AB})$.

Show that using two bits of key per qubit of message is optimal, i.e., for any alternative (but still reversible) definition of the encryption operation $\mathcal{E}_k$, (8.14) can only be satisfied for any state $|\psi\rangle$ if the key k is at least 2 bits.

Exercise 8.6. Bit commitment → solution

Hooke's law, regarding the stiffness of springs, was first published in 1676 as an anagram 'ceiiinosssttuu', and only unscrambled in 1678 to read (in Latin) *ut tensio sic vis* (as the extension so the force).[90] The use of anagrams was apparently common at the time, as it gave a means of committing to a particular result without immediately revealing it. This bought time to build upon the discovery, while still being able to claim priority.

This is the idea of the bit commitment cryptographic protocol. The protocol consists of two steps. In step one, Alice commits a bit to Bob, giving him some physical system that contains the bit, but which conceals it from him. We could think of the bit written on a piece of paper and stored in an impenetrable safe. In the second step, Alice unveils the bit, giving Bob the key to the safe.

Consider the following bit commitment protocol for a bit b, using quantum systems. To commit, Alice generates a random string $X = \{0,1\}^n$ and encodes every bit into a qubit using a basis $B_0 = \{|0\rangle, |1\rangle\}$ if $b = 0$ or $B_1 = \{|+\rangle, |-\rangle\}$ if $b = 1$. These qubits are sent to Bob and he stores them. To unveil the bit, Alice sends b and X to Bob and he will validate the process by applying measurements on his states in basis B_b and comparing the results with X.

a) Show that Bob has no information about b before it is revealed, i.e. the protocol is concealing.

b) Show that if Alice commits honestly to 0, the probability of her unveiling a 1 without Bob noticing the cheat is equal to 2^{-n}.

c) Give a strategy that allows Alice to cheat perfectly, i.e. that allows her to unveil 0 or 1 in such a way that Bobs probability of detecting her cheating is zero.
Hint: Consider the steering phenomenon.

Exercise 8.7. Data hiding → solution

Suppose you have two agents, Alice and Bond, at your service. You want them to deliver a secret (classical) message to your ally Charlie. You will give Alice and Bond two different states (i.e. an encryption of your message), so that they cannot extract the secret message unless they are physically together. Specifically, data hiding is what you want: states that are easily distinguishable by doing a certain class of operations, such as a global measurement on Alice and Bond's systems together, but they are nearly indistinguishable under a different, restricted class of operations, such as local operations and classical communication (LOCC). Formally, we say that a family of states $\{\rho^i\}_i$ is ε-secure under a set of operations $\mathbb{E}$ if

$$\delta(\mathcal{E}(\rho^i), \mathcal{E}(\rho^j)) < \varepsilon, \quad \forall i,j, \quad \forall \mathcal{E} \in \mathbb{E}.$$

In this exercise we will consider a data hiding scheme which is secure under LOCC and so the original message can only be recovered if global measurements on the joint system are allowed. Consider a $2d$-qubit Hilbert space, $\mathscr{H}_A \otimes \mathscr{H}_B$, and the computational basis of both spaces. Consider the projectors onto the symmetric and antisymmetric subspaces of $\mathscr{H}_A \otimes \mathscr{H}_B$,

$$\Pi^S = \frac{1}{2}\sum_{i<j}\Big(|i\rangle_A|j\rangle_B + |j\rangle_A|i\rangle_B\Big)\Big(\langle i|_A\langle j|_B + \langle j|_A\langle i|_B\Big) + \sum_i |i\rangle_A|i\rangle_B\langle i|_A\langle i|_B,$$
$$\Pi^A = \frac{1}{2}\sum_{i<j}\Big(|i\rangle_A|j\rangle_B - |j\rangle_A|i\rangle_B\Big)\Big(\langle i|_A\langle j|_B - \langle j|_A\langle i|_B\Big).$$

You will encode only one bit of information, b, giving Alice and Bond each their d—qubit part of ρ^b_{AB}, with

$$\rho^{b=0} = \frac{2}{d(d+1)}\Pi^S, \qquad \rho^{b=1} = \frac{2}{d(d-1)}\Pi^A.$$

a) Show that $\rho^{b=0}$ and $\rho^{b=1}$ are valid density operators and explain how you would proceed to recover b if you had access to Alice and Bond's systems (together).

b) Consider the flip operator in basis $\{|i\rangle_A|j\rangle_B\}_{ij}$,

$$F = \Pi^S - \Pi^A = \sum_{i,j}|i\rangle_A|j\rangle_B\langle j|_A\langle i|_B.$$

Show that, for all operators $M_A \in \mathrm{End}(\mathscr{H}_A), N_B \in \mathrm{End}(\mathscr{H}_B)$,

$$\mathrm{Tr}[F(M_A \otimes N_B)] = \mathrm{Tr}(M_A N_B).$$

In particular, for all pure states $|x\rangle_A, |y\rangle_B$, $\mathrm{Tr}[F|xy\rangle\langle xy|] = |\langle x|y\rangle|^2$.

c) Suppose that Alice and Bond perform local projective measurements in arbitrary bases $\{|x\rangle_A\}$ and $\{|y\rangle_B\}$ respectively. We call the joint probability distribution of the outcomes P_{XY} when they measure state $\rho^{b=0}$ and Q_{XY} when they measure $\rho^{b=1}$. We want them to be unable to determine which state they measured, i.e., to distinguish the two distributions, so we want to show that $\delta(P_{XY}, Q_{XY})$ is small. Remember that

$$P_{XY}(x,y) = \mathrm{Tr}(|xy\rangle\langle xy|\rho^{b=0}), \qquad Q_{XY}(x,y) = \mathrm{Tr}(|xy\rangle\langle xy|\rho^{b=1}).$$

Use the results from b) to show that $\delta(P_{XY}, Q_{XY}) \leq \frac{2}{d+1}$.

Hint: start from the trace distance as

$$\delta(P_{XY}, Q_{XY}) = \sum_{x,y\in\mathscr{S}} P_{XY}(x,y) - Q_{XY}(x,y),$$

with $\mathscr{S} = \{(x,y) : P_{XY}(x,y) > Q_{XY}(x,y)\}$.

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 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 $\mathbb{1}$. 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 $SS^*=S^*S=\mathbb{1}$, and *self-adjoint* (or *Hermitian*) 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 $U^*U=\mathbb{1}_{\mathcal{H}}$. 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 *bra-ket notation*, where vectors are interpreted as operators. More precisely, we can associate any vector $v\in\mathcal{H}$ with an endomorphism $|v\rangle\in\mathrm{Hom}(\mathbb{C},\mathcal{H})$, called *ket* and defined as

$$|v\rangle:\quad \gamma\mapsto\gamma v, \tag{A.2}$$

[1] Augustin-Louis Cauchy, 1789 – 1857, French mathematician.

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*[2] *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 $u, v \in \mathscr{H}$,

$$\langle u| \circ |v\rangle = (u, v). \tag{A.4}$$

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 $\mathbb{1} \in \mathrm{End}(\mathscr{H})$ can be written as

$$\mathbb{1} = \sum_i |b_i\rangle\langle b_i| \tag{A.6}$$

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.

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

$$S = \mathbb{1}\, S\, \mathbb{1} \tag{A.8}$$

$$= \sum_{j,k} |b_j\rangle\langle b_j|S|b_k\rangle\langle b_k| \tag{A.9}$$

$$= \sum_{j,k} S_{j,k}|b_j\rangle\langle b_k|. \tag{A.10}$$

[2] Frigyes Riesz, 1880 – 1956, Hungarian mathematician.

Now the basis dependence is plain to see. Matrix representations can be given for more general operators $S \in \mathrm{Hom}(\mathscr{H}, \mathscr{H}')$ by the same technique:

$$S = \mathbb{1}_{\mathscr{H}'} \, S \, \mathbb{1}_{\mathscr{H}} \tag{A.11}$$
$$= \sum_{j,k} |b'_j\rangle\langle b'_j|S|b_k\rangle\langle b_k| \tag{A.12}$$
$$= \sum_{j,k} S_{j,k} |b'_j\rangle\langle b_k|. \tag{A.13}$$

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

$$|v\rangle = \sum_j v_j |b_j\rangle. \tag{A.14}$$

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

$$\langle u|Sv\rangle = \langle S^* u|v\rangle. \tag{A.15}$$

In terms of matrix representatives, reading the above from right to left we have

$$(S^* u)_j^* \cdot v_j = u_j^* \cdot (Sv)_j \tag{A.16}$$
$$= \sum_{jk} u_j^* S_{jk} v_k \tag{A.17}$$
$$= \sum_{jk} ([S_{jk}]^* u_j)^* v_k \tag{A.18}$$
$$= \sum_{jk} ([S_{jk}]^\dagger u_k)^* v_j. \tag{A.19}$$

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 $\mathscr{H}_A$ and $\mathscr{H}_B$, we may formally define their product $u \times v$, which is an element of the *Cartesian*[3] *product* $\mathscr{H}_A \times \mathscr{H}_B$. 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

[3] René Descartes, 1596 – 1650, French philosopher and mathematician.

enforce this sort of linearity on $\mathscr{H}_A \times \mathscr{H}_B$. 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 $\mathscr{H}_A \times \mathscr{H}_B$ 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 $\mathscr{H}_A \otimes \mathscr{H}_B$.

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

$$(S \otimes T)(u \otimes v) := (Su) \otimes (Tv) \tag{A.22}$$

for any $u \in \mathscr{H}_A$ and $v \in \mathscr{H}_B$. 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.

$$\mathrm{Hom}(\mathscr{H}_A, \mathscr{H}'_A) \otimes \mathrm{Hom}(\mathscr{H}_B, \mathscr{H}'_B) \simeq \mathrm{Hom}(\mathscr{H}_A \otimes \mathscr{H}_B, \mathscr{H}'_A \otimes \mathscr{H}'_B). \tag{A.23}$$

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 $u \in \mathscr{H}_A$ and $v \in \mathscr{H}_B$.

A.5 Trace and partial trace

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

$$\mathrm{Tr}(S) := \sum_i \langle b_i|S|b_i\rangle, \tag{A.25}$$

where $\{|b_i\rangle\}_i$ is any orthonormal basis of $\mathscr{H}$. 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 $S, T \in \mathrm{End}(\mathscr{H})$ 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 $\mathrm{End}(\mathscr{H}_A \otimes \mathscr{H}_B)$ on a product space $\mathscr{H}_A \otimes \mathscr{H}_B$ onto the endomorphisms $\mathrm{End}(\mathscr{H}_A)$ on $\mathscr{H}_A$. (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 $S \in \mathrm{End}(\mathscr{H}_A)$ and $T \in \mathrm{End}(\mathscr{H}_B)$.

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 $T_A \otimes \mathbb{1}_B$ where $T_A \in \mathrm{End}(\mathscr{H}_A)$. That is, for any operator $S_{AB} \in \mathrm{End}(\mathscr{H}_A \otimes \mathscr{H}_B)$,

$$\mathrm{Tr}_B\big(S_{AB}(T_A \otimes \mathbb{1}_B)\big) = \mathrm{Tr}_B(S_{AB})T_A \tag{A.31}$$

and

$$\mathrm{Tr}_B\big((T_A \otimes \mathbb{1}_B)S_{AB}\big) = T_A \mathrm{Tr}_B(S_{AB}). \tag{A.32}$$

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 $S_{ABC} \in \mathrm{End}(\mathscr{H}_A \otimes \mathscr{H}_B \otimes \mathscr{H}_C)$,

$$\mathrm{Tr}_{AB}(S_{ABC}) = \mathrm{Tr}_A(\mathrm{Tr}_B(S_{ABC})). \tag{A.34}$$

A.6 Decompositions of operators and vectors

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

$$S = UDV^*. \tag{A.35}$$

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

$$S = \sqrt{SS^*}U \tag{A.36}$$

and

$$S = U\sqrt{S^*S}. \tag{A.37}$$

Spectral decomposition. Let $S \in \mathrm{End}(\mathscr{H})$ be normal and let $\{|b_i\rangle\}_i$ be an orthonormal basis of $\mathscr{H}$. Then there exists a unitary $U \in \mathrm{End}(\mathscr{H})$ and an operator $D \in \mathrm{End}(\mathscr{H})$ 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 $S \in \mathrm{End}(\mathscr{H})$, there exists a basis $\{|b_i\rangle\}_i$ of $\mathscr{H}$ 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 inner product* between two operators $S, T \in \mathrm{End}(\mathscr{H})$ 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 $S, T \in \mathrm{End}(\mathscr{H})$. If $S \geq 0$ and $T \geq 0$ then*

$$\mathrm{Tr}(ST) \geq 0. \tag{A.43}$$

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 $\phi \in \mathscr{H}$,

$$\langle \phi | V^* V | \phi \rangle = ||V\phi||^2 \geq 0. \tag{A.45}$$

□

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* $S \in \mathrm{End}(\mathscr{H})$,

$$||S||_1 = \max_U |\mathrm{Tr}(US)| \tag{A.49}$$

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

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[4] 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 $U := V^*$. □

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

$$E_{i,j} := \begin{cases} \frac{1}{\sqrt{2}}|e_i\rangle\langle e_j| + \frac{1}{\sqrt{2}}|e_j\rangle\langle e_i| & \text{if } i < j \\ \frac{i}{\sqrt{2}}|e_i\rangle\langle e_j| - \frac{i}{\sqrt{2}}|e_j\rangle\langle e_i| & \text{if } i > j \\ |e_i\rangle\langle e_i| & \text{otherwise} \end{cases} \tag{A.53}$$

[4] Karl Hermann Amandus Schwarz, 1843 – 1921, German mathematician.

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).

A.9 Norm inequalities

Let $x \in \mathbb{R}^n$ then $||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2$.

Proof. By definition of the 1 norm, we can write

$$||x||_1^2 = \left(\sum_i |x_i|\right)^2 = \sum_i x_i^2 + \sum_{i \neq j} |x_i||x_j| \geq \sum_i x_i^2 = ||x||_2^2. \tag{A.56}$$

Some simple algebra gives

$$0 \leq \frac{1}{2}\sum_{ij}(|x_i| - |x_j|)^2 = \frac{1}{2}\sum_{ij}\left(|x_i|^2 + |x_j|^2 - 2|x_i||x_j|\right) = n\sum_i |x_i|^2 - \sum_{i,j}|x_i||x_j| = n||x||_2^2 - ||x||_1^2,$$

which proves the assertion. □

A.10 A useful operator inequality

Lemma A.10.1. *For all* $0 \leq S \leq \mathbb{1}$, $T \geq 0$, *and any constant* $a \geq 0$,

$$\mathbb{1} - (S+T)^{-1/2} S (S+T)^{-1/2} \leq (1+a)(\mathbb{1} - S) + (2 + a + a^{-1})T. \tag{A.57}$$

B

Solutions to Exercises

B.1 Exercises from Chapter 2

2.1 Statistical distance

a) The lower bound follows immediately from the fact that $S = \emptyset$ (or $S = X$) is an event. The upper bound follows because $P[S] \leq 1$ and $Q[S] \geq 0$ for every event S. To see the triangle inequality, consider the following chain of inequalities

$$\begin{aligned}
\delta(P,Q) &= \sup_{S\subseteq X} |P[S]-R[S]-Q[S]+R[S]| && \text{(B.1)}\\
&\leq \sup_{S\subseteq X} \{|P[S]-R[S]|+|R[S]-Q[S]|\} && \text{(B.2)}\\
&\leq \sup_{S\subseteq X} |P[S]-R[S]| + \sup_{S\subseteq X} |R[S]-Q[S]| && \text{(B.3)}\\
&= \delta(P,R)+\delta(R,Q). && \text{(B.4)}
\end{aligned}$$

b) Supposing the die is chosen uniformly at random, Bayes' rule (the rule for conditional probability) implies that, given the outcome of the throw, the die is more likely to be the one which has the greater probability for the observed outcome. Thus, the optimal strategy is to guess the die was the one which was more likely to yield the observed outcome of the throw. More formally, defining the event $S = \{x \in \mathscr{X} : P_X(x) \geq Q_X(x)\}$ (the results that are more likely with die P), the best strategy is report P was the actual die for all outcomes $x \in S$ and Q otherwise.

The probability that the guess is correct, again assuming the choice of die is uniform, is

$$P_{\text{correct}} = \frac{1}{2}P[S]+\frac{1}{2}Q[\bar{S}] = \frac{1}{2}(P[S]+1-Q[\mathscr{S}]) = \frac{1}{2}(1+\delta(P,Q)),$$

by definition (2.32) of the statistical distance.

c) For a finite alphabet we can write

$$\delta(P,Q) = \max_{S\subseteq X} |\sum_{x\in S} P(x)-Q(x)|. \tag{B.5}$$

Evidently, the optimal S only includes x for which $P(x) \geq Q(x)$ (or vice versa). Since $X = S \cup \bar{S}$ we have

$$\begin{aligned}
&\sum_{x\in S} P(x)-Q(x)+\sum_{x\in\bar{S}} P(x)-Q(x) = 0 \quad \text{and} && \text{(B.6)}\\
&\sum_{x\in S} P(x)-Q(x)-\sum_{x\in\bar{S}} P(x)-Q(x) = \sum_{x\in X} |P(x)-Q(x)| && \text{(B.7)}
\end{aligned}$$

and therefore the desired equality (2.33) holds.

2.2 Jensen's inequality

One can prove this inequality by induction, starting from the definition of the convex function as the base of the induction. Then for the induction step we have, given a probability distribution $\{p_1, \ldots, p_{n+1}\}$:

$$f(\sum_{k=1}^{n+1} p_k x_k) = f((1-p_{n+1})\sum_{k=1}^{n} \frac{p_k x_k}{1-p_{n+1}} + p_{n+1}x_{n+1}).$$

Now by the base of induction we have:

$$f((1-p_{n+1})\sum_{k=1}^{n} \frac{p_k x_k}{1-p_{n+1}} + p_{n+1}x_{n+1}) \leq (1-p_{n+1}) f(\sum_{k=1}^{n} \frac{p_k x_k}{1-p_{n+1}}) + p_{n+1} f(x_{n+1}).$$

By the induction assumption that the Jensen's inequality holds for the case n probabilities in the distribution (here the distribution $\{\frac{p_k}{1-p_{n+1}}\}$), we have:

$$(1-p_{n+1}) f(\sum_{k=1}^{n} \frac{p_k x_k}{1-p_{n+1}}) + p_{n+1} f(x_{n+1}) \leq (1-p_{n+1}) \sum_{k=1}^{n} \frac{p_k}{1-p_{n+1}} f(x_k) + p_{n+1} f(x_{n+1}).$$

$$\rightarrow f(\sum_{k=1}^{n+1} p_k x_k) \leq (1-p_{n+1}) \sum_{k=1}^{n} \frac{p_k}{1-p_{n+1}} f(x_k) + p_{n+1} f(x_{n+1}) = \sum_{k=1}^{n+1} p_k f(x_k).$$

2.3 Weak law of large numbers

a) Multiply each term in the expression for $P[A \geq \varepsilon]$ by $a/\varepsilon \geq 1$ to obtain

$$P[A \geq \varepsilon] = \sum_{a \geq \varepsilon} P_A(a) \leq \sum_{a \geq \varepsilon} \frac{a P_a(a)}{\varepsilon} \leq \sum_{a} \frac{a P_a(a)}{\varepsilon} = \frac{\langle A \rangle}{\varepsilon}.$$

b) Set $A = (X-\mu)^2$ and use Markov's inequality.

c) Defining $X = \frac{1}{n}\sum X_i$, the expectation value is still μ, but the variance becomes σ^2/n. Using Chebyshev's inequality we get

$$P\left[\left(\frac{1}{n}\sum_i X_i - \mu\right)^2 \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon},$$

and the weak law follows for any fixed $\varepsilon > 0$.

2.4 Conditional probabilities: Knowing more does not always help

Start by defining R to be the event that it rains and R' that the radio predicts rain. Then, $P[R \cap R']$ is the probability that it rains and the radio has predicted rain, while $P[R|R']$ is the conditional probability of rain given a prediction of rain by the radio. The problem statement amounts to the definitions $P[R] = 80\%$, $P[R|R'] = 100\%$, and $P[R \cap R'] + P[\bar{R} \cap \bar{R}'] = 80\%$.

a) The best thing your grandfather can do is to say it will rain every morning – this way he will win 80% of the time. As for you, clearly if the radio predicts rain you should, too. The question is what to do if the radio does not predict rain. But we have

$$P[R] = P[R|R']P[R'] + P[R|\bar{R}']P[\bar{R}'] = 80\% \quad \text{and} \tag{B.8}$$

$$P[R \cap R'] + P[\bar{R} \cap \bar{R}'] = P[R|R']P[R'] + P[\bar{R}|\bar{R}']P[\bar{R}'] = 80\%, \tag{B.9}$$

which implies that $P[R|\bar{R}'] = P[\bar{R}|\bar{R}']$. Thus, when the radio does not predict rain, it actually delivers no useful information, since the probability of rain is 50%. Any strategy for this case is equally good.

b) Both you and your grandfather will be correct on approximately 80% of the days – this is easy to see since one of your optimal strategies is to copy your grandfather and say it will always rain.

B.2 Exercises from Chapter 3

3.1 Hadamard gate

a) A matrix U is unitary when $U^*U = \mathbb{1}$. In fact, $H^* = H$, so we just need to verify that $H^2 = \mathbb{1}$, which is the case.

b) Since $H^2 = \mathbb{1}$, its eigenvalues must be ± 1. If both eigenvalues were equal, it would be proportional to the identity matrix. Thus, one eigenvalue is $+1$ and the other -1. By direct calculation we can find that the (normalized) eigenvectors are

$$|\lambda_\pm\rangle = \pm\frac{\sqrt{2\pm\sqrt{2}}}{2}|0\rangle + \frac{1}{\sqrt{2(2\pm\sqrt{2})}}|1\rangle \tag{B.10}$$

c) The eigenbasis of $\sigma_{\hat{x}}$ is formed by the two states $|\hat{x}_\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)$. From the form of H given in (3.56), it is clear that we can express H as

$$H = |\hat{x}_+\rangle\langle 0| + |\hat{x}_-\rangle\langle 1| \quad \text{or} \tag{B.11}$$
$$H = |0\rangle\langle\hat{x}_+| + |1\rangle\langle\hat{x}_-| \tag{B.12}$$

The latter form follows immediately from the first since $H^\dagger = H$. Finally, we can express the $\sigma_{\hat{z}}$ basis $|0/1\rangle$ in terms of the $\sigma_{\hat{x}}$ basis as $|0\rangle = \frac{1}{\sqrt{2}}(|\hat{x}_+\rangle + |\hat{x}_-\rangle)$ and $|1\rangle = \frac{1}{\sqrt{2}}(|\hat{x}_+\rangle - |\hat{x}_-\rangle)$. Thus, if we replace $|0\rangle$ and $|1\rangle$ by these expressions in the equation for H we find

$$H = |0\rangle\langle\hat{x}_+| + |1\rangle\langle\hat{x}_-| = \tfrac{1}{\sqrt{2}}\left(|\hat{x}_+\rangle\langle\hat{x}_+| + |\hat{x}_-\rangle\langle\hat{x}_+| + |\hat{x}_+\rangle\langle\hat{x}_-| - |\hat{x}_-\rangle\langle\hat{x}_-|\right). \tag{B.13}$$

Evidently, H has exactly the same representation in the $\sigma_{\hat{x}}$ basis! In retrospect, we should have anticipated this immediately once we noticed that H interchanges the $\sigma_{\hat{z}}$ and $\sigma_{\hat{x}}$ bases.

For $\sigma_{\hat{y}}$, we can proceed differently. What is the action of H on the $\sigma_{\hat{y}}$ eigenstates? These are $|\hat{y}_\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm i|1\rangle)$. Thus,

$$H|\hat{y}_\pm\rangle = \tfrac{1}{\sqrt{2}}(H|0\rangle \pm iH|1\rangle) \tag{B.14}$$
$$= \tfrac{1}{2}(|0\rangle + |1\rangle \pm i|0\rangle \mp i|1\rangle) \tag{B.15}$$
$$= \left(\frac{1\pm i}{2}\right)|0\rangle + \left(\frac{1\mp i}{2}\right)|1\rangle \tag{B.16}$$
$$= \tfrac{1}{\sqrt{2}}e^{i\pm\frac{\pi}{4}}\left(|0\rangle + \left(\frac{1\mp i}{1\pm i}\right)|1\rangle\right) \tag{B.17}$$
$$= \tfrac{1}{\sqrt{2}}e^{i\pm\frac{\pi}{4}}(|0\rangle \mp i|1\rangle) \tag{B.18}$$
$$= e^{i\pm\frac{\pi}{4}}|\hat{y}_\mp\rangle \tag{B.19}$$

Therefore, the Hadamard operation just swaps the two states in the basis (note that if we used a different phase convention for defining the $\sigma_{\hat{y}}$ eigenstates, there would be extra phase factors in this equation). So, $H = \begin{pmatrix} 0 & e^{-i\frac{\pi}{4}} \\ e^{i\frac{\pi}{4}} & 0 \end{pmatrix}$ in this basis.

d) All unitary operators on a qubit are rotations of the Bloch sphere by some angle about some axis. Since $H^2 = \mathbb{1}$, it must be a π rotation. Because the $\hat{y}$-axis is interchanged under H, the axis must lie somewhere in the $\hat{x}$-$\hat{z}$ plane. Finally, since H interchanges the $\sigma_{\hat{x}}$ and $\sigma_{\hat{z}}$ bases, it must be a rotation about the $\hat{m} = \frac{1}{\sqrt{2}}(\hat{x} + \hat{z})$ axis.

3.2 State distinguishability

a) The the probability of correctly guessing, averaged over Alice's choice of the state is

$$p_{\text{guess}} = \tfrac{1}{2}(|\langle\psi_0||\phi_0\rangle|^2 + |\langle\psi_1||\phi_1\rangle|^2) \tag{B.20}$$

To optimize the choice of measurement, suppose $|\psi_0\rangle = \alpha|0\rangle + \beta|1\rangle$ for some $\alpha, \beta \in \mathbb{C}$ such that $|\alpha|^2 + |\beta|^2 = 1$. Then $|\psi_1\rangle = -\beta^*|0\rangle + \alpha^*|1\rangle$ is orthogonal as intended. Using this in (B.20) gives

$$p_{\text{guess}} = \frac{1}{2}\left(\left|\frac{\alpha^* + \beta^*}{\sqrt{2}}\right|^2 + \left|\frac{i\alpha - \beta}{\sqrt{2}}\right|^2\right) \tag{B.21}$$

$$= \tfrac{1}{2}(1 + 2\text{Re}\left[\left(\tfrac{1-i}{2}\right)\alpha\beta^*\right]). \tag{B.22}$$

If we express α and β as $\alpha = ae^{i\theta}$ and $\beta = be^{i\eta}$ for real a, b, θ, η, then we get

$$p_{\text{guess}} = \tfrac{1}{2}(1 + 2ab\,\text{Re}\left[\left(\tfrac{1-i}{2}\right)e^{i(\theta-\eta)}\right]). \tag{B.23}$$

To maximize, we ought to choose $a = b = \frac{1}{\sqrt{2}}$, and we may also set $\eta = 0$ since only the difference $\theta - \eta$ is relevant. Now we have

$$p_{\text{guess}} = \tfrac{1}{2}(1 + \text{Re}\left[\left(\tfrac{1-i}{2}\right)e^{i\theta}\right]) \tag{B.24}$$

$$= \tfrac{1}{2}(1 + \tfrac{1}{\sqrt{2}}\text{Re}\left[e^{-i\pi/4}e^{i\theta}\right]), \tag{B.25}$$

from which it is clear that the best thing to do is to set $\theta = \pi/4$ to get $p_{\text{guess}} = \frac{1}{2}(1 + \frac{1}{\sqrt{2}}) \approx 85.4\%$. The basis states making up the measurement are $|\psi_0\rangle = \frac{1}{\sqrt{2}}(e^{i\pi/4}|0\rangle + |1\rangle)$ and $|\psi_1\rangle = \frac{1}{\sqrt{2}}(-|0\rangle + e^{-i\pi/4}|1\rangle)$.

b) The point of this exercise is to show that thinking in terms of the Bloch sphere is a lot more intuitive than just taking a brute force approach as we did in the solution of the previous exercise. Let $\hat{n}_0$ and $\hat{n}_1$ be the Bloch vectors of the two states. Call $\hat{m}$ the Bloch vector associated with one of the two basis vectors of the measurement, specifically the one which indicates that the state is $|\phi_0\rangle$ (the other is associated with $-\hat{m}$). The guessing probability takes the form

$$p_{\text{guess}} = \tfrac{1}{2}(|\langle\psi_0||\phi_0\rangle|^2 + |\langle\psi_1||\phi_1\rangle|^2) \tag{B.26}$$

$$= \tfrac{1}{2}\left(\tfrac{1}{2}(1 + \hat{n}_0 \cdot \hat{m}) + \tfrac{1}{2}(1 - \hat{n}_1 \cdot \hat{m})\right) \tag{B.27}$$

$$= \tfrac{1}{4}(2 + \hat{m} \cdot (\hat{n}_0 - \hat{n}_1)) \tag{B.28}$$

The optimal $\hat{m}$ lies along $\hat{n}_0 - \hat{n}_1$ and has unit length, i.e.

$$\hat{m} = \frac{\hat{n}_0 - \hat{n}_1}{\sqrt{(\hat{n}_0 - \hat{n}_1)\cdot(\hat{n}_0 - \hat{n}_1)}} \tag{B.29}$$

$$= \frac{\hat{n}_0 - \hat{n}_1}{\sqrt{2 - 2\cos\theta}}. \tag{B.30}$$

Therefore,

$$p_{\text{guess}} = \tfrac{1}{4}\left(2 + \sqrt{2 - 2\cos\theta}\right) \tag{B.31}$$

$$= \tfrac{1}{2}\left(1 + \sqrt{\frac{1-\cos\theta}{2}}\right) \tag{B.32}$$

$$= \tfrac{1}{2}\left(1 + \sin\tfrac{\theta}{2}\right). \tag{B.33}$$

Finally, we should check that this gives sensible results. When $\theta = 0$, $p_{\text{guess}} = \frac{1}{2}$, as it should. On the other hand, the states $|\phi_k\rangle$ are orthogonal for $\theta = \pi$, and indeed $p_{\text{guess}} = 1$ in this case. In the previous exercise we investigated the case $\theta = \frac{\pi}{2}$ and here we immediately find $p_{\text{guess}} = \frac{1}{2}(1 + \frac{1}{\sqrt{2}})$, as before.

3.3 Fidelity

a) First let's just try to guess the result. The unknown state $|\psi\rangle$ is somewhere on the Bloch sphere, and we might as well orient the sphere so that this direction is the $\hat{z}$ direction. The fidelity of $|\psi\rangle$ with any other state $|\phi\rangle$ is given by

$$|\langle\psi||\phi\rangle|^2 = \text{Tr}[P_\psi P_\phi] = \tfrac{1}{2}(1 + \cos\theta), \tag{B.34}$$

where θ is the angle between the two states on the Bloch sphere. Any state in the $\hat{x}$-$\hat{y}$ plane has a fidelity of $\frac{1}{2}$, and since a random state is as likely to lie in the upper hemisphere as in the lower, i.e. $\theta = \frac{\pi}{2} + \alpha$ and $\theta = \frac{\pi}{2} - \alpha$ are equally-likely, the average fidelity ought to be $\frac{1}{2}$. A simple integration confirms this guess:

$$\langle F\rangle = \frac{1}{4\pi}\int_0^{2\pi}\phi\int_0^{\pi}\theta\sin\theta\,\frac{1}{2}(1+\cos\theta) = \frac{1}{4}\int_0^{\pi}\theta\sin\theta = \frac{1}{2}. \tag{B.35}$$

b) Given the outcome $|k\rangle$, the fidelity is $F_k = |\langle k||\psi\rangle|^2$ and this occurs with probability $p_k = |\langle k||\psi\rangle|^2$, so averaging over the measurement outcome gives $F = \sum_k p_k F_k = \sum_k |\langle k||\psi\rangle|^4$. Now we average over $|\psi\rangle = \cos\frac{\theta}{2}|0\rangle + \sin\frac{\theta}{2}e^{i\phi}|1\rangle$:

$$\langle F\rangle = \frac{1}{4\pi}\int_0^{2\pi}\phi\int_0^{\pi}\theta\sin\theta\left(\cos^4\frac{\theta}{2} + \sin^4\frac{\theta}{2}\right) = \frac{2}{3}. \tag{B.36}$$

Thus making the measurement increases the fidelity of the guess.

3.4 Indirect measurement

Since U is unitary, it preserves inner products, and therefore

$$\langle\phi_1|\phi_2\rangle = \langle\phi_1|\phi_2\rangle\langle\text{blank}|\text{blank}\rangle = \langle\phi_1|\phi_2\rangle\langle\beta_1|\beta_2\rangle \qquad \Rightarrow \quad \langle\beta_1|\beta_2\rangle = 1$$

This means the states $|\beta_j\rangle$ are identical, and are therefore completely indistinguishable. The implication only holds if $\langle\phi_1|\phi_2\rangle \neq 0$, i.e. except when the states are orthogonal, since in that case an extra factor $\langle\beta_1|\beta_2\rangle = 0$ will not violate the equality. Of course, orthogonal states can already be distinguished by direct, nondisturbing measurement of the associated projectors $P_j = |\phi_j\rangle\langle\phi_j|$.

3.5 Broken measurement

Start with the Schmidt decomposition of $|\Psi\rangle_{AB}$:

$$|\Psi\rangle_{AB} = \sum_k \sqrt{p_k}|\alpha_k\rangle|\beta_k\rangle.$$

Bob's measurement projectors P_j can be expanded in his Schmidt basis as $P_j = \sum_{k\ell} c^j_{k\ell}|\beta_k\rangle\langle\beta_\ell|$. In order for Alice's measurement to replicate Bob's, the probabilities of the various outcomes must be identical, which is to say

$$\langle\Psi|(P_j)_B|\Psi\rangle_{AB} = \langle\Psi|(P'_j)_A|\Psi\rangle_{AB} \quad \Rightarrow \quad \sum_k p_k\langle\alpha_k|P'_j|\alpha_k\rangle = \sum_k p_k\langle\beta_k|P_j|\beta_k\rangle.$$

Thus Alice should choose $P'_j = \sum_{k\ell} c^j_{k\ell}|\alpha_k\rangle\langle\alpha_\ell|$. The post-measurement states when Alice or Bob measures are given by

$$|\Psi'_j\rangle = \sum_{k\ell}\sqrt{p_k}c^j_{k\ell}|\alpha_\ell\rangle|\beta_k\rangle \qquad \text{and} \qquad |\Psi_j\rangle = \sum_{k\ell}\sqrt{p_k}c^j_{k\ell}|\alpha_k\rangle|\beta_\ell\rangle,$$

respectively. Neither is in Schmidt form, but note that they are related by a simple swap operation $|\alpha_j\rangle_A|\beta_k\rangle_B \leftrightarrow |\alpha_k\rangle_A|\beta_j\rangle_B$, which is unitary; call it W_{AB} so that $|\Psi'_j\rangle = W|\Psi_j\rangle$. Now let $U'_j \otimes V'_j$ be unitary operators which transform $|\Psi_j\rangle$ to Schmidt form in the $|\alpha_j\rangle|\beta_k\rangle$ basis. That is, $(U'_j \otimes V'_j)|\Psi_j\rangle = \sum_k \sqrt{p^j_k}|\alpha_k\rangle|\beta_k\rangle$, and it follows that $W(U'_j\otimes V'_j)|\Psi_j\rangle = (U'_j\otimes V'_j)|\Psi_j\rangle$. Therefore $V'_j\otimes U'_j$ takes $|\Psi'_j\rangle$ to Schmidt form:

$$(V'_j \otimes U'_j)|\Psi'_j\rangle = W W^\dagger(V'_j \otimes U'_j)W|\Psi_j\rangle = W(U'_j \otimes V'_j)|\Psi_j\rangle = \sum_k \sqrt{p^j_k}|\alpha_k\rangle|\beta_k\rangle,$$

and thus

$$\begin{aligned}
& (U'_j \otimes V'_j)|\Psi_j\rangle = (V'_j \otimes U'_j)|\Psi'_j\rangle \\
\Rightarrow \quad & (U'_j \otimes V'_j)(\mathbb{1} \otimes P_j)|\Psi\rangle = (V'_j \otimes U'_j)(P'_j \otimes \mathbb{1})|\Psi\rangle \\
\Rightarrow \quad & (\mathbb{1} \otimes P_j)|\Psi\rangle = (U'^\dagger_j V'_j \otimes V'^\dagger_j U'_j)(P'_j \otimes \mathbb{1})|\Psi\rangle.
\end{aligned}$$

3.6 Remote copy

Suppose Alice copies $|\psi\rangle_A$ to her half of the maximally entangled state $|\Phi\rangle_{A'B}$ using the CNOT gate $U_{Cnot}|j,k\rangle = |j, j\oplus k\rangle$. This results in

$$\begin{aligned}
U^{AA'}_{CNOT}|\psi\rangle_A|\Phi\rangle_{A'B} &= \tfrac{1}{\sqrt{2}}(a|000\rangle + a|011\rangle + b|110\rangle + b|101\rangle)_{AA'B} \\
&= \tfrac{1}{\sqrt{2}}[(a|00\rangle + b|11\rangle)_{AB}|0\rangle_{A'} + (a|01\rangle + b|10\rangle)_{AB}|1\rangle_{A'}] \\
&= \tfrac{1}{\sqrt{2}}(|\Psi\rangle_{AB}|0\rangle_{A'} + (\sigma_x)_B|\Psi\rangle_{AB}|1\rangle_{A'}).
\end{aligned}$$

As in teleportation, this creates the desired output state, up to the action of a Pauli operator on Bob's system which is indexed by an orthogonal state at Alice's end. By measuring system A' and telling Bob the result (using just one bit since there are only two outcomes) he can undo the Pauli operator to create $|\Psi\rangle_{AB}$.

3.7 Measurements on a bipartite state

a) In accordance with the postulates, Alice describes B by the postmeasurement state as in (3.2). Computing this we find, for any θ,

$$ {}_A\langle\theta|\Phi\rangle_{AB} = \tfrac{1}{\sqrt{2}}(\cos\theta\langle 0| + \sin\theta\langle 1|)(|00\rangle + |11\rangle) \tag{B.37} $$

$$ = \tfrac{1}{\sqrt{2}}(\cos\theta|0\rangle + \sin\theta\langle 1|) = \tfrac{1}{\sqrt{2}}|\theta\rangle_B. \tag{B.38} $$

Thus, she describes Bob's state as $|\theta\rangle$ or $|\frac{\pi}{2} - \theta\rangle$ depending on the result of her measurement, either of which is equally-likely as the other.

b) The calculation from the previous part shows that the outcomes of such a measurement on $|\Phi\rangle_{AB}$ are both equally-likely, no matter the value of θ. So, the probability distribution of his outcomes should be uniform, as Alice's mere implementation of the measurement should not affect any observable quantity at his end.

But we must check this is consistent with the postulates. Conditioned on Alice's measurement result, the state of B is either $|\theta\rangle$ or $|\frac{\pi}{2} - \theta\rangle$, but Bob does not know which, so he must average over the two possibilities. Alice, who knows the result, does not need to average. The probability of obtaining $|0\rangle$ in his measurement, given the state $|\theta\rangle$, is simply $\cos^2\theta$. The average probability is then $\frac{1}{2}\cos^2\theta + \frac{1}{2}\cos^2(\frac{\pi}{2} - \theta) = \frac{1}{2}$, as hoped.

3.8 The Hilbert-Schmidt inner product

a) The orthonormal bases $\left\{|i\rangle_{R/Q}\right\}_i$ are arbitrary but fixed. We can expand any operator A as $A = \sum_{kl} a_{kl}|k\rangle\langle l|$, and its transpose as $A^T = \sum_{kl} a_{kl}|l\rangle\langle k|$. We then have

$$ A \otimes \mathbb{1}|\Omega\rangle = \sum_{kli}\Big(a_{kl}\,|k\rangle\langle l| \otimes \mathbb{1}\Big)|i\rangle\,|i\rangle = \sum_{kli} a_{kl}\,\delta_{li}\,|k\rangle|i\rangle = \sum_{kl} a_{kl}\,|k\rangle|l\rangle $$

Similarly,

$$ \mathbb{1} \otimes A^T|\Omega\rangle = \sum_{kli}\Big(\mathbb{1} \otimes a_{kl}\,|l\rangle\langle k|\Big)|i\rangle|i\rangle = \sum_{kli} a_{kl}\,\delta_{ki}\,|i\rangle|l\rangle = \sum_{kl} a_{kl}\,|k\rangle|l\rangle. $$

b) From the first part it follows that

$$ A \otimes B|\Omega\rangle = (\mathbb{1} \otimes B)(A \otimes \mathbb{1})|\Omega\rangle = (\mathbb{1} \otimes B)(\mathbb{1} \otimes A^T)|\Omega\rangle = (\mathbb{1} \otimes BA^T)|\Omega\rangle, $$

so we can write

$$ \langle\Omega|A \otimes B|\Omega\rangle = \langle\Omega|\mathbb{1} \otimes BA^T|\Omega\rangle. $$

Now we use the structure of $|\Omega\rangle$ to write

$$ \sum_{ij}\langle i|\langle i|\mathbb{1} \otimes BA^T|j\rangle|j\rangle = \delta_{ij}\langle i|BA^T|k\rangle = \mathrm{Tr}BA^T = \mathrm{Tr}A^TB. $$

3.9 Teleportation redux

a)

$$\begin{aligned}\left(U_A \otimes \overline{U}_B\right)|\Phi\rangle_{AB} &= \frac{1}{\sqrt{2}} \sum_{jk\ell mt} U_{jk}\overline{U}_{\ell m}(|j\rangle\langle k|_A \otimes |\ell\rangle\langle m|_B)|t,t\rangle_{AB} \\ &= \frac{1}{\sqrt{2}} \sum_{j\ell t} U_{jt}\overline{U}_{\ell t}|j,\ell\rangle_{AB} = \frac{1}{\sqrt{2}} \sum_{j\ell}|j,\ell\rangle_{AB} \sum_t U_{jt}(U^*)_{t\ell} \\ &= \frac{1}{\sqrt{2}} \sum_{j\ell}|j,\ell\rangle_{AB}[UU^*]_{j\ell} = \frac{1}{\sqrt{2}} \sum_j |j,j\rangle_{AB} = |\Phi\rangle_{AB}\end{aligned}$$

b)

$$_A\langle\psi|\Phi\rangle_{AB} = \frac{1}{\sqrt{2}} \sum_{jk} \psi^*_{jA}\langle j|k\rangle_A|k\rangle_B = \frac{1}{\sqrt{2}} \sum_k \psi^*_k|k\rangle_B = \frac{1}{\sqrt{2}}|\psi^*\rangle_B \tag{B.39}$$

c) In the first case we have

$$\begin{aligned}|\psi'_j\rangle_B &= {}_{A'A}\langle\Phi_j|(|\psi\rangle_{A'}U_B|\Phi\rangle_{AB}) = {}_{A'A}\langle\Phi|\left(\mathbb{1}_{A'} \otimes (\sigma_j^\dagger)_A \otimes U_B\right)|\psi\rangle_{A'}|\Phi\rangle_{AB} \\ &= {}_{A'A}\langle\Phi|\left(\mathbb{1}_{A'} \otimes \mathbb{1}_A \otimes (U\sigma_j^*)_B\right)|\psi\rangle_{A'}|\Phi\rangle_{AB} = \tfrac{1}{2}(U\sigma_j^*)_B|\psi\rangle_B.\end{aligned}$$

After Bob receives Alice's message and applies σ_j^T they end up with the state $|\psi''_j\rangle = (\sigma_j^T U\sigma_j^*)|\psi\rangle$. For the second case

$$|\psi'_j\rangle_B = {}_{A'A}\langle\Phi_j|U_{A'}(|\psi\rangle_{A'}U_B|\Phi\rangle_{AB}) = {}_{A'A}\langle\Phi_j|(|U\psi\rangle_{A'}U_B|\Phi\rangle_{AB}) = \tfrac{1}{2}(\sigma_j^*U)_B|\psi\rangle_B.$$

Now Bob's correction operation produces $|\psi''\rangle = U|\psi\rangle$. This is an important result, because it shows that it is possible to perform an arbitrary single-qubit operation solely by measuring an appropriately prepared state.

d) Work with the Schmidt decomposition: $|\psi\rangle_{A_1A_2} = \sum_k \sqrt{p_k}|\alpha_k\rangle_{A_1}|\beta_k\rangle_{A_2}$. Then following the same calculation above we get

$$\begin{aligned}|\psi'_j\rangle_{A_1B} &= {}_{A_2A}\langle\Phi_j|\left(|\psi\rangle_{A_1A_2}|\Phi\rangle_{AB}\right) = \sum_k \sqrt{p_k}\ {}_{A_2A}\langle\Phi_j|(|\alpha_k\rangle_{A_1}|\beta_k\rangle_{A_2}|\Phi\rangle_{AB}) \\ &= \sum_k \sqrt{p_k}|\alpha_k\rangle_{A_1}\ {}_{A_2A}\langle\Phi_j|(|\beta_k\rangle_{A_2}|\Phi\rangle_{AB}) = \tfrac{1}{2}\sum_k \sqrt{p_k}|\alpha_k\rangle_{A_1}(\sigma_j^*)_B|\beta_k\rangle_B \\ &= \tfrac{1}{2}(\sigma_j^*)_B|\psi\rangle_{A_1B}.\end{aligned}$$

Once again Bob can undo the σ_j^* on system B and thus teleportation can also faithfully transfer part of a larger, entangled system.

3.10 "All-or-nothing" violation of local realism

a) Observe that the three operators commute, since X and Y anticommute. Since the state is invariant under permutations of the three systems, we only need to check that it is an eigenstate of the first operator, since the others are generated from it by permutation. Both X and Y flip bits in the standard basis, but Y adds an extra $-\mathrm{i}$ if the input is $|0\rangle$ and i if $|1\rangle$. Thus $XYY|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}(-\mathrm{i})^2|111\rangle - (\mathrm{i}^2)|000\rangle = |\mathrm{GHZ}\rangle$.

b) Measuring Y on any two systems determines the X value on the third, so absent any "spooky action at a distance", the X value should be well-defined. Similarly, measurements of X and Y on any two determine the Y value of the third, so it should also be well-defined. For X measurements on each spin, the product $x_1x_2x_3 = 1$ since $x_1x_2x_3 = (x_1y_2y_3)(y_1x_2y_3)(y_1y_2x_3)$ (if x_j and y_k all take the values ± 1.)

c) Measuring X on each system and taking the product is the same as measuring $X_1X_2X_3$. $|\mathrm{GHZ}\rangle$ is clearly an eigenstate of this operator with eigenvalue -1, so $X_1X_2X_3 = -1$.

B.3 Exercises from Chapter 4

4.1 The Bloch ball

a) We have

$$\rho_1 = \tfrac{1}{2}(\mathbb{1} + \tfrac{1}{2}\sigma_x) = \frac{1}{4}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

whose eigenvalues are $\{1/4, 3/4\}$, corresponding to the vectors $|+\rangle$ and $|-\rangle$, respectively.

Meanwhile,

$$\rho_2 = \tfrac{1}{2}(\mathbb{1} + \tfrac{1}{\sqrt{2}}(\sigma_x + \sigma_z)) = \frac{1}{2\sqrt{2}}\begin{pmatrix} 1+\sqrt{2} & 1 \\ 1 & \sqrt{2}-1 \end{pmatrix}.$$

Observe that $\frac{1}{\sqrt{2}}(\sigma_x + \sigma_z)$ is in fact the Hadamard gate H from Exercise 3.1, whose eigenvalues are ± 1. Thus, the eigenvalues of ρ_2 are $\{0, 1\}$, corresponding to the eigenvectors of the H.

b) The Pauli matrices are Hermitian and the vector $\vec{r}$ is real, therefore ρ is Hermitian. As the Pauli operators are traceless, the unit trace condition is also immediately satisfied. To show positivity, we must compute the eigenvalues. Note that for a 2×2 Hermitian matrix, the trace is the sum of the eigenvalues, while the determinant is their product. Hence, $\lambda_+ + \lambda_- = 1$ and $\lambda_+\lambda_- = \frac{1}{2}\det\begin{pmatrix} 1+r_z & r_x - i\,r_y \\ r_x + i\,r_y & 1-r_z \end{pmatrix}$. We can conclude that

$$\lambda_\pm = \frac{1}{2}(1 + |\vec{r}|),$$

so that the condition $|\vec{r}| \leq 1$ ensures positivity of ρ.

c) First observe that any 2×2 Hermitian operator can be expressed as a linear combination of the Pauli operators and identity operator with real coefficients. For normalization to hold, the coefficient of the identity operator must be $1/2$. All that remains is positivity, which as we have seen, implies that $|\vec{r}| \leq 1$.

d) The surface of the ball is defined by $|\vec{r}| = 1$, which leads to valid density operators with eigenvalues $\{0, 1\}$, i.e. all pure states.

4.2 Partial trace

a) Let us express ρ_{AB} in terms of bases for A and B as

$$\rho_{AB} = \sum_{jklm} C_{jl;km} |j\rangle\langle k|_A \otimes |l\rangle\langle m|_B.$$

Then $\rho_A = \sum_{jkl} C_{jl;kl} |j\rangle\langle k|_A$. Clearly ρ_A is Hermitian since from Hermiticity of ρ_{AB} is must hold that $\overline{C}_{km;jk} = C_{jl;km}$. For positivity of ρ_A, consider an arbitrary $|\psi\rangle_A$ and its tensor product with any one of the basis states $|l\rangle_B$. Then, from positivity of ρ_{AB} we have $\langle\psi| \otimes \langle l|\rho|psi\rangle \otimes |l\rangle \geq 0$ and therefore

$$\begin{aligned} 0 &\leq \sum_l \langle\psi| \otimes \langle l|\rho|\psi\rangle \otimes |l\rangle \\ &= \sum_l (\langle\psi| \otimes \langle l|)(\sum_{jkl'm} C_{jl';km} |j\rangle\langle k| \otimes |l'\rangle\langle m|)(|\psi\rangle \otimes |l\rangle) \\ &= \sum_{jkl} C_{jl;kl} \langle\psi|j\rangle\langle k|\psi\rangle \\ &= \langle\psi|\rho_A|\psi\rangle. \end{aligned}$$

The normalization condition of ρ_A is simply the same as that of ρ_{AB}.

b) For ρ_A we have

$$\begin{aligned} \rho_A &= \mathrm{Tr}_B[|\Psi\rangle\langle\Psi|_{AB}] \\ &= \sum_{jkj'k'} C_{jk} \bar{C}_{j'k'} |j\rangle\langle j'|\delta_{kk'} \\ &= \sum_{jj'} |j\rangle\langle j'| \sum_k C_{jk} \bar{C}_{j'k} \\ &= \sum_{jj'} |j\rangle\langle j'|[CC^\dagger]_{jj'} \end{aligned}$$

For ρ_B the calculation is entirely analogous.

c) For each one we obtain the maximally-mixed state $\frac{1}{2}\mathbb{1}_A$.

d) Clearly $P_X(x) = 1/2$. We can represent P_{XY} by the state

$$\rho_{XY} = \tfrac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|).$$

The partial trace is then just the maximally-mixed state $\frac{1}{2}\mathbb{1}$.

4.3 Canonical purifications

a) Tracing out B, we obtain

$$\mathrm{Tr}_B[|\psi\rangle\langle\psi|_{AB}] = \sqrt{\rho}\,\mathrm{Tr}_B[U_B|\Omega\rangle\langle\Omega|U_B^*]\sqrt{\rho} = \sqrt{\rho}\mathbb{1}\sqrt{\rho} = \rho.$$

b) By Proposition 4.1.1, any two purifications of ρ_A are related by unitaries (or isometries) on B. Since applying another unitary to B gives a state of the same form, all purifications must have this form.

4.4 Decompositions of density matrices

a) By Proposition 4.1.2 we have $\sqrt{p_\ell}|\phi_\ell\rangle = \sum_k \sqrt{\lambda_k} U_{k\ell}|k\rangle$ for some unitary matrix $U_{k\ell}$. Taking the norm of each expression results in

$$p_\ell = \sum_k \lambda_k |U_{k\ell}|^2$$

since $|k\rangle$ is an orthonormal basis. Thus $\vec{\lambda}$ majorizes $\vec{p}$. Note that we cannot turn this argument around to say that $\vec{p}$ majorizes λ. Since starting from $\sqrt{\lambda_k}|k\rangle = \sum_\ell \sqrt{p_\ell} U^\dagger_{k\ell}|\phi_\ell\rangle$ we cannot easily compute the norm of the righthand side because the $|\phi_k\rangle$ are not orthogonal.

b) $\vec{u}$ is majorized by every other distribution $\vec{p}$ (of length less or equal to n) since we can use the doubly stochastic matrix $T_{jk} = 1/n$ for all j,k to produce $\vec{u} = T\vec{p}$. Therefore, to find a decomposition in which all the weights are identical, we need to find a unitary matrix whose entries all have the same magnitude, namely $1/\sqrt{n}$. One choice that exists in every dimension is the Fourier transform $F_{jk} = \frac{1}{\sqrt{n}}\omega^{jk}$, where $\omega = \exp(2\pi i/n)$. The vectors in the decomposition are therefore

$$|\phi_\ell\rangle = \sum_k \sqrt{\lambda_k}\omega^{k\ell}|k\rangle.$$

4.5 Generalized measurement by direct (tensor) product

a) Name the output states $|\phi_{00}\rangle_{AB}$ and $|\phi_{01}\rangle_{AB}$, respectively. Although the specification of U is not complete, we have the pieces we need, and we can write $U_{AB} = \sum_{jk}|\phi_{jk}\rangle\langle jk|$ for some states $|\phi_{10}\rangle$ and $|\phi_{11}\rangle$. The measurement operators A_k are defined implicitly by

$$U_{AB}|\psi\rangle_A|0\rangle_B = \sum_k (A_k)_A|\psi\rangle_A|k\rangle_B.$$

Thus $A_k = {}_B\langle k|U_{AB}|0\rangle_B = \sum_j {}_B\langle k|\phi_{j0}\rangle_{AB}\langle j|_A$, which is an operator on system A, even though it might not look like it at first glance. We then find

$$A_0 = \frac{2}{\sqrt{6}}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_1 = \frac{1}{\sqrt{6}}\begin{pmatrix} \sqrt{3} & 1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \frac{1}{\sqrt{6}}\begin{pmatrix} \sqrt{3} & -1 \\ 0 & 0 \end{pmatrix}.$$

b) The corresponding POVM elements are given by $E_j = A^* A_j$:

$$E_0 = \frac{2}{3}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad E_1 = \frac{1}{6}\begin{pmatrix} 3 & \sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}, \quad E_2 = \frac{1}{6}\begin{pmatrix} 3 & -\sqrt{3} \\ -\sqrt{3} & 1 \end{pmatrix}.$$

They are each rank one (which can be verified by calculating the determinant). The POVM elements project onto trine states $|1\rangle, (\sqrt{3}|0\rangle \pm |1\rangle)/2$.

c) The averaged post-measurement state is given by $\rho' = \sum_j A_j \rho A^*$. In this case we have $\rho' = \mathrm{diag}(2/3, 1/3)$.

4.6 Geometry of POVMs

a) We expand E in its eigenbasis and write

$$\begin{aligned} E &= \lambda\, |0\rangle\langle 0| + \sum_{i\neq 0} \lambda_i\, |i\rangle\langle i| \\ &= \lambda\, |0\rangle\langle 0| + (1-\lambda_0) \sum_{i\neq 0} \lambda_i\, |i\rangle\langle i| + \lambda \sum_{i\neq 0} \lambda_i\, |i\rangle\langle i| \\ &= \lambda \underbrace{\left(|0\rangle\langle 0| + \sum_{i\neq 0} \lambda_i\, |i\rangle\langle i|\right)}_{E_1} + (1-\lambda) \underbrace{\sum_{i\neq 0} \lambda_i\, |i\rangle\langle i|}_{E_2}. \end{aligned}$$

Hence we can write the POVM $\{E, \mathbb{1} - E\}$ as a convex combination of the POVMs $\{E_1, \mathbb{1} - E_1\}$ and $\{E_2, \mathbb{1} - E_2\}$.

b) Let E be an orthogonal projector on some subspace $V \in \mathscr{H}$ and let $|\psi\rangle \in V^\perp$. If we assume that E can be written as the convex combination of two positive operators then

$$\begin{aligned} 0 &= \langle\psi|E|\psi\rangle \\ &= \lambda\langle\psi|E_1|\psi\rangle + (1-\lambda)\langle\psi|E_2|\psi\rangle. \end{aligned}$$

However, both terms on the right hand side are non-negative, thus they must vanish identically. Since $|\psi\rangle$ was arbitrary we conclude that $E_1 = E_2 = E$.

c) The element-wise convex combination of elements an be interpreted as using two different measurement devices with probability α and $1-\alpha$, but not knowing which measurement device was used. In contrast to that, a simple convex concatenation of sets would be interpreted as using two different measurement devices with probability α and $1-\alpha$, but keeping track of which measurement device was used. This is because by definition of a POVM, each POVM element corresponds to a specific measurement outcome. If the two POVMs are concatenated, we can still uniquely relate the measurement outcome to the corresponding measurement device.

4.7 Some common quantum channels

Dephasing The dephased output is the same as measuring the state in the standard basis: $\mathrm{diag}(\rho_{00}, \rho_{11}) = \sum_{j=0}^{1} P_j \rho P_j$ for $P_j = |j\rangle\langle j|$. Thus possible Kraus operators are $A_2 = \sqrt{1-p}\,\mathbb{1}$, $A_j = \sqrt{p} P_j$, $j = 0, 1$. But we can find a representation with fewer Kraus operators. Notice that $\sigma_z \rho \sigma_z = \begin{pmatrix} \rho_{00} & -\rho_{01} \\ -\rho_{10} & \rho_{11} \end{pmatrix}$. Thus $(\rho + \sigma_z \rho \sigma_z)/2 = \mathrm{diag}(\rho_{00}, \rho_{11})$ and $\rho' = \sum_{j=0}^{1} A_j \rho A_j^*$ for $A_0 = \sqrt{1-p/2}\,\mathbb{1}$ and $A_1 = \sqrt{p/2}\,\sigma_z$.

Two is the minimal number of Kraus operators, since for one Kraus operator the trace preserving condition becomes $A^*A = \mathbb{1}$ and implies $A = \mathbb{1}$ for qubit-to-qubit channels. The action of the dephasing channel is to shrink the $x-y$ components of the Bloch vector. We have found examples of both projective and unitary Kraus operators.

Depolarizing Since the action of conjugation by σ_z destroys off-diagonal elements in the basis in which σ_z is diagonal, we could try separately conjugating by each Pauli operator and see what happens. The result is that $\mathcal{E}(\rho) = \sum_{j=1}^{4} A_j \rho A_j^*$ for $A_0 = \sqrt{1-3p/2}\mathbb{1}$, $A_k = \sqrt{p/2}\sigma_k$, $k = 1,2,3$.

The number of Kraus operators is at least the rank of the Choi state. Applying the depolarizing channel to half of a maximally entangled state $|\Phi\rangle_{AB}$ results in the mixture

$$\rho'_{AB} = (1-p)|\Phi\rangle\langle\Phi|_{AB} + p\,\tfrac{1}{4}\mathbb{1}_{AB}. \tag{B.40}$$

This state has rank four unless $p = 0$. The action of the depolarizing channel is to shrink all components of the Bloch vector. The Kraus operators we have found are all (proportional to) unitaries.

Amplitude damping Finally, for the amplitude damping channel, we can read off the Kraus operators from the unitary action since $U|\psi\rangle|0\rangle = \sum_k A_k|\psi\rangle|k\rangle$. Therefore $A_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}$ and $A_1 = \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix}$.

The action of the amplitude damping channel is to map the Bloch vector towards the north pole of the Bloch sphere (or whichever direction is associated with the state $|0\rangle$.)

4.8 Classical channels as CPTP maps

a) We have

$$P_Y(0) = \sum_x P_X(x) P_{Y|X=x}(0) = q(1-p) + (1-q)p$$
$$P_Y(1) = qp + (1-q)(1-p),$$

which can be expressed as a quantum state $\rho_y = [q(1-p)+(1-q)p]\,|0\rangle\langle 0| + [qp+(1-q)(1-p)]\,|1\rangle\langle 1| \in \mathcal{L}(\mathcal{H}_Y)$.

b) We take four operators, corresponding to the four different "branches" of the channel,

$$M_{0\to 0} = \sqrt{1-p}|0\rangle\langle 0|$$
$$M_{0\to 1} = \sqrt{p}|1\rangle\langle 0|$$
$$M_{1\to 0} = \sqrt{p}|0\rangle\langle 1|$$
$$M_{1\to 1} = \sqrt{1-p}|1\rangle\langle 1|.$$

To check that this works for the classical state ρ_X, we compute

$$\begin{aligned}
\mathcal{E}(\rho_X) &= \sum_{xy} M_{x\to y}\,\rho_X\,M_{x\to y}^* \\
&= \sum_{xy} M_{x\to y}\left[q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|\right] M_{x\to y}^* \\
&= (1-p)\,|0\rangle\langle 0|\Big[q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|\Big]|0\rangle\langle 0| + p\,|1\rangle\langle 0|\Big[q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|\Big]|0\rangle\langle 1| \\
&\quad + p\,|0\rangle\langle 1|\Big[q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|\Big]|1\rangle\langle 0| + (1-p)\,|1\rangle\langle 1|\Big[q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|\Big]|1\rangle\langle 1| \\
&= q(1-p)\,|0\rangle\langle 0| + qp\,|1\rangle\langle 1| + (1-q)p\,|0\rangle\langle 0| + (1-q)(1-p)\,|1\rangle\langle 1|.
\end{aligned}$$

As intended, $\mathcal{E}(\rho_X) = \rho_Y$.

c) In general, we can express the state in the computational basis as $\rho_X = \sum_{ij} \alpha_{ij} |i\rangle\langle j|$, with the usual conditions (positivity, normalization). Applying the map gives us

$$\begin{aligned}
\mathcal{E}(\rho_X) =& \sum_{xy} M_{x\to y} \Big[\sum_{ij} \alpha_{ij} |i\rangle\langle i|j \Big] M^*_{x\to y} \\
=&(1-p)\,|0\rangle\langle 0| \Big[\sum_{ij} \alpha_{ij} |i\rangle\langle i|j \Big] |0\rangle\langle 0| + p\,|1\rangle\langle 0| \Big[\sum_{ij} \alpha_{ij} |i\rangle\langle i|j \Big] |0\rangle\langle 1| \\
&+ p\,|0\rangle\langle 1| \Big[\sum_{ij} \alpha_{ij} |i\rangle\langle i|j \Big] |1\rangle\langle 0| + (1-p)\,|1\rangle\langle 1| \Big[\sum_{ij} \alpha_{ij} |i\rangle\langle i|j \Big] |1\rangle\langle 1| \\
=&\alpha_{11}(1-p)\,|0\rangle\langle 0| + \alpha_{11} p\,|1\rangle\langle 1| + \alpha_{22} p\,|0\rangle\langle 0| + \alpha_{22}(1-p)\,|1\rangle\langle 1|.
\end{aligned}$$

Using $\alpha_{11} := \alpha, \alpha_{22} = 1-\alpha$, we get $\mathcal{E}(\rho_X) = [\alpha(1-p) + (1-\alpha)p]\,|0\rangle\langle 0| + [\alpha p + (1-\alpha)(1-p)]\,|1\rangle\langle 1|$. The channel ignores the off-diagonal terms of ρ_X: it acts as a measurement on the computational basis followed by the classical binary symmetric channel.

d) We generalize the previous result as

$$\begin{aligned}
\mathcal{E}_W(\rho_X) =& \sum_{x,y} P_{Y|X=x}(y)\,|y\rangle\langle x|\rho_X|x\rangle\langle y| \\
=& \sum_{x,y} E_{x\to y} \rho_X E^* x \to y, \quad E_{x\to y} = \sqrt{P_{Y|X=x}(y)}\,|y\rangle\langle x|.
\end{aligned}$$

To see that this works, take a classical state $\rho_X = \sum_x P_X(x)\,|x\rangle\langle x|$ as input,

$$\begin{aligned}
\mathcal{E}_W(\rho_X) =& \sum_{x,y} P_{Y|X=x}(y)\,|y\rangle\langle x| \Big(\sum_{x'} P_X(x')\,|x'\rangle\langle x'| \Big) |x\rangle\langle y| \\
=& \sum_{x,y} P_{Y|X=x}(y)\,P_X(x)\,|y\rangle\langle y| \\
=& \sum_{y} P_y(y)\,|y\rangle\langle y|.
\end{aligned}$$

4.9 Unital channels

Let $B_k = V^* A_k U$. Then it's easy to verify that $\Lambda' = \sum_k B_k \Lambda B_k^*$ and that $\sum_k B_k B_k^* = \sum_k B_k^* B_k = \mathbb{1}$. Now consider the component form of each of these equations:

$$\begin{aligned}
(\Lambda')_\ell &= \sum_{k,n} (B_k)_{\ell n} (B_k^*)_{n\ell} (\Lambda_n), \\
\delta_{\ell m} &= \sum_{k,n} (B_k)_{\ell n} (B_k^*)_{nm}, \\
\delta_{\ell m} &= \sum_{k,n} (B_k^*)_{\ell n} (B_k)_{nm}.
\end{aligned}$$

We only need one index for Λ and Λ' since they are diagonal. Defining $D_{\ell n} = \sum_k (B_k)_{\ell n} (B_k^*)_{n\ell}$, we have $\Lambda' = D\Lambda$ (thinking of Λ and Λ' as vectors), and the two conditions on the B_k imply that D is doubly stochastic.

4.10 The Choi isomorphism

a) First define the components of the Bloch representation by

$$\rho = \frac{1}{2}\begin{pmatrix} 1+z & x-iy \\ x+iy & 1-z \end{pmatrix}, \qquad x^2+y^2+z^2 \leq 1.$$

We apply the map to this state and get

$$\rho' = \frac{1}{2}\begin{pmatrix} 1+2\alpha x & 2\alpha z \\ 2\alpha z & 1-2\alpha x \end{pmatrix}.$$

The mapping is trace-preserving, hence it is positive if and only if the determinant of ρ' is positive for all allowed values of x, y and z. The determinant is given by

$$\begin{aligned} \det(\rho') &= \frac{1}{4}(1-4\alpha^2x^2-4\alpha^2z^2) \\ &\geq \frac{1}{4}-\alpha^2. \end{aligned}$$

The map is positive for $0 \leq \alpha \leq \frac{1}{2}$.

b) Using the Bell basis we have

$$(\mathscr{E}_\alpha \otimes \mathscr{I})[|\Phi\rangle\langle\Phi|] = \tfrac{1}{4}\mathbb{1}_{AB} + \alpha(|\Phi_x\rangle\langle\Phi_z| + |\Phi_z\rangle\langle\Phi_x|).$$

We could expand this in a local basis $|j\rangle|k\rangle$ and compute the eigenvalues to determine positivity. But we can just as well use the Bell basis, since the state is already partially expressed in this basis. The identity operator is the equal mixture of all the Bell state projectors, and so in the basis $\{|\Phi_x\rangle, |\Phi_z\rangle, |\Phi\rangle, |\Phi_y\rangle\}$ we have

$$(\mathscr{E}_\alpha \otimes \mathscr{I})[|\Phi\rangle\langle\Phi|] \simeq \frac{1}{4}\begin{pmatrix} 1 & 4\alpha & 0 & 0 \\ 4\alpha & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For complete positivity the determiniant of the output must be nonnegative, which we can easily see is only true for $0 \leq \alpha \leq \frac{1}{4}$.

c) To construct Kraus operators, we use (4.57). At $\alpha = 1/4$ there are only three eigenvectors of the Choi state with nonzero eigenvalues, $|\Phi\rangle$, $|\Phi_y\rangle$, and $\frac{1}{\sqrt{2}}(|\Phi_x\rangle + |\Phi_z\rangle) = \frac{1}{2}(|00\rangle + |01\rangle + |10\rangle - |11\rangle)$, corresponding to eigenvalues 1/4, 1/4, and 1/2, respectively. For $|\Phi\rangle$ we have the Kraus operator M_0 such that

$$M_0|\phi\rangle = \tfrac{\sqrt{2}}{2}{}_A\langle\bar{\phi}|\Phi\rangle = \tfrac{1}{2}|\phi\rangle,$$

using the results of Exercise 3.9. Hence, $M_0 = \frac{1}{2}\mathbb{1}$.

Similarly, for $|\Phi_y\rangle$ we get M_1 with

$$M_1|\phi\rangle = \tfrac{\sqrt{2}}{2}{}_A\langle\bar{\phi}|\Phi_y\rangle = \tfrac{\sqrt{2}}{2}{}_A\langle\bar{\phi}|\sigma_y|\Phi\rangle = \tfrac{1}{2}\sigma_y|\phi\rangle,$$

and thus $M_1 = \frac{1}{2}\sigma_y$. Since the last eigenvector is a superposition of $|\Phi_x\rangle$ and $|\Phi_z\rangle$, the same reasoning implies that $M_2 = \frac{1}{2}(\sigma_x + \sigma_z)$.

B.4 Exercises from Chapter 5

5.1 Minimum-error state discrimination

a) No matter how many outcomes the physical apparatus produces, Bob will invariably group them into two sets corresponding to the guess of ρ_j. Hence without loss of generality we can just regard the POVM whose elements are the sums of the two groups as a new POVM.

b) The probability of error is just

$$p_{\text{error}} = p_1 \text{Tr}[\rho_1 E_2] + p_2 \text{Tr}[\rho_2 E_1] = p_1 \text{Tr}[\rho_1] + \text{Tr}[(p_2\rho_2 - p_1\rho_1)E_1].$$

Using the eigenvalues and eigenvectors of $p_2\rho_2 - p_1\rho_1$ yields the desired result.

c) The quantities $\langle e_i|E_1|e_i\rangle$ are nonnegative, but the λ can be negative. The best choice is to choose E_1 to project onto the subspace corresponding to negative eigenvalues: $E_1 = \sum_{i:\lambda_i<0} |e_i\rangle\langle e_i|$.

d) Let $B = p_2\rho_2 - p_1\rho_1$. Then $\text{Tr}[B] = \sum_{i:\lambda_i\geq 0} \lambda_i + \sum_{i:\lambda_i<0} \lambda_i$, while $||B||_1 = \sum_{i:\lambda_i\geq 0} \lambda_i - \sum_{i:\lambda_i<0} \lambda_i$. Thus $\sum_{i:\lambda_i<0} \lambda_i = (\text{Tr}[B] - ||B||_1)/2$. Substitution into the equation for p_{error} gives the desired result.

5.2 Unambiguous state discrimination

a) Since the two signal states are pure, they span a two-dimensional subspace and without loss of generality we can restrict the support of the POVM elements to this subspace—an effective qubit. Suppose E_j are rank-one operators $E_j = \alpha_k |\xi_j\rangle\langle\xi_j|$ (if they aren't, decompose them into a set of rank-one operators). Then we want to fulfill $0 = \Pr(E_j|\rho_k) = \alpha_k |\langle\xi_j|\phi_k\rangle|^2$, which can only work if $|\xi_j\rangle = |\phi_k^\perp\rangle$. That is, $|\xi_0\rangle$ is the state orthogonal to $|\phi_1\rangle$ and *vice versa*; the unambiguous measurement works by rejecting rather than confirming one of the two hypotheses. Thus $E_j = \alpha_j |\phi_k^\perp\rangle\langle\phi_k^\perp|$ for $j \neq k$ and some $0 \leq \alpha_k \leq 1$.

b) Since $\langle\phi_1|\phi_2\rangle \neq 0$ in general, $\sum_{j=1}^2 E_j \neq \mathbb{1}$, and therefore a third measurement element is needed. This outcome tells Bob nothing about which signal was sent, so it is an inconclusive result $E_?$.

c) We know that a general unambiguous discrimination POVM has the form

$$E_0 = \alpha_0 |\phi_1^\perp\rangle\langle\phi_1^\perp|, \qquad E_1 = \alpha_1 |\phi_0^\perp\rangle\langle\phi_0^\perp|, \qquad E_? = \mathbb{1} - E_0 - E_1.$$

The sum-to-unity constraint is enforced by the form of $E_?$ and $E_{0/1}$ are positive by construction, so the only outstanding constraint is that $E_?$ be positive. Symmetry between the signal states implies that $\alpha_0 = \alpha_1$, leaving

$$\mathbb{1} - \alpha(|\phi_0^\perp\rangle\langle\phi_0^\perp| + |\phi_1^\perp\rangle\langle\phi_1^\perp|) \geq 0.$$

Thus we should choose the largest value of α consistent with this constraint. We can find a closed-form expression in terms of Bloch-sphere quantities. Let $|\phi_j\rangle$ have Bloch vector $\hat{n}_j$, meaning $|\phi_j^\perp\rangle$ has Bloch vector $-\hat{n}_j$. Then the constraint becomes

$$\mathbb{1} - \tfrac{1}{2}\alpha(\mathbb{1} - \hat{n}_1 \cdot \vec{\sigma} + \mathbb{1} - \hat{n}_0 \cdot \vec{\sigma}) = (1-\alpha)\mathbb{1} + \alpha(\hat{n}_0 + \hat{n}_1) \cdot \vec{\sigma} \geq 0.$$

We know the eigenvalues of a general expression in terms of the Pauli operators and identity from the lecture on qubits, namely $\lambda_\pm = (1-\alpha) \pm \alpha|\hat{n}_0 + \hat{n}_1|$. Thus, the largest possible α is

$$\alpha = \frac{1}{1+|\hat{n}_0 + \hat{n}_1|}.$$

When the $|\phi_j\rangle$ are orthogonal, $\hat{n}_0 + \hat{n}_1 = 0$ and the unambiguous measurement goes over into the usual projection measurement.

5.3 Decoupling

a) Clearly $|\Phi\rangle_{AA'}$ is a purification of the maximally mixed state $\frac{\mathbb{1}_A}{d_A}$. The partial trace over A' and B' of the product state $|\Phi\rangle_{AA'} \otimes |\psi\rangle_{BB'}$ will just produce the tensor product of the partial trace of $|\Phi\rangle_{AA'}$ and that of $|\psi\rangle_{BB'}$ over B', i.e. the state ρ_{AB}.

b) From (5.40) we have $F(\sigma_{AB}, \frac{\mathbb{1}_A}{d_A} \otimes \rho_B) \geq 1-\varepsilon$. Then, using the properties of fidelity there exists a purification $|\phi\rangle_{ABA'B'}$ of σ_{AB} such that $F(|\phi\rangle_{ABA'B'}, |\Psi\rangle_{AA'}) = F(\sigma_{AB}, \frac{\mathbb{1}_A}{d_A} \otimes \rho_B)$. Using (5.41) yields the desired result for this particular purification.

5.4 Entanglement and channel distinguishability

a) This follows immediately from the definition, since the set of input states considered in the maximization for δ_{1-1} is a subset of those for δ itself.

b) First we calculate $\delta_{1-1}(\mathcal{E}_p, \mathcal{I})$:

$$\begin{aligned}\delta_{1-1}(\mathcal{E}_p, \mathcal{I}) &= \max_{\rho_A} \delta(p\tfrac{1}{2}\mathbb{1}_A + (1-p)\rho_A, \rho_A) \\ &= \max_{\rho_A} \tfrac{1}{2}\mathrm{Tr}\left|p\tfrac{1}{2}\mathbb{1} - p\rho_A\right| \\ &= \max_{\rho_A} \tfrac{p}{2}\mathrm{Tr}\left|\tfrac{1}{2}\mathbb{1} - \rho_A\right|.\end{aligned}$$

Note that $\mathrm{Tr}|\mathbb{1}/2 - \rho|$ is the distance between the Bloch vector of ρ and the center of the Bloch sphere. This implies that its maximum occurs when ρ_A is a pure state, and so the radius is 1. Therefore $\delta_{1-1}(\mathcal{E}_p, \mathcal{I}) = \frac{1}{2}p$.

To compute or bound δ, first note that the channel can be thought of as leaving the input state untouched with probability $1-p$ or else taking the trace of the state and then creating the new state $\frac{1}{2}\mathbb{1}$ with probability p. Therefore,

$$\begin{aligned}\delta(\mathcal{E}, \mathcal{I}) &= \max_{\rho_{AR}} \delta(\mathcal{E} \otimes \mathcal{I}(\rho_{AR}), \mathcal{I} \otimes \mathcal{I}(\rho_{AR})) \\ &= \max_{\rho_{AR}} \tfrac{1}{2}\mathrm{Tr}|p\tfrac{1}{2}\mathbb{1}_A \otimes \rho_R + (1-p)\rho_{AR} - \rho_{AR}| \\ &= \frac{p}{2} \max_{\rho_{AR}} \mathrm{Tr}|\tfrac{1}{2}\mathbb{1}_A \otimes \rho_R - \rho_{AR}|.\end{aligned}$$

We can find a lower bound by choosing $\rho_{AR} = |\Phi\rangle\langle\Phi|_{AR}$. In that case $\rho_R = \frac{1}{2}\mathbb{1}$ and thus

$$\begin{aligned}\delta(\mathcal{E},\mathcal{I}) &\geq \tfrac{p}{2}\mathrm{Tr}|\tfrac{1}{4}\mathbb{1}_{AR} - |\Phi\rangle\langle\Phi|_{AR}| \\ &= \frac{p}{2}(3\times\tfrac{1}{4} - (-\tfrac{3}{4})) = \tfrac{3}{4}p.\end{aligned}$$

Thus we have found a channel for which $\delta(\mathcal{E},\mathcal{F}) > \delta_{1-1}(\mathcal{E},\mathcal{F})$, i.e. the general inequality relating the two quantities is strict.

B.5 Exercises from Chapter 6

6.1 Properties of the von Neumann entropy

The first two properties are clear by inspection. The third follows from positivity of the relative entropy by taking $\sigma = \frac{1}{d}\mathbb{1}$, where $\mathbb{1}$ is the identity operator on $\mathrm{supp}\rho$ and $d = |\mathrm{supp}\rho|$: Then, $D(\rho_A,\sigma_A) = \log d - H(A)_\rho \geq 0$.

To prove the fourth let $\rho_A = \sum_k p_k(\rho_k)_A$. Then $\sum_k p_k D(\rho_k,\rho) = H(A)_\rho - \sum_k p_k H(A)_{\rho_k} \geq 0$. Since the relative entropy is only zero when the arguments are identical, at least on the support of the first argument, for equality to hold it must be the case that restricting ρ to the support of ρ_k gives back ρ_k itself. Thus all the states must be disjoint for concavity to hold with equality.

Finally, let $\rho' = \sum_k P_k \rho P_k$. Observe that $[P_k,\rho'] = 0$. Thus,

$$D(\rho_A,\rho'_A) = -H(A)_\rho - \sum_k \mathrm{Tr}[P_k(\rho_A\log\rho'_A)P_k] = H(A)_{\rho'} - H(A)_\rho \geq 0.$$

6.2 Optimization in the conditional von Neumann entropy

Expanding $D(\rho_{AB},\mathbb{1}_A\otimes\sigma_B)$ we have

$$\begin{aligned}D(\rho_{AB},\mathbb{1}_A\otimes\sigma_B) &= \mathrm{Tr}[\rho_{AB}\log\rho_{AB}] - \mathrm{Tr}[\rho_{AB}\log(\mathbb{1}_A\otimes\sigma_B)] \\ &= \mathrm{Tr}[\rho_{AB}\log\rho_{AB}] - \mathrm{Tr}[\rho_{AB}(\mathbb{1}_A\otimes\log\sigma_B)] \\ &= \mathrm{Tr}[\rho_{AB}\log\rho_{AB}] - \mathrm{Tr}[\rho_B\log\sigma_B] \\ &= \mathrm{Tr}[\rho_{AB}\log\rho_{AB}] - \mathrm{Tr}[\rho_B\log\sigma_B] + \mathrm{Tr}[\rho_B\log\rho_B] - \mathrm{Tr}[\rho_B\log\rho_B] \\ &= D(\rho_{AB},\rho_A\otimes\rho_B) + D(\rho_B,\sigma_B).\end{aligned}$$

Thus, $\max_\sigma -D(\rho_{AB},\rho_A\otimes\sigma_B) = -D(\rho_{AB},\rho_A\otimes\rho_B) + \max_\sigma -D(\rho_B,\sigma_B)$. The maximum of the last expression is clearly zero, which holds for $\sigma_B = \rho_B$.

6.3 Quantum mutual information

a) From the chain rule for mutual information we have $I(A:B) = H(A) + H(B) - H(AB) \leq H(A) + H(B) \leq 2$ for bipartite qubit systems. The Bell state achieves this bound.

b) Write $I(A:B) = H(A) - H(A|B)$ and use the bound $H(A|B) \geq 0$ since AB is a CQ state.

c) The AB subsystem is just a classically-correlated state, as is the ABC system, which accounts for the first two expressions. However, $ABCD$ is not a classical state and so $I(A:B|CD) = H(CD|A) - H(CD|AB) = -H(CD|AB) = 1$.

6.4 Data processing for classical mutual information

a) First observe that

$$\frac{P_{XY|Z=z}(x,y)}{P_{Y|Z=z}(y)} = \frac{P_{XYZ}(x,y,z)}{P_{YZ}(y,z)} = P_{X|Y=y,Z=z}(x),$$

which implies $I(X{:}Y|Z) = H(X|Z) - H(X|YZ)$. Then

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

b) There are only two ways to expand this expression:

$$I(X{:}YZ) = I(X{:}Z) + I(X{:}Y|Z) = I(X{:}Y) + I(X{:}Z|Y).$$

Since X and Z are conditionally independent given Y, $I(X{:}Z|Y) = 0$. Meanwhile, $I(X{:}Y|Z) \geq 0$, since it is a mixture (over Z) of positive quantities $I(X{:}Y|Z = z)$. Therefore $I(X{:}Y) \geq I(X{:}Z)$.

6.5 Fano's inequality

a) The random variables X, Y, and $\widehat{X}$ form a Markov chain, so we can use the data processing inequality. It leads directly to $H(X|\widehat{X}) \geq H(X|Y)$.

b) $H(E|X,\widehat{X}) = 0$ since E is determined from X and $\widehat{X}$. $H(E|\widehat{X}) \leq H(E) = h_2(P_{\text{error}})$ since conditioning reduces entropy.

$$\begin{aligned} H(X|E,\widehat{X}) &= H(X|E=0,\widehat{X})p(E=0) + H(X|E=1,\widehat{X})p(E=1) \\ &= 0(1-P_{\text{error}}) + H(X|E=1,\widehat{X})P_{\text{error}} \leq P_{\text{error}} \log|X| \end{aligned}$$

Putting this together we have

$$H(X|Y) \leq H(X|\widehat{X}) \leq h_2(P_{\text{error}}) + P_{\text{error}} \log|X| \leq 1 + P_{\text{error}} \log|X|,$$

where the last inequality follows since $h_2(x) \leq 1$. Rearranging terms gives the Fano inequality.

6.6 Squashed entanglement of separable states

Any separable state can be written $\rho_{AB} = \sum_k p_k (\sigma_k)_A \otimes (\gamma_k)_B$, which we can extend to $\rho_{ABE} = \sum_k p_k (\sigma_k)_A \otimes (\gamma_k)_B \otimes |k\rangle\langle k|_E$. This is a CQ state with E classical, so we can interpret $I(A:B|E)$ as the mixture of $I(A:B)_k$ over the values of k in E. But each of these terms is zero since AB is a product state given the value of k.

6.7 A sufficient entanglement criterion

a) Applying $\mathscr{F} \otimes \mathscr{I}$ to each state in the mixture defining the separable state results in a valid product state since $\mathscr{F}$ is positive. Mixtures of positive states are positive, so the output is positive.

b) Clearly the identity remains unchanged by the operation. We found the partial transpose of $|\Phi\rangle$ in (4.47) so the output state is

$$\rho' = \frac{1}{4}\begin{pmatrix} 2-\varepsilon & 0 & 0 & 0 \\ 0 & \varepsilon & 2-2\varepsilon & 0 \\ 0 & 2-2\varepsilon & \varepsilon & 0 \\ 0 & 0 & 0 & 2-\varepsilon \end{pmatrix}.$$

Clearly there is a doubly-degenerate eigenvalue $(2-\varepsilon)/4$ and for the other two we only need to look at the 2×2 middle block. Clearly its eigenvectors are $|\pm\rangle$, with eigenvalues $(2-\varepsilon)/4$ and $3\varepsilon - 2$, respectively. The latter is negative for $\varepsilon > 2/3$, and therefore we can conclude that the state is certainly not separable for $2/3 < \varepsilon \leq 1$.

B.6 Exercises from Chapter 7

7.1 Optimality of superdense coding

Suppose X and Y are perfectly correlated, but random 2-bit random variables. Then, $I(X:Y) = 2$. Suppose now that Y is transmitted through the putative superdense coding scheme using a units of entanglement and b units of quantum communication. Before the decoding step, the receiver has a system B consisting of $a+b$ qubits. Supposing the protocol works as intended, the Holevo bound, Corollary 7.3.1 implies $a + b \geq 2$.

7.2 Classical resource inequalities

a) Since E is independent of A and B by the security condition, $I(A:B|E) = I(A:B)$. Since A and B are perfectly correlated but each uniformly distributed, $I(A:B|E) = l$.

b) After any putative LOCC operation, we would have $I(A':B'|E') \leq l$. From the previous part we know that if A' and B' make a length-l' secret key, then $I(A':B'|E') = l'$, whence it follows that $l' \leq l$.

c) By the chain rule we have $(A,X:B|E) = H(AX|E) - H(AX|BE) = H(A|E) + H(X|AE) - H(A|BE) - H(X|ABE) = I(A:B|E) + I(X:B|AE)$. But since X is independent of everything else, the second term vanishes.

d) This follows by monotonicity.

e) The proof is entirely similar to the monotonicity of squashed entanglement, Proposition 6.6.1.

7.3 Classical channel capacities

For any classical random variable Y we have

$$I(X:Y) = H(Y) - H(Y|X) = H(Y) - \sum_x P_X(x) H(Y|X=x),$$

where $H(Y|X=x) = H(Y)_{P_{Y|X=x}}$. The BSC and BEC are symmetric channels, in the sense that $H(Y|X=x)$ is independent of x. That is, the output has the same entropy conditioned on the input, for any input. In the case of the BSC the entropy equals $h_2(p) = -p\log p - (1-p)\log(1-p)$, often called the *binary entropy*. This also holds for the BEC. Therefore, the capacity of either channel is

$$\max_{P_X} H(Y) - h_2(p),$$

the only difference being how Y is related to X. To determine the optimal distribution for either case, we can explicitly optimize, or note the fact that P_Y is a convex mixture of distributions and use the concavity of entropy. That is, $P_Y = \sum_x P_X(x) P_{Y|X=x}$ and therefore the optimal input distribution must be $P_X \sim$uniform. This reasoning holds for any symmetric channel in the sense defined above, but still does not tell us the value of the capacity.

But we need only evaluate $H(Y)$ for X uniform. Under the BSC, a uniform input is mapped to a uniform output, so $H(Y) = 1$ and the capacity is $1 - h_2(p)$. Under the BEC, the uniform input

is mapped to the distribution with values $((1-p)/2, p, (1-p)/2)$. Computing the entropy we find $H(Y) = h_2(p) + (1-p)$, and thus the capacity is simply $1-p$.

7.4 Classical capacity of the depolarizing channel

a) These are nearly the same as in the BSC: $P_{Y|X=x}(x) = 1 - p + p/2 = 1 - p/2$.

b) By the previous exercise we know $q = 1/2$ is optimal and the capacity is $1 - h_2(p/2)$.

c)

B.7 Exercises from Chapter 8

8.1 One-time pad

First note that

$$\begin{aligned} I(C:M) - I(C:M|K) &= I(M:K) - I(M:K|C) \\ &= I(K:C) - I(K:C|M), \end{aligned}$$

and that mutual information is non-negative. We introduce $x = I(C:M|K)$, $y = I(M:K|C)$ and $z = I(K:C|M)$ and, using $I(C:M) = 0$, we get

$$x - I(C;M) = x = y - I(M;K) = z - I(K;C). \tag{B.41}$$

Using the two conditions, we write

$$\begin{aligned} H(M) &= H(M|C,K) + I(C:M) + I(K:M|C) = y, \quad \text{and} \\ H(K) &= H(K|M,C) + I(M:K) + I(M:C|K) \geq y - x + z. \end{aligned}$$

However, since $y \geq x$ and $z \geq x$ (from (B.41)), we get $H(K) \geq H(M)$.

8.2 Secrecy and correctness

Let $\tilde{\rho}_{ABE}$ be the tripartite state held be the distinguisher after interacting with the ideal system $\sigma_E \mathscr{K}$ for an optimal simulator σ_E, and let Γ be the positive operator which projects the AB system on all states with $A \neq B$. Then

$$\varepsilon \geq \delta(\rho_{ABE}, \tilde{\rho}_{ABE}) \geq \delta(\rho_{AB}, \tilde{\rho}_{AB}) \geq \mathrm{Tr}[\Gamma(\rho_{AB} - \tilde{\rho}_{AB})] = \Pr[A \neq B]_\rho.$$

The last equality holds because by construction of the ideal key resource $\mathscr{K}$, $\mathrm{Tr}(\Gamma \tilde{\rho}_{AB}) = 0$ (for any simulator σ_E).

Let $p_{\text{key}} \rho_{AE}^{\text{key}}$ be the state of the real AE system held by the distinguisher after projecting on the subspace in which a key is generated, and let $\tilde{p}_{\text{key}} \tau_A \otimes \tilde{\rho}_E^{\text{key}}$ be the state of the ideal AE system for the same projection. Note that we cannot assume that $p_{\text{key}} = \tilde{p}_{\text{key}}$ or $\rho_E^{\text{key}} = \tilde{\rho}_E^{\text{key}}$, since we do not know how the simulator σ_E works.

Since

$$\varepsilon \geq \delta(\rho_{ABE}, \tilde{\rho}_{ABE}) \geq \delta(p_{\text{key}} \rho_{AE}^{\text{key}}, \tilde{p}_{\text{key}} \tau_A \otimes \tilde{\rho}_E^{\text{key}}) \geq \delta(p_{\text{key}} \rho_E^{\text{key}}, \tilde{p}_{\text{key}} \tilde{\rho}_E^{\text{key}}),$$

we have

$$\begin{aligned} p_{\text{key}}\delta\left(\rho_{AE}^{\text{key}}, \tau_A \otimes \rho_E^{\text{key}}\right) &= \delta\left(p_{\text{key}}\rho_{AE}^{\text{key}}, p_{\text{key}}\tau_A \otimes \rho_E^{\text{key}}\right) \\ &\leq \delta\left(p_{\text{key}}\rho_{AE}^{\text{key}}, \tilde{p}_{\text{key}}\tau_A \otimes \tilde{\rho}_E^{\text{key}}\right) + \delta\left(\tilde{p}_{\text{key}}\tau_A \otimes \tilde{\rho}_E^{\text{key}}, p_{\text{key}}\tau_A \otimes \rho_E^{\text{key}}\right) \\ &\leq \varepsilon + \varepsilon. \end{aligned}$$

8.3 Min-entropy chain rule

Let $\rho_{XZE} = \sum_{x,z} p_{x,z}|x\rangle\langle x| \otimes |z\rangle\langle z| \otimes \rho_E^{x,z}$ and let $\{\Gamma_x^z\}_x$ be the optimal measurement of the E system to guess x given that $Z = z$. A possible strategy for guessing XZ given E is to pick z uniformly at random then apply the measurement $\{\Gamma_x^z\}_x$ to the E system. This strategy would succeed with probability

$$\sum_{x,z} p_{x,z} \frac{1}{|\mathscr{Z}|} \text{Tr}(\Gamma_x^z \rho_E^{x,z}) = \frac{1}{|\mathscr{Z}|} p_{\text{guess}}(X|ZE)_\rho.$$

We thus have

$$p_{\text{guess}}(X|E)_\rho \geq p_{\text{guess}}(XZ|E)_\rho \geq \frac{1}{|\mathscr{Z}|} p_{\text{guess}}(X|ZE)_\rho,$$

hence

$$H_{\min}(X|ZE)_\rho \geq H_{\min}(X|E)_\rho - \log|\mathscr{Z}|.$$

To prove the smooth version, let $\bar{\rho}_{XE} \in \mathscr{B}^\varepsilon(\rho_{XE})$ be the state which maximizes $H_{\min}(X|E)_{\bar{\rho}}$. Let $\bar{\rho}_{XZE}$ be an extension of $\bar{\rho}_{XE}$ such that $P(\rho_{XZE}, \bar{\rho}_{XZE}) = P(\rho_{XE}, \bar{\rho}_{XE})$. By the property of the purified distance, such a state is guaranteed to exist. Then

$$H_{\min}^\varepsilon(X|ZE)_\rho \geq H_{\min}(X|ZE)_{\bar{\rho}} \geq H_{\min}(X|E)_{\bar{\rho}} - \log|\mathscr{Z}| = H_{\min}^\varepsilon(X|E)_\rho - \log|\mathscr{Z}|.$$

8.4 Privacy amplification with smooth min-entropy

Let $\bar{\rho}_{XE} \in \mathscr{B}^{\bar{\varepsilon}}(\rho_{XE})$ be the state which maximizes $H_{\min}(X|E)_{\bar{\rho}}$. Then

$$\delta\left(\bar{\rho}_{F(X,Y)YE}, \tau_U \otimes \tau_Y \otimes \bar{\rho}_E\right) \leq \varepsilon.$$

Furthermore,

$$\begin{aligned} \delta(\rho_{F(X,Y)YE}, \bar{\rho}_{F(X,Y)YE}) \leq \delta(\rho_{XE} \otimes \tau_Y, \bar{\rho}_{XE} \otimes \tau_Y) \leq P(\rho_{XE}, \bar{\rho}_{XE}), \\ \delta(\tau_U \otimes \tau_Y \otimes \rho_E, \tau_U \otimes \tau_Y \otimes \bar{\rho}_E) \leq P(\rho_{XE}, \bar{\rho}_{XE}). \end{aligned}$$

The result follows from two uses of the triangle inequality.

8.5 Quantum one-time pad

Consider the following encoding of 2 bit messages in the Bell states:

$$|\beta_{00}\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}, \qquad |\beta_{01}\rangle = \frac{|01\rangle + |10\rangle}{\sqrt{2}},$$
$$|\beta_{10}\rangle = \frac{|00\rangle - |11\rangle}{\sqrt{2}}, \qquad |\beta_{11}\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}},$$

We then encrypt the first qubit of the Bell pair with a (reversible) scheme satisfying (8.14). The resulting cipher is $\rho_C = \tau_A \otimes \tau_B$, regardless of the original message. This is a perfect encryption of a two bit message, satisfying $I(M:C) = 0$ and $H(M|CK) = 0$. In Exercise 11.1 we proved that any

scheme satisfying these two conditions must have $H(K) > H(M)$, hence the key must be at least 2 bits.

8.6 Bit commitment

a) Obviously $\rho^{b=0} = \rho^{b=1} = \frac{1}{2}\mathbb{1}$ for this algorithm. Thus, Bob has no information about b.

b) If Bob measures in the wrong basis, he hits the correct X_i with probability 1/2. So the probability that he gets the whole string right is given by 2^{-n}.

c) Alice prepares a singlet state

$$|\Psi_4\rangle = \tfrac{1}{\sqrt{2}}(|01\rangle - |10\rangle) = \tfrac{1}{\sqrt{2}}(|+-\rangle - |-+\rangle)$$

and sends half of it to Bob. He gets the state $\mathbb{1}/2$ as he is supposed to. However, when revealing, Alice first measures the qubit she kept in the B_b basis and send the negated results to Bob. His measurements will then always agree with the data he got and he does not know Alice cheated.

8.7 Data hiding

a) Both Π^S and Π^A are projectors (they have the form $\sum_i |\phi_i\rangle\langle\phi_i|$, for orthonormal $\{|\phi_i\rangle\}_i$), so $\rho^{b=0}$ and $\rho^{b=1}$ are Hermitian and positive semi-definite. As for normalization, we have

$$\rho^{b=0} = \frac{2}{d(d+1)}\left(\overbrace{\sum_i}^{d \text{ terms}} |ii\rangle\langle ii| + \frac{1}{2}\overbrace{\sum_j\sum_{i<j}}^{\frac{d(d-1)}{2} \text{ terms}} |ij\rangle\langle ij| + |ji\rangle\langle ji| + |ij\rangle\langle ij|ji + |ji\rangle\langle ji|ij\right)$$

$$\mathrm{Tr}(\rho^{b=0}) = \frac{2}{d(d+1)}\left(d + \frac{1}{2}\left[\frac{d(d-1)}{2} + \frac{d(d-1)}{2}\right]\right) = 1;$$

and

$$\rho^{b=1} = \frac{2}{d(d-1)}\left(\frac{1}{2}\overbrace{\sum_j\sum_{i<j}}^{\frac{d(d-1)}{2} \text{ terms}} |ij\rangle\langle ij| + |ji\rangle\langle ji| - |ij\rangle\langle ij|ji - |ji\rangle\langle ji|ij\right)$$

$$\mathrm{Tr}(\rho^{b=1}) = \frac{2}{d(d-1)}\left(\frac{1}{2}\left[\frac{d(d-1)}{2} + \frac{d(d-1)}{2}\right]\right) = 1.$$

If we had access to both systems, we could perform the global measurement described by the POVM $\left\{\Pi^S, \Pi^A, \mathbb{1} - \Pi^S - \Pi^A\right\}$. The probabilities of the three possible outcomes are $(1,0,0)$ if the state is $\rho^{b=10}$ and $(0,1,0)$ if the state is $\rho^{b=1}$, so we could recover the value of b with certainty.

b) We expand the operators in the basis of the flip operator,

$$M = \sum_{i,j} x_{ij}|i\rangle\langle i|j, \qquad N = \sum_{k,\ell} y_{k\ell}|k\rangle\langle k|\ell.$$

Applying the flip operator, we have

$$F(M_A \otimes N_B) = \sum_{i,j'} |i'j'\rangle\langle i'j'|j'i'\left(\sum_{i,j,k,\ell} x_{ij}\, y_{k\ell}\, |ik\rangle\langle ik|j\ell\right)$$
$$= \sum_{i,j,k,\ell} x_{ij}\, y_{k\ell}\, |ki\rangle\langle ki|j\ell.$$

Now we take the trace,

$$\mathrm{Tr}[F(M_A \otimes N_B)] = \sum_{i',j'} \langle i',j'|\left(\sum_{i,j,k,\ell} x_{ij}\, y_{k\ell}\, |ki\rangle\langle ki|j\ell\right)|i',j'\rangle$$
$$= \sum_{i,j} x_{ij} y_{ji}.$$

On the other hand,

$$M_A N_B = \left(\sum_{i,j} x_{ij}|i\rangle\langle i|j\right)\left(\sum_{k,\ell} y_{k\ell}|k\rangle\langle k|\ell\right) = \sum_{i,j,\ell} x_{ij} y_{j\ell}|i\rangle\langle i|\ell,$$

$$\mathrm{Tr}(M_A N_B) = \sum_{i',j'} \langle i'|\left(\sum_{i,j,\ell} x_{ij} y_{j\ell}|i\rangle\langle i|\ell\right)|i'\rangle = \sum_{i,j} x_{ij} y_{ji},$$

which proves our claim. In the particular case of pure states, $M = |x\rangle\langle x|, N = |y\rangle\langle y|$, we can take the trace using an on. basis $\{|x_i\rangle\}_i$, such that $|x_0\rangle = |x\rangle$,

$$\mathrm{Tr}(MN) = \sum_i \langle x_i|x\rangle\langle x|y\rangle\langle y|x_i\rangle = \langle x|y\rangle\langle y|x\rangle = |\langle x|y\rangle|^2.$$

c)

$$\begin{aligned}
\delta(P_{XY}, Q_{XY}) &= \sum_{x,y\in\mathcal{S}} P_{XY}(x,y) - Q_{XY}(x,y) \\
&= \sum_{x,y\in\mathcal{S}} \mathrm{Tr}(|xy\rangle\langle xy|\rho^{b=0}) - \mathrm{Tr}(|xy\rangle\langle xy|\rho^{b=1}) \\
&= \sum_{x,y\in\mathcal{S}} \mathrm{Tr}(|xy\rangle\langle xy|[\rho^{b=0} - \rho^{b=1}]) \\
&= \sum_{x,y\in\mathcal{S}} \mathrm{Tr}\left(|xy\rangle\langle xy|\left[\frac{2}{d(d+1)}\Pi^S - \frac{2}{d(d-1)}\Pi^A\right]\right) \qquad \text{Note: } \frac{2}{d(d-1)} = \frac{2}{d(d+1)} + \\
&= \frac{2}{d(d+1)} \sum_{x,y\in\mathcal{S}} \mathrm{Tr}\left(|xy\rangle\langle xy|\left[\Pi^S - \Pi^A\right]\right) - \frac{4}{d(d-1)(d+1)} \sum_{x,y\in\mathcal{S}} \mathrm{Tr}(|xy\rangle\langle xy|\Pi^A) \\
&\leq \frac{2}{d(d+1)} \sum_{x,y\in\mathcal{S}} \mathrm{Tr}\,(F|xy\rangle\langle xy|) \qquad \text{Because } \Pi^A \text{projector } \Rightarrow 0 \leq \mathrm{Tr}(|xy\rangle\langle xy|\Pi^A) \leq 1 \\
&\leq \frac{2}{d(d+1)} \sum_{x}^{d}\sum_{y}^{d} |\langle x|y\rangle|^2 \qquad \text{Note: } |x\rangle = \sum_{y}^{d} \langle y|x\rangle|y\rangle \Rightarrow |\langle x|x\rangle|^2 = \sum_{y} |\langle y|x\rangle|^2 \\
&= \frac{2}{d(d+1)} \sum_{x}^{d} \langle x|x\rangle^2 = \frac{2}{d+1}.
\end{aligned}$$

Bibliography

[1] R. Landauer, "Information is physical", Physics Today **44**, 23 (1991).

[2] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (Cambridge University Press, Sept. 2000).

[3] E. G. Rieffel and W. H. Polak, *Quantum computing: a gentle introduction*, 1st ed. (The MIT Press, Mar. 2011).

[4] S. M. Barnett, *Quantum information*, Oxford master series in physics. Atomic, optical and laser physics 16 (Oxford University Press, Oxford, 2009).

[5] B. Schumacher and M. Westmoreland, *Quantum processes systems, and information* (Cambridge University Press, Apr. 2010).

[6] M. Hayashi, *Quantum information theory: an introduction* (Springer, June 2006).

[7] A. S. Holevo, *Probabilistic and statistical aspects of quantum theory (publications of the scuola normale superiore / monographs*, 1st Edition. (Edizioni della Normale, June 2011).

[8] A. S. Holevo, *Quantum systems, channels, information a mathematical introduction*, English (De Gruyter, Berlin, 2012).

[9] A. Peres, *Quantum theory: concepts and methods*, Vol. 72, Fundamental Theories of Physics (Kluwer Academic Publishers, New York, 2002).

[10] M. Wilde, *Quantum information theory* (Cambridge University Press, Cambridge, UK. ; New York, 2013).

[11] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. (Wiley-Interscience, July 2006).

[12] D. J. C. MacKay, *Information theory, inference & learning algorithms*, 1st (Cambridge University Press, June 2002).

[13] N. D. Mermin, *Quantum computer science: an introduction*, 1st ed. (Cambridge University Press, Sept. 2007).

[14] S. Aaronson, *Quantum computing since democritus* (Cambridge University Press, 2013).

[15] D. H. Mellor, *Probability: a philosophical introduction*, English (Routledge, London; New York, 2005).

[16] E. T. Jaynes, *Probability theory: the logic of science* (Cambridge University Press, June 2003).

[17] S. M. Ross, *A first course in probability*, English (Pearson Prentice Hall, Upper Saddle River, N.J., 2010).

[18] A. Gut, *An intermediate course in probability*, 2nd ed, Springer texts in statistics (Springer, Dordrecht : New York, NY, 2009).

[19] R. Durrett, *Probability : theory and examples*, English (Cambridge University Press, Cambridge; New York, 2010).

[20] B. Fristedt and L. Gray, *A modern approach to probability theory*, en (Birkhäuser Boston, Boston, MA, 1997).

[21] A. S. Davis, "Markov chains as random input automata", The American Mathematical Monthly **68**, 264–267 (1961).

[22] R. P. Feynman, R. B. Leighton, and M. L. Sands, *The feynman lectures on physics, vol. 3: quantum mechanics*, English (Addison-Wesley Pub. Co., Reading, Mass., 1963).

[23] A. Einstein, B. Podolsky, and N. Rosen, "Can quantum-mechanical description of physical reality be considered complete?", Physical Review **47**, 777 (1935).

[24] E. Schrödinger, "Discussion of probability relations between separated systems", Mathematical Proceedings of the Cambridge Philosophical Society **31**, 555–563 (1935).

[25] A. Einstein, *Letter to schrödinger*, German, EAC 22-47, June 1935.

[26] D. Howard, "Einstein on locality and separability", Studies in History and Philosophy of Science Part A **16**, 171–201 (1985).

[27] A. Peres, "Unperformed experiments have no results", American Journal of Physics **46**, 745 (1978).

[28] P. A. M. Dirac, *The principles of quantum mechanics*, 4th ed. (Oxford University Press, USA, 1967).

[29] J. v. Neumann, *Mathematical foundations of quantum mechanics*, Investigations in Physics 2 (Princeton University Press, 1955).

[30] L. E. Ballentine, *Quantum mechanics: a modern development* (World Scientific Publishing Company, Mar. 1998).

[31] B. C. Hall, *Quantum theory for mathematicians* (Springer, New York, 2013).

[32] L. A. Takhtajan, *Quantum mechanics for mathematicians*, Graduate studies in mathematics v. 95 (American Mathematical Society, Providence, R.I, 2008).

[33] D. M. Greenberger, K. Hentschel, and F. Weinert, eds., *Compendium of quantum physics: concepts, experiments, history, and philosophy* (Springer, Heidelberg ; New York, 2009).

[34] R. I. G. Hughes, *The structure and interpretation of quantum mechanics* (Harvard University Press, Cambridge, Mass, 1989).

[35] G. Jaeger, *Entanglement, information, and the interpretation of quantum mechanics*, The frontiers collection (Springer, Berlin, 2009).

[36] C. G. Timpson, *Quantum information theory and the foundations of quantum mechanics*, 1st ed, Oxford philosophical monographs (Clarendon Press, Oxford, 2013).

[37] B.-G. Englert, "Fringe visibility and which-way information: an inequality", Physical Review Letters **77**, 2154 (1996).

[38] J. Preskill, "Lecture notes for phys 229", 2004.

[39] M. A. Naimark, "Spectral functions of a symmetric operator", Russian, Izvestiya Akademii Nauk SSSR. Seriya Matematicheskaya **4**, 277–318 (1940).

[40] W. F. Stinespring, "Positive functions on c*-algebras", Proceedings of the American Mathematical Society **6**, 211–216 (1955).

[41] K.-E. Hellwig and K. Kraus, "Pure operations and measurements", Communications in Mathematical Physics **11**, 214–220 (1969).

[42] K.-E. Hellwig and K. Kraus, "Operations and measurements. II", Communications in Mathematical Physics **16**, 142–147 (1970).

[43] A. Jamiołkowski, "Linear transformations which preserve trace and positive semidefiniteness of operators", Reports on Mathematical Physics **3**, 275–278 (1972).

[44] J. de Pillis, "Linear transformations which preserve hermitian and positive semidefinite operators", Pacific Journal of Mathematics **23**, 129–137 (1967).

[45] M.-D. Choi, "Completely positive linear maps on complex matrices", Linear Algebra and its Applications **10**, 285–290 (1975).

[46] E. B. Davies, *Quantum theory of open systems* (Academic Press, London, 1976).

[47] K. Kraus, *States, effects, and operations: fundamental notions of quantum theory*, Lecture notes in physics 190 (Springer-Verlag, Berlin, 1983).

[48] C. W. Helstrom, "Detection theory and quantum mechanics", Information and Control **10**, 254–291 (1967).

[49] C. W. Helstrom, *Quantum detection and estimation theory*, Vol. 123, Mathematics in Science and Engineering (Academic, London, 1976).

[50] D. Bures, "An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras", Transactions of the American Mathematical Society **135**, 199–212 (1969).

[51] A. Uhlmann, "The "transition probability" in the state space of a *-algebra", Reports on Mathematical Physics **9**, 273–279 (1976).

[52] C. A. Fuchs, "Distinguishability and accessible information in quantum theory", PhD thesis (University of New Mexico, Jan. 1996).

[53] A. Y. Kitaev, "Quantum computations: algorithms and error correction", Russian Mathematical Surveys **52**, 1191–1249 (1997).

[54] V. Paulsen, *Completely bounded maps and operator algebras*, Vol. 78, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Jan. 2003).

[55] P. Hausladen and W. K. Wootters, "A 'pretty good' measurement for distinguishing quantum states", Journal of Modern Optics **41**, 2385 (1994).

[56] V. P. Belavkin, "Optimal multiple quantum statistical hypothesis testing", Stochastics **1**, 315 (1975).

[57] O. Christensen, *An introduction to frames and riesz bases* (Birkhäuser, Boston, 2003).

[58] O. Christensen, *Frames and bases: an introductory course*, 1st ed. (Birkhäuser Boston, July 2008).

[59] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge University Press, Mar. 2004).

[60] R. T. Rockafellar, *Convex analysis* (Princeton Univ Pr, June 1970).

[61] A. Barvinok, *A course in convexity* (American Mathematical Society, Nov. 2002).

[62] J. v. Tiel, *Convex analysis: an introductory text* (Wiley, Chichester [West Sussex] ; New York, 1984).

[63] S. Kullback and R. A. Leibler, "On information and sufficiency", The Annals of Mathematical Statistics **22**, 79–86 (1951).

[64] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", The Annals of Mathematical Statistics **23**, 493–507 (1952).

[65] I. Csiszár, "Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten", Magyar. Tud. Akad. Mat. Kutato Int. Kozl (Publication of the Mathematical Institute of the Hungarian Academy of Sciences) **8**, 85–108 (1963).

[66] T. Morimoto, "Markov processes and the h-theorem", Journal of the Physical Society of Japan **18**, 328–331 (1963).

[67] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another", Journal of the Royal Statistical Society. Series B (Methodological) **28**, 131–142 (1966).

[68] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory", English, IEEE Transactions on Information Theory **52**, 4394–4412 (2006).

[69] I. Csiszar and P. Shields, *Information theory and statistics: a tutorial* (Now Publishers Inc, Dec. 2004).

[70] S. Amari and H. Nagaoka, *Methods of information geometry*, Translations of mathematical monographs v. 191 (American Mathematical Society, Providence, RI, 2000).

[71] M. Ohya and D. Petz, *Quantum entropy and its use*, Corrected (Springer, May 2004).

[72] D. Petz, *Quantum information theory and quantum statistics*, 1st ed. (Springer, Dec. 2007).

[73] J. W. Gibbs, "Elementary principles in statistical mechanics", (C. Scribner's sons, 1902), p. 207.

[74] C. E. Shannon, "A mathematical theory of communication", Bell System Technical Journal **27**, 379–423 (1948).

[75] M. Tribus and E. C. McIrvine, "Energy and information", Scientific American **224**, 178–184 (1971).

[76] H. Araki and E. H. Lieb, "Entropy inequalities", en, Communications in Mathematical Physics **18**, 160–170 (1970).

[77] E. H. Lieb and M. B. Ruskai, "Proof of the strong subadditivity of quantum-mechanical entropy", Journal of Mathematical Physics **14**, 1938–1941 (1973).

[78] R. V. L. Hartley, "Transmission of information", en, Bell System Technical Journal **7**, 535–563 (1928).

[79] M. Blum, "Coin flipping by telephone a protocol for solving impossible problems", SIGACT News **15**, DOI:10.1145/1008908.1008911, 23–27 (1983).

[80] D. R. Stinson, *Cryptography: theory and practice* (CRC Press, 2005).

[81] C. E. Shannon, "Communication theory of secrecy systems", Bell System Technical Journal **28**, 656–715 (1949).

[82] S. Wiesner, "Conjugate coding", SIGACT News **15**, 10.1145/1008908.1008920, 78–88 (1983).

[83] C. H. Bennett and G. Brassard, "Quantum cryptography: public key distribution and coin tossing", Proceedings International Conference on Computers, Systems & Signal Processing (1984).

[84] D. Mayers, "Unconditionally secure quantum bit commitment is impossible", Phys. Rev. Lett. **78**, DOI:10.1103/PhysRevLett.78.3414, 3414–3417 (1997).

[85] P. W. Shor and J. Preskill, "Simple proof of security of the BB84 quantum key distribution protocol", Phys. Rev. Lett. **85**, DOI:10.1103/PhysRevLett.85.441, 441–444 (2000).

[86] E. Biham, M. Boyer, P. O. Boykin, T. Mor, and V. Roychowdhury, "A proof of the security of quantum key distribution", English, Journal of Cryptology **19**, DOI:10.1007/s00145-005-0011-3, 381–439 (2006).

[87] R. Renner, "Security of quantum key distribution", PhD thesis, ETH Zurich, arXiv:0512258 (2005).

[88] M. Tomamichel, C. C. W. Lim, N. Gisin, and R. Renner, "Tight finite-key analysis for quantum cryptography", Nature Communications **3**, 634 (2012).

[89] A. K. Ekert, "Quantum cryptography based on bell's theorem", Phys. Rev. Lett. **67**, DOI:10.1103/PhysRevLett.67 661–663 (1991).

[90] F. F. Centore, *Robert hooke's contributions to mechanics*, en (Springer Netherlands, Dordrecht, 1970).