Quantum Information Theory

Matthias Christandl, Maris Ozols, Michael Walter and Freek Witteveen

Spring 2024

Last updated: May 10, 2024 at 09:08

Preface

What is *quantum information*? To answer this question we must at least understand the basic structure of *quantum mechanics*, but we must also come to terms with the notion of *information*.

Information

The concept of information is ubiquitous in our age. Giving a precise definition of the meaning of the term is not so obvious. A basic intuition is that information is related to knowledge, and it can be transferred.

In the 1930s and 1940s, Turing and Shannon abstracted the *concept of information* from its *physical carrier* with the goal of a universal theory of information and computation that would apply to all physical systems. In the universal theory of information, the basic unit is the *bit*. A bit is something which can take two values. In practice, say in a computer, this could be whether a current is present or not, or whether a tiny magnet is pointing upward or downward, etc. However, for the purpose of information theory, the physical details are completely immaterial, and we simply mark the two states with values `0` or `1`.

In information theory we have in mind some process, a *source*, generating sequences of symbols. For instance, this could be you, typing on your computer. How can we use bits to measure information? Let us make this concrete with the fundamental example of *compression*. Suppose you have written a document on your computer. You can save the document by encoding each symbol into bits. If you wrote a text of length n using an alphabet of k symbols, this would naively require $n\lceil \log k \rceil$ bits (since you need $\lceil \log k \rceil$ bits to encode a single symbol from the alphabet). However, as you will probably be familiar with, you can also ask your computer to *compress* the text file into a smaller number of bits. This procedure is such that you can recover the original document from the compressed file by some algorithm. This suggests that a reasonable way to think about the amount of information present in your text document may be given by the smallest number of bits you can compress the file to.

A second basic aspect of information concerns transmission. Physical communication channels (such as cables or electromagnetic waves) are typically noisy: if one sends out a specific signal it may get corrupted along the way. However, one can *correct* the errors by adding redundancy to the signal. Information theory studies how to add as little redundancy as possible for reliable communication. Information theory is essential to the functioning of electronic telecommunication at high rates.

Quantum theory

The concept of information, as developed by Shannon, relies on an abstraction where the physical carrier of the information is no longer relevant. However, already before the advent of information theory, it had become clear that at small scales nature behaves *quantum mechanically*. As analog of the bit, quantum mechanics has the *qubit* as its minimal system.

When studying actual physical systems of certain (very small) currents or magnets, physicists have found that such systems behave in a fundamentally different way, and are described by the rules of quantum mechanics. So, when we want to know the *fundamental theory of information* we need to study sources producing quantum states and decide what information means in this context. Not only is this the fundamental theory of information, it also holds a promise to be useful for various practical applications, in combination with quantum computation. In quantum mechanics, the generalization of the bit is the *qubit*. Here, we should think of the two outcomes as vectors which we will call $|0\rangle$ and $|1\rangle$

$$|0\rangle := \begin{pmatrix}1\\0\end{pmatrix}, \qquad |1\rangle := \begin{pmatrix}0\\1\end{pmatrix}$$

and the qubit can be in a *superposition*

$$\alpha|0\rangle + \beta|1\rangle = \begin{pmatrix}\alpha\\\beta\end{pmatrix}.$$

In this case, it turns out that we have to allow the numbers α and β to be *complex* numbers $\alpha, \beta \in \mathbb{C}$. The correct normalization condition in this case is $|\alpha|^2 + |\beta|^2 = 1$. We can not directly 'observe' the state of the qubit. If we *measure* whether the qubit is in state $|0\rangle$ or $|1\rangle$, we find outcome 0 with probability $p_0 = |\alpha|^2$ and outcome 1 with probability $p_1 = |\beta|^2$. However, we can also choose a different basis of $\mathbb{C}^2$ and measure in that basis, and in that case the measurement probabilities are to be calculated in that basis. An important difference with classical models is that measurement *changes* or even destroys the quantum state. We will see that this leads to substantially different properties than a classical bit with probabilities p_0 and p_1.

An introduction to quantum information theory

There are at least two motivations to study a theory of quantum information. The first reason is *fundamental.* Since physical reality is quantum mechanical to the best of our knowledge, any fundamental theory of information needs to account for quantum mechanics. In this course we will see that even as we abstract away the details of the physical models this leads to different notions of information. This is not only important for our understanding of information, but can also be a guide to understand quantum physics itself! Concepts from quantum information theory have for example been applied in many-body physics (how does information propagate in quantum mechanical materials?) and quantum gravity thought experiments (what happens to information when it falls in a black hole?).

A second reason for studying quantum information theory is more practical: it has important applications! A well-known application is quantum key distribution, in which quantum bits are exchanged to enable secure communication in a way that is classically impossible. At the same time, we need to understand quantum information when we want to design cryptographic standards of the future that withstand attackers that are in possession of a quantum computer. Moreover, as the engineering challenge of building fault-tolerant quantum computers is overcome, *quantum communication* will be required to link multiple quantum computers. Principles of quantum information theory are also central to building quantum computers in the first place and to designing quantum algorithms to run on them.

Quantum theory was developed in the first half of the 20th century. Already early on, in the 1930s, unexpected information-theoretic properties of quantum theory came to light in the form of the EPR paradox. However, it was only in the second half of the 20th century, in the light of experimental progress in the manipulation of quantum systems, that the development of a theory of quantum information and computation accelerated.

In this course we will develop the *quantum theory of information* from first principles. We will begin by developing a formalism describing *quantum systems* which gives a unified way to describe both quantum mechanics and classical probability. We will introduce the fundamental objects of quantum information theory by providing a set of axioms for arbitrary finite dimensional quantum systems. Just as for classical information theory, the physical details of how such quantum systems arise will not concern us: we will treat the underlying physics as a black box providing us the basic 'rules' of quantum mechanics. In this course we will be guided by a number of natural questions about the nature of quantum mechanics.

The first six lectures of the course are spent on an exposition of the formalism of quantum states and quantum operations, providing an answer to the question

What is a quantum system and what constitutes a quantum information processing procedure?

Since we are classical agents ourselves, one could be tempted to think that since we only observe the classical outcomes of protocols and experiments, we only have to care about classical correlations.

What is the difference between quantum and classical correlations?

Fundamentally quantum correlation is known as *entanglement*. We will see how so-called *Bell games* allow one to distinguish between classical correlations and correlations arising from entangled quantum systems. Entanglement turns out to be an important resource for information theoretic tasks, and we will see two basic communication examples known as *teleportation* and *superdense coding*.

In the second half of the course, we introduce *quantum information theory*. The first main question we address is

What is quantum information?

We answer this by relating the notion of quantum information to *compression* of quantum sources into qubits. This leads to the notion of *entropy*, which is a measure for the amount of information in a source. We then study the properties of various entropic quantities, leading to a 'calculus' of quantum information theory.

The next broad question we will address is

How do we obtain reliable protocols from noisy resources?

This question is also one of the central problems of classical information theory, as discussed above: how can we communicate reliably and at a high rate over a noisy channel? As a concrete example, we will see how in the quantum setting we can obtain 'pure' entanglement from a noisy source. We study a more advanced quantum information processing task called *state merging*, which unifies a number of important protocols. After that, we study the rate at which we can reliably send quantum information over a noisy quantum channel.

The final question we address is

How can quantum mechanics be used for privacy?

We will see how the special properties of quantum mechanics can be used for secure cryptography. The intuition is that one can in general not obtain classical information about a quantum system without disturbing it, and that this allows one to detect eavesdroppers.

Of course, our answers to each of these questions is only a beginning! They are the starting point to deep and diverse fields of research with many beautiful results and open questions.

General references

There are many good books and lecture notes on quantum information theory. A modern classic is *Quantum computation and quantum information* by Nielsen and Chuang [31], which starts from the basics and offers an excellent introduction to both quantum computation and quantum information. Newer textbooks are *Quantum information theory* by Wilde [47] which covers a wide range of topics in quantum Shannon theory, the more mathematically oriented *The theory of quantum information* by Watrous [45] and *Quantum information theory* by Renes [38] which offers an operational perspective and covers topics in one-shot information theory. A useful set of physics lecture notes on quantum computation and information are due to Preskill [37].

The standard reference work for classical information theory is [12]. A charming and insightful textbook is [29]. Finally, [17] is a popular science book on information theory and is recommended reading.

Structure of the lectures

Each lecture has at its start a table with a brief summary of the main concepts and results of the lecture. At the end of every lecture there is an outlook, containing references both to sources for a more detailed treatment of the material as well as references to more further results related to the topic of the lecture. We also provide citations to some of the original works where these concepts were developed, without attempting to give a comprehensive overview. These references are intended to allow interested students to read further or find inspiration for a thesis topic and are optional reading. Crucially, there are also exercises at the end of each lecture! The exercises range from basic computations to more advanced problems which introduce new concepts and are roughly ordered by conceptual difficulty. The computational questions are such that one does not require a computer or calculator to solve them. Doing exercises is the best way to learn the subject; you are encouraged to try as many as possible!

Prerequisites

This course has been offered in master programs in Mathematics, Quantum Information Science, Computer Science, IT Security, and Physics at the Universities of Amsterdam, Bochum, and Copenhagen. A basic knowledge of linear algebra is required. We briefly introduce relevant facts from linear algebra in Appendix A. Ideally, students have previously taken an introductory course in quantum computing, quantum mechanics, or similar.

Acknowledgements

These lecture notes are based on previous sets of lecture notes by the authors for courses in Copenhagen and Amsterdam.

Summer 2024: Thanks to Maximilian Andersen and Vincent Krämer for pointing out corrections.

Spring 2023: Thanks to Tommaso Aschieri, Daniel Guerrero Domínguez, Jacob Fronk, Jonas Kruip, Miquel Llanos, Frederik Mols, Lea Northcote Sørensen and Matthew Teynor for pointing out corrections. Special thanks to Dylan Harley for helping to prepare many of the exercises.

Fall 2023: Thanks to Emil Hasse Henningsen, Thomas Mørk, Natacha Nielsen, Taro Spirig, Patrick Michael Olsen Sturm, Antonios Tsepas and Juliette Vlieghe for pointing out corrections.

Spring 2022: Thanks to Floris Westerman for spotting typos.

Spring 2021: Thanks to Christiaan van Asperen, Tim Blankenstein, Kjartan van Driel, Dylan Feenstra, Arend-Jan Quist, Misha Schram, and Jens de Vries for pointint out corrections. Special

thanks to Dmitry Grinko.

Spring 2020: Thanks to Khallil Berrekkal, Robert Cañellas, Egor Cickovskij, Jasper van Egeraat, Titas Geryba, Thijs Jenneskens, Dimitrios Loupas, Andrea Mazzocco for pointing out corrections. Special thanks to Harold Nieuwboer and Alvaro Piedrafita.

Spring 2019: Special thanks to Arjan Cornelissen.

Contents

Lecture 1

Bit and qubit: quantum states

Concept	Math translation
Quantum system	Hilbert space $\mathcal{H}$
Quantum state	A density matrix is a positive (semidefinite) operator ρ with trace 1. Pure states are given by $\rho = \lvert\psi\rangle\langle\psi\rvert$ for $\lvert\psi\rangle \in \mathcal{H}$. Classical probability distributions are diagonal density matrices.
Measurement	$\mu(x)$ with $\sum_x \mu(x) = \mathbb{1}$ and $\mu(x) \geq 0$. Probability of outcome x is $\mathrm{tr}[\mu(x)\rho]$.

We start with some intuition of what a (quantum) bit is supposed to be. Later in this lecture we will develop a precise version!

The basic unit of classical information theory is the *bit*. A bit is something which can take two values. In practice, say in a computer, this could be whether a current is present or not, or whether a tiny magnet is pointing upward or downward, etc. However, for the purpose of information theory, the physical details are completely immaterial, and we simply mark the two states with values `0` or `1`. A single bit is a random variable that can take value `0` with probability p_0, and value `1` with probability p_1. For this to make sense, we require the normalization $p_0 + p_1 = 1$, since with probability 1 we have one of the two outcomes.

When studying actual physical systems of certain (very small) currents or magnets, physicists have found that such systems behave in a fundamentally different way than ordinary random variable, and are described by the rules of quantum mechanics. So, when we want to know the *fundamental theory of information* we need to study sources producing quantum states and decide what information means in this context.

In quantum mechanics, the generalization of the bit is the *qubit*. Here, we should think of the two outcomes as vectors which we will call $\lvert 0\rangle$ and $\lvert 1\rangle$

$$\lvert 0\rangle := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \lvert 1\rangle := \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and the qubit can be in a *superposition*

$$\alpha\lvert 0\rangle + \beta\lvert 1\rangle = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

It turns out that we have to allow the numbers α and β to be *complex* numbers $\alpha, \beta \in \mathbb{C}$. The correct normalization condition in this case is $|\alpha|^2 + |\beta|^2 = 1$. We can not directly 'observe' the state of the qubit. If we *measure* whether the qubit is in state $|0\rangle$ or $|1\rangle$, we find outcome 0 with probability $p_0 = |\alpha|^2$ and outcome 1 with probability $p_1 = |\beta|^2$. However, we can also choose a different basis of $\mathbb{C}^2$ and measure in that basis, and in that case the measurement probabilities are to be calculated in that basis. We will see that this leads to substantially different properties than a classical bit with probabilities p_0 and p_1. In order to understand this, and to develop a theory of quantum information, we will begin by developing a formalism describing *quantum systems* which gives a unified way to describe both quantum mechanics and classical probability. In fact, what we described above was a *pure quantum state*, and we will define a more general notion which encompasses both pure quantum states and classical random variables. We will introduce the fundamental objects of quantum information theory by providing a set of axioms for arbitrary finite dimensional quantum systems. Just as for classical information theory, the physical details of how such quantum systems arise will not concern us: we will treat the underlying physics as a black box providing us the basic 'rules' of quantum mechanics.

Hilbert space

The state space of a quantum mechanical system is a complex *Hilbert space* $\mathcal{H}$. In this course we will restrict to *finite dimensional* complex Hilbert spaces,[1] which are simply complex vector spaces together with an inner product $\langle \cdot | \cdot \rangle$. We postulate this as our first axiom of quantum mechanics:

Axiom 1 (Hilbert space)**.** To every quantum system we associate a Hilbert space $\mathcal{H}$.

If the dimension of the Hilbert space is d we may identify $\mathcal{H} \cong \mathbb{C}^d$, with the standard inner product. A quantum system with Hilbert space $\mathbb{C}^d$ can be thought of as one which has d possible distinct 'states'. In Section 1.1 we will see precisely what we mean by a state.

We will use *bra-ket* notation. We denote vectors in the Hilbert space as $|\psi\rangle$. The Hilbert space of dual vectors $\mathcal{H}^*$ consists of linear maps $\mathcal{H} \to \mathbb{C}$. We use the notation $\langle \psi | \in \mathcal{H}^*$ where $\langle \psi |$ is the functional on $\mathcal{H}$ mapping

$$|\phi\rangle \mapsto \langle \psi | \phi \rangle.$$

The vector $|\psi\rangle$ is called the 'ket' and $\langle \psi |$ the 'bra'. Note that this notation is such that applying a 'bra' $\langle \psi |$ to a 'ket' $|\phi\rangle$ gives the 'bracket' inner product $\langle \psi | \phi \rangle$, so

$$\langle \psi | | \phi \rangle = \langle \psi | \phi \rangle.$$

This is perhaps a little abstract, but remember that you can always think of $|\psi\rangle$ as a *column* vector and $\langle \psi |$ as a *row* vector. If we have Hilbert space $\mathbb{C}^d$, then

$$|\psi\rangle = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{d-1} \end{pmatrix}$$

[1] If the vector space is infinite dimensional, there is the additional condition of completeness, which requires that any Cauchy sequence has a converging subsequence. This condition is always satisfied in finite dimensional complex inner product spaces.

where $\psi_i \in \mathbb{C}$. Then

$$\langle\psi| = \begin{pmatrix} \overline{\psi_0} & \overline{\psi_1} & \cdots & \overline{\psi_{d-1}} \end{pmatrix}$$

and

$$\langle\psi|\phi\rangle = \sum_{i=1}^{d} \overline{\psi_i}\phi_i.$$

We introduce the following notation for the standard basis of $\mathbb{C}^d$:

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad |1\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \cdots \qquad |d-1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

So, if

$$|\psi\rangle = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{d-1} \end{pmatrix}$$

then we may also write this as

$$|\psi\rangle = \sum_{i=0}^{d-1} \psi_i |i\rangle.$$

Finally, it will sometimes be useful to think of $|\psi\rangle$ as the linear map $\mathbb{C} \to \mathcal{H}$ which maps $z \in \mathbb{C}$ to $z|\psi\rangle$. This is natural, because a d-dimensional column vector can also be seen as a $d \times 1$ matrix.

Example 1.1. The simplest nontrivial Hilbert space is $\mathbb{C}^2$. This is known as a *qubit*, and it has a basis given by

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

This is the quantum version of a (classical) bit which can take values 0 and 1.

Most of the mathematics involved in quantum information theory (and quantum mechanics more generally) is linear algebra. In Appendix A we review the background in linear algebra that we need in these lectures. Here is a table with the concepts reviewed and the notation we will typically use for them throughout the lectures:

Concept	Notation
Hilbert space	$\mathcal{H}, \mathcal{K}, \ldots$
Inner product	$\langle v \vert w \rangle$
Standard basis for $\mathbb{C}^d$	$\vert 0\rangle, \vert 1\rangle, \ldots, \vert d-1\rangle$
Linear maps (operators, matrices) from $\mathcal{H}$ to $\mathcal{K}$	$M, N, \ldots \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$, $\mathrm{Lin}(\mathcal{H}) = \mathrm{Lin}(\mathcal{H}, \mathcal{H})$
Positive (semidefinite) matrices	$P, Q, \ldots \in \mathrm{PSD}(\mathcal{H})$, $P, Q \geq 0$
Unitary matrices and isometries	$U, V, \ldots \in \mathrm{U}(\mathcal{H})$ or $\mathrm{Isom}(\mathcal{H}, \mathcal{K})$
Identity matrix	$\mathbb{1}$
Adjoint (conjugate transpose) and transpose	$M^\dagger$, M^{T}
Trace	$\mathrm{tr}[M]$

We will use the words 'linear map', '(linear) operator' and 'matrix' more or less interchangeably, and denote the set of linear operators between $\mathcal{H}$ and $\mathcal{K}$ (or matrices, after choosing a basis) by $\mathrm{Lin}(\mathcal{H}, \mathcal{K})$, or $\mathrm{Lin}(\mathcal{H})$ if $\mathcal{H} = \mathcal{K}$. Important classes of linear operators are:

- Hermitian matrices: $M \in \mathrm{Lin}(\mathcal{H})$ is Hermitian (or self-adjoint) if $M^\dagger = M$.
- Positive matrices: $P \in \mathrm{Lin}(\mathcal{H})$ is positive (or positive semidefinite) if $\langle \psi | P | \psi \rangle \geq 0$ for all $\psi \in \mathcal{H}$. The set of positive matrices is denoted by $\mathrm{PSD}(\mathcal{H})$.
- Unitary matrices: $U \in \mathrm{Lin}(\mathcal{H})$ is unitary if $U^\dagger U = \mathbb{1}$ (so $U^{-1} = U^\dagger$). The set of unitary matrices is denoted by $\mathrm{U}(\mathcal{H})$.
- Isometries: $V \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ is an isometry if $V^\dagger V = \mathbb{1}$. The set of isometries is denoted by $\mathrm{Isom}(\mathcal{H}, \mathcal{K})$. When $\mathcal{H} = \mathcal{K}$ then these are the same as the unitaries: $\mathrm{Isom}(\mathcal{H}, \mathcal{H}) = \mathrm{U}(\mathcal{H})$.
- Projections: $P \in \mathrm{Lin}(\mathcal{H})$ is a projection if $P = P^\dagger$ and $P^2 = P$.

In Appendix A we review these notions and notation in more detail. Even if you are already familiar with the relevant linear algebra, it may be helpful to have a look at this appendix to familiarize yourself with the conventions and notation we choose! Especially helpful facts are the *spectral theorem* for Hermitian matrices, in Theorem A.1 and the *characterization of positive matrices* in Lemma A.2.

1.1 Quantum states

We now introduce the fundamental object of quantum information theory. We give a fairly abstract definition, after which we will explain why this definition is sensible and how it relates to perhaps more familiar notions of probabilities and quantum states.

Definition 1.2. A *density matrix* (or density operator) is a positive operator $\rho \in \mathrm{PSD}(\mathcal{H})$ with $\mathrm{tr}[\rho] = 1$. We denote the set of density matrices on $\mathcal{H}$ by

$$\mathrm{S}(\mathcal{H}) = \{\rho \in \mathrm{PSD}(\mathcal{H}) : \mathrm{tr}[\rho] = 1\}.$$

The importance of such density matrices is our next axiom of quantum theory:

Axiom 2 (Quantum states)**.** The state of a quantum system with Hilbert space $\mathcal{H}$ is described by a density matrix $\rho \in \mathrm{S}(\mathcal{H})$.

In light of Axiom 2 we will refer to a density matrix ρ as a *quantum state* and use the terms 'density matrix' and 'quantum state' interchangeably. While Axiom 2 is quite compact, it already contains a lot of information! To unpack this, we will start by investigating two important special cases of quantum states.

Classical states

Our formalism should generalize probability and information theory, and it should at least contain probabilities as a special case. What do we mean by 'probability theory'? We restrict to probability theory on finite sets (corresponding to our working assumption of finite dimensional Hilbert spaces). If we denote by Σ the finite set of outcomes, then a probability distribution assigns a real number $p(x) \geq 0$ to each outcome $x \in \Sigma$, and these numbers need to sum to 1. In other words, the collection of probability distributions on Σ is

$$\mathrm{P}(\Sigma) = \left\{ p : \Sigma \to \mathbb{R}_{\geq 0} \text{ such that } \sum_{x \in \Sigma} p(x) = 1 \right\}.$$

Often we also write $p_x = p(x)$ and say that $\{p_x\}_{x \in \Sigma}$ is a probability distribution.

Example 1.3. A die has outcomes in the set $\Sigma = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. If we have a fair die, then we have the probability distribution $p = \{p_x\}_{x \in \Sigma}$ with

$p($⚀$) = p($⚁$) = p($⚂$) = p($⚃$) = p($⚄$) = p($⚅$) = {}^1\!/_6.$

How can we relate probability distributions to quantum states? Suppose that $\mathcal{H}$ is a Hilbert space and we have basis vectors $|x\rangle \in \mathcal{H}$ for all possible outcomes $x \in \Sigma$. Then we can associate to any probability distribution $p \in \mathrm{P}(\Sigma)$ the quantum state $\rho = \sum_{x \in \Sigma} p(x) |x\rangle\langle x| \in \mathrm{S}(\mathcal{H})$. Often such a Hilbert space and basis comes about naturally.

For example, the standard basis of the Hilbert space of a qubit, $\mathcal{H} = \mathbb{C}^2$, is labeled by the possible values of an ordinary bit, $\Sigma = \{0, 1\}$, and the standard basis of $\mathcal{H} = \mathbb{C}^d$ is labeled by $\Sigma = \{0, 1, \dots, d-1\}$. In general, if Σ is any finite set, we write $\mathcal{H} = \mathbb{C}^\Sigma$ for a Hilbert space with an orthonormal basis $|x\rangle$ labeled by the elements $x \in \Sigma$. The vectors in $\mathcal{H}$ are "formal" linear combinations of these basis vectors:

$$\mathbb{C}^\Sigma = \left\{ |v\rangle = \sum_{x \in \Sigma} v_x |x\rangle : v_x \in \mathbb{C} \right\},$$

with inner product

$$\langle v|w\rangle = \sum_{x \in \Sigma} \overline{v_x} w_x.$$

When $\Sigma = \{0, \dots, d-1\}$ then this is nothing but $\mathbb{C}^d$. We call the basis $|x\rangle$ labeled by elements $x \in \Sigma$ the *standard basis* or computational basis.

In quantum information, it can be useful to work with Hilbert spaces that come with such a distinguished basis. Sometimes this is just a calculation tool, but more often than not the basis states are used to store some classical data (such as a message that one would like to transmit, encrypt, or compute with, the outcomes of a measurement, etc.). For this reason we call quantum states that arise from probability distributions "classical" quantum states:

Definition 1.4 (Classical states)**.** Let Σ be a finite set. A quantum state ρ on $\mathcal{H} = \mathbb{C}^\Sigma$ is called *classical* if it is of the form

$$\rho = \sum_{x\in\Sigma} p(x)|x\rangle\langle x| \tag{1.1}$$

where $p \in \mathrm{P}(\Sigma)$ is an arbitrary probability distribution. In other words, the classical states are precisely those with a density matrix that is diagonal with respect to the standard basis.

For example, the uniform die of Example 1.3 would be described by the density matrix

$$\rho = \frac{1}{6}\mathbb{1} = \frac{1}{6}\big(|⚀\rangle\langle⚀| + |⚁\rangle\langle⚁| + |⚂\rangle\langle⚂| + |⚃\rangle\langle⚃| + |⚄\rangle\langle⚄| + |⚅\rangle\langle⚅|\big) = \begin{pmatrix} 1/6 & & & \\ & 1/6 & & \\ & & \ddots & \\ & & & 1/6 \end{pmatrix}$$

on the Hilbert space $\mathbb{C}^\Sigma$ with basis labeled by $\Sigma = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$. In general, we need to choose an order of the elements in Σ to write down a matrix representation. Here this did not matter, because the probability distribution is uniform and so all probabilities are the same.

Pure states

Given a vector $|\psi\rangle \in \mathcal{H}$ which has unit norm, so $\langle\psi|\psi\rangle = 1$, we can let $\rho = |\psi\rangle\langle\psi|$. It is clear that $\rho \in \mathrm{PSD}(\mathcal{H})$ by Lemma A.2. It is also normalized, as

$$\mathrm{tr}[\rho] = \mathrm{tr}[|\psi\rangle\langle\psi|] = \langle\psi|\psi\rangle = 1.$$

Therefore, ρ is a quantum state. Such states are called *pure states*.

Definition 1.5. A quantum state $\rho \in \mathrm{S}(\mathcal{H})$ is a *pure state* if there exists $|\psi\rangle \in \mathcal{H}$ such that $\rho = |\psi\rangle\langle\psi|$. A state which is not pure is called *mixed*.

If we have a pure state $\rho = |\psi\rangle\langle\psi|$ we will (in a slight abuse of language) also refer to $|\psi\rangle$ as a pure state.

There is a redundancy in $|\psi\rangle$ as a description of the state: multiplying $|\psi\rangle$ with a *phase* does not change the associated density matrix, since if we let $|\phi\rangle = e^{i\theta}|\psi\rangle$ for $\theta \in \mathbb{R}$, then $|\phi\rangle\langle\phi| = |\psi\rangle\langle\psi|$. To get rid of this redundancy is one reason to consider the density operator ρ instead of the vector $|\psi\rangle$.

One can develop a theory of quantum mechanics just using pure states (and maybe you have seen this in a previous course on quantum mechanics or quantum computing!). You can think of a pure state as one which is in a 'deterministic' quantum state, i.e. there is no classical randomness. This also motivated why it is desirable to also consider states that are mixed.

Note that the pure state $\rho = |\psi\rangle\langle\psi|$ is a projection onto the one-dimensional space spanned by $|\psi\rangle$. You may prove the following lemma in Exercise 1.18.

Lemma 1.6. *Let $\rho \in \mathrm{S}(\mathcal{H})$. The following are equivalent:*

(a) ρ is a pure state.

(b) ρ has rank one.

(c) ρ is a projection.

Example 1.7. As an example of pure qubit states, consider $|0\rangle$ and $|1\rangle$ on the qubit Hilbert space $\mathcal{H} = \mathbb{C}^2$. The corresponding density matrices are

$$|0\rangle\langle 0| = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad |1\rangle\langle 1| = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

More generally, if we let

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \text{ with } \alpha, \beta \in \mathbb{C}, |\alpha|^2 + |\beta|^2 = 1$$

be an arbitrary normalized vector, we get

$$\rho = |\psi\rangle\langle\psi| = \begin{pmatrix} |\alpha|^2 & \alpha\overline{\beta} \\ \overline{\alpha}\beta & |\beta|^2 \end{pmatrix}.$$

For instance, $|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle)$ gives density matrix

$$\rho = \frac{1}{2}\begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix}.$$

This should be contrasted with the *classical* states of a qubit, which take the form

$$\rho = \begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}.$$

General states

We now consider a general quantum state $\rho \in \mathrm{S}(\mathcal{H})$. By Lemma A.2, because ρ is positive, we may write an eigendecomposition

$$\rho = \sum_{i=1}^{d} p_i |\psi_i\rangle\langle\psi_i| \tag{1.2}$$

where $p_i \geq 0$ are the eigenvalues and where the $|\psi_i\rangle$ form a basis of eigenvectors. Moreover, since we may compute the trace using the basis $\{|\psi_i\rangle\}_{i=1}^d$ (in which ρ is a diagonal matrix), the fact that $\mathrm{tr}[\rho] = 1$ is equivalent to the condition

$$\sum_{i=1}^{d} p_i = 1.$$

In other words, the eigenvalues $p = \{p_i\}_{i=1}^d$ define a probability distribution. This leads to the following interpretation of the state ρ: it describes a situation where we have the pure state $|\psi_i\rangle$ with probability p_i. For instance, you may think of some device which prepares a quantum state if we push a button. What the device does is that, upon pushing the button, it samples (classically!) according to the probability distribution p. If it samples outcome i it prepares the *pure* state $|\psi_i\rangle$. This situation would be described by the density matrix in Eq. (1.2). However, there is an important caveat here: this interpretation is *not unique*! Any decomposition

$$\rho = \sum_{j \in J} q_j |\phi_j\rangle\langle\phi_j|$$

where the $|\phi_j\rangle$ are pure states (but not necessarily pairwise orthogonal!) and $q_j \geq 0$ gives rise to such an interpretation! Indeed, if we again compute

$$1 = \operatorname{tr}[\rho] = \sum_{j \in J} q_j \underbrace{\operatorname{tr}[|\phi_j\rangle\langle\phi_j|]}_{=1} = \sum_{j \in J} q_j$$

we see that $\{q_j\}_{j \in J}$ is a probability distribution. Hence we can also think of ρ as also describing the situation where have the pure state $|\phi_j\rangle$ with probability q_j. This is a first fundamental difference between probability distributions and quantum states.

Remark 1.8. The decomposition in Eq. (1.2) looks similar to our definition of classical states. However, what is crucial for classical states is that they are diagonal in the *standard* basis, whereas the basis in Eq. (1.2) depends on the state ρ and will in general be a different basis.

Example 1.9. For any Hilbert space $\mathcal{H}$ of dimension d we may define the *maximally mixed state*

$$\tau = \frac{1}{d}\mathbb{1}.$$

This can be decomposed as

$$\tau = \sum_{j=1}^{d} \frac{1}{d} |e_j\rangle\langle e_j|$$

for any choice of basis $\{|e_j\rangle\}_{j=1}^{d}$ of $\mathcal{H}$. In other words, if we pick a basis vector uniformly at random then we get the maximally mixed state, independent of which basis we consider.

Recall that a subset X of a (real or complex) vector space V is called *convex* if for any two $x_0, x_1 \in X$ it holds that for all $0 \leq t \leq 1$

$$x_t = tx_1 + (1-t)x_0 \in X.$$

This means that if we draw a line segment between the points x_0 and x_1, this line segment is contained in X. We say that x_t is a *convex combination* of x_0 and x_1. The *extreme points* of a convex set X are the elements $x \in X$ which have the following property: if one writes x as a convex combination

$$x = tx_1 + (1-t)x_0 \text{ for } x_0, x_1 \in X,$$

with $0 < t < 1$, then $x_0 = x_1 = x$. In other words, an extreme point is one that cannot be written as a nontrivial convex combination of elements in X.

The set of probability distributions $\mathrm{P}(\Sigma)$ is a convex set known as the probability simplex. For a bit, $\mathrm{P}(\{0,1\}) \subset \mathbb{R}^2$ is simply the line segment connecting the deterministic distributions $\binom{1}{0}$ and $\binom{0}{1}$, which are the two extreme poins. Similarly, $\mathrm{P}(\{0,1,2\}) \subset \mathbb{R}^3$ is a triangle, and so forth. What is the shape of the set of quantum states?

Lemma 1.10. *The set* $\mathrm{S}(\mathcal{H}) \subset \mathrm{Lin}(\mathcal{H})$ *is convex. The set of its extreme points coincides with the set of pure states.*

The proof is Exercise 1.10.

Qubits and the Bloch sphere

As we saw in Example 1.1 and Example 1.7 the most basic quantum system is a system with Hilbert space $\mathbb{C}^2$ (a one-dimensional Hilbert space is trivial), which we call a *qubit*. One of the most useful skills for a quantum information theorist is to have a good grasp of this basic quantum system, as it will function as a building block for larger quantum systems. So, we will now very explicitly parametrize and visualize qubit states. The following matrices

$$\mathbb{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \qquad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

form a basis of the real vector space of Hermitian matrices. The matrices X, Y, and Z are traceless. These matrices are called the *Pauli matrices*. They have a number of properties which are very useful in computations. First of all, they are Hermitian as well as unitary, and therefore they square to the identity:

$$X^2 = Y^2 = Z^2 = \mathbb{1}.$$

Moreover, X, Y, Z have eigenvalues ±1. They anticommute and multiply in the following cyclic manner

$$XY = -YX = iZ, \quad YZ = -YZ = iX, \quad ZX = -XZ = iY.$$

This implies that the trace of a product of two different Pauli operators is zero, e.g. $\operatorname{tr}[XY] = 0$. Check for yourself that these properties are indeed valid!

We can expand an arbitrary Hermitian operator $\rho \in \operatorname{Lin}(\mathbb{C}^2)$ with $\operatorname{tr}[\rho] = 1$ as

$$\rho = \frac{1}{2}(\mathbb{1} + xX + yY + zZ) = \frac{1}{2}\begin{pmatrix} 1+z & x-iy \\ x+iy & 1-z \end{pmatrix}$$

where $x, y, z \in \mathbb{R}$ (using that X, Y, Z, I are a basis of the real vector space of Hermitian matrices, and X, Y, Z are traceless).

When is $\rho \geq 0$ and therefore a quantum state? If $\lambda_1, \lambda_2 \in \mathbb{R}$ denote the two eigenvalues of ρ, then $\operatorname{tr}[\rho] = \lambda_1 + \lambda_2 = 1$. This means that at least one of λ_1, λ_2 must be positive. Therefore, $\rho \geq 0$ if and only if $\lambda_1\lambda_2 = \det(\rho) \geq 0$. We compute the determinant to be

$$\det(\rho) = \frac{1}{4}\left((1+z)(1-z) - (x+iy)(x-iy)\right) = \frac{1}{4}(1 - x^2 - y^2 - z^2),$$

so $\rho \geq 0$ if and only if the vector[2] $\vec{r} = (x, y, z)$ has norm at most 1. In other words, the states of a qubit are parametrized by a solid ball of radius one in $\mathbb{R}^3$. This is known as the *Bloch ball*, and the vector $\vec{r}$ is called the *Bloch vector*.

When is the state pure? The answer follows from Lemma 1.6: this is the case if and only if we have one eigenvalue equal to 1 and one equal to 0, so if and only if $\det(\rho(\vec{r})) = 0$, which is equivalent to the vector $\vec{r}$ being a unit vector. Thus, the pure states of a qubit are parameterized by the unit sphere in $\mathbb{R}^3$, called the *Bloch sphere*.[3] We summarize:

[2] We write $\vec{r}$ for vectors in the real vector space $\mathbb{R}^3$ and write the inner product between $\vec{r}$ and $\vec{s}$ as $\vec{r} \cdot \vec{s}$ to avoid confusion with *complex* Hilbert spaces for which we use bra-ket notation.

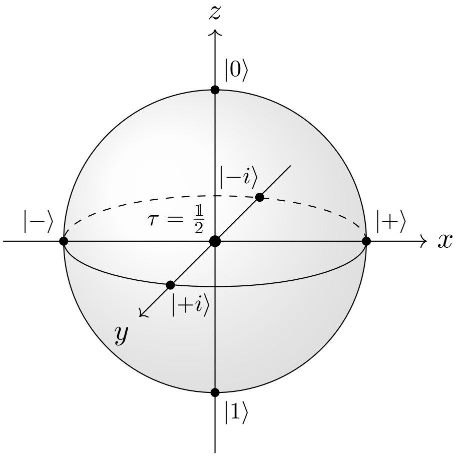
[3] Note that this is different from the characterization of pure states by unit vectors in $\mathbb{C}^2$, which would correspond to the unit sphere in $\mathbb{C}^3 \cong \mathbb{R}^4$. If you want to be mathematical about this, the Bloch sphere corresponds to an identification of the complex projective space $\mathbb{CP}^1$ with the two-dimensional real sphere S^2.

Lemma 1.11 (Bloch ball and sphere). *Any state $\rho \in \mathrm{S}(\mathbb{C}^2)$ of a qubit can be written in the form*

$$\rho = \frac{1}{2}\left(I + r_x X + r_y Y + r_z Z\right), \tag{1.3}$$

where $\vec{r} = \left(\begin{smallmatrix} r_x \\ r_y \\ r_z \end{smallmatrix}\right)$ is an arbitrary vector of norm $\|\vec{r}\| \leq 1$. Moreover, ρ is pure if and only if $\|\vec{r}\| = 1$. We denote the density matrix corresponding to $\vec{r}$ by $\rho(\vec{r})$.

We can visualize the situation as follows:



The interior of the Bloch ball represents mixed states. In particular, the center $\vec{r} = (0,0,0)$ corresponds to the maximally mixed state $\rho(\vec{r}) = \tau = \frac{\mathbb{1}}{2}$. A vector $\vec{r} = (0,0,2p-1)$ on the z-axis with $p \in [0,1]$ corresponds to the classical state

$$\rho = \frac{1}{2}\left(I + (2p-1)Z\right) = \frac{1}{2}\begin{pmatrix} 1 + (2p-1) & 0 \\ 0 & 1 - (2 - p1) \end{pmatrix} = \begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}.$$

This is a classical bit taking values 0 and 1 with probabilities p and $1-p$, respectively.

Let us see which states are on the intersection of the axes with the sphere:

z-axis:	$\vec{r} = (0,0,1):\ |0\rangle$	$\vec{r} = (0,0,-1):\ |1\rangle$
x-axis:	$\vec{r} = (1,0,0):\ |+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$	$\vec{r} = (-1,0,0):\ |-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$
y-axis:	$\vec{r} = (0,1,0):\ |+i\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle)$	$\vec{r} = (0,-1,0):\ |-i\rangle = \frac{1}{\sqrt{2}}(|0\rangle - i|1\rangle)$.

This gives three different bases of $\mathbb{C}^2$, which we will refer to as the X, Y, and Z bases. This is because the X-basis, with $\vec{r} = (\pm 1, 0, 0)$, consists of the eigenvectors $|\pm\rangle$ of the Pauli X matrix. Similarly, $\vec{r} = (0, \pm 1, 0)$ corresponds to the eigenvectors $|\pm i\rangle$ of the Pauli Y matrix, and the Z-basis, with $\vec{r} = (0,0,\pm 1)$, is given by the eigenvectors $|0\rangle, |1\rangle$ of the Pauli Z matrix.

1.2 Measurements

When we manipulate quantum states, and we 'observe' them, in practice using some measurement device, the information we obtain is classical. So, our formalism requires a notion of *measurements*, which converts quantum systems to classical outcomes. We will first give a rather formal definition, and then describe a special case for pure states which you may already be familiar with or at least will provide further intuition.

Definition 1.12 (Measurement)**.** A *measurement* or *positive operator valued measure* (POVM) on a Hilbert space $\mathcal{H}$ with a set of outcomes Ω is a collection of operators $\mu = \{\mu(x) \in \mathrm{PSD}(\mathcal{H})\}_{x\in\Omega}$ such that

$$\sum_{x\in\Omega} \mu(x) = \mathbb{1}.$$

We denote by $\mathrm{Meas}(\mathcal{H}, \Omega)$ the set of measurements on $\mathcal{H}$ with outcomes in Ω.

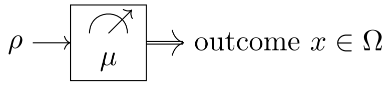
Often we will also write $\mu_x = \mu(x)$ and say that $\{\mu_x\}_{x\in\Omega}$ is a measurement. Note that this definition implies that for each x we have $0 \leq \mu(x) \leq \mathbb{1}$. This in turn means that each $\mu(x)$ is Hermitian and has eigenvalues between 0 and 1.

This looks similar to a probability distribution, except that $\mu(x)$ is an operator and not a number. However, we can apply the measurement to a quantum state to obtain a probability distribution of classical outcomes. This is described in the following axiom:

Axiom 3 (Measurement)**.** When performing a measurement $\mu \in \mathrm{Meas}(\mathcal{H}, \Omega)$ on a quantum state $\rho \in \mathrm{S}(\mathcal{H})$, we observe outcome $x \in \Omega$ with probability

$$p(x) = \mathrm{tr}[\mu(x)\rho].$$

We will use the following graphical notation for a measurement, where time reads from left to right. Lines or arrows indicate the presence of a quantum system, and double lines or arrows indicate that we have classical outcomes:



Example 1.13. Again, we take a qubit. We consider outcomes 0 and 1, and let

$$\mu(0) = |0\rangle\langle 0| \qquad \mu(1) = |1\rangle\langle 1|.$$

It is clear that this defines a measurement. If we measure the state $\rho = |\psi\rangle\langle\psi|$ for $|\psi\rangle = |0\rangle$, it is easy to see that the probability of outcomes is

$$p(0) = 1 \qquad p(1) = 0.$$

If we instead measure $|\psi\rangle = |+\rangle$, we find

$$p(0) = \mathrm{tr}\left[|0\rangle\langle 0|\frac{1}{2}\left(|0\rangle + |1\rangle\right)\left(\langle 0| + \langle 1|\right)\right] = \frac{1}{2}$$

and

$$p(1) = \mathrm{tr}\left[|1\rangle\langle 1|\frac{1}{2}\left(|0\rangle + |1\rangle\right)\left(\langle 0| + \langle 1|\right)\right] = \frac{1}{2}.$$

As a final example, consider the maximally mixed state τ. Then

$$p(0) = \mathrm{tr}\left[|0\rangle\langle 0|\frac{1}{2}\mathbb{1}\right] = \frac{1}{2} = \mathrm{tr}\left[|1\rangle\langle 1|\frac{1}{2}\mathbb{1}\right] = p(1).$$

We see that this measurement does not distinguish between $|+\rangle$ and the maximally mixed state!

Basis measurements and projective measurements

We now generalize the construction of a measurement in Example 1.13. Choose an arbitrary basis $\{|e_x\rangle\}_{x\in\Omega}$ of $\mathcal{H}$. Then we can construct a measurement with set of outcomes Ω by taking

$$\mu(x) = |e_x\rangle\langle e_x| \quad \text{for } x \in \Omega.$$

It is clear that these operators define a measurement since for any basis

$$\sum_{x\in\Omega} |e_x\rangle\langle e_x| = \mathbb{1}.$$

Such a measurement is called a *basis measurement* and is the most important special case of a measurement. In particular, for a quantum system with standard basis $|x\rangle$ labeled by $x \in \Sigma$ we can always perform a *standard basis measurement* by taking $\Omega = \Sigma$ and $\mu(x) = |x\rangle\langle x|$.

Lemma 1.14 (Born's rule)**.** *If we have a quantum state $\rho \in \mathrm{S}(\mathcal{H})$ and we perform a basis measurement $\mu(x) = |e_x\rangle\langle e_x|$, we observe outcome $x \in \Omega$ with probability*

$$p(x) = \langle e_x|\rho|e_x\rangle.$$

In particular, if $\rho = |\psi\rangle\langle\psi|$ is a pure state with

$$|\psi\rangle = \sum_{x\in\Omega} \psi_x |e_x\rangle,$$

then

$$p(x) = |\psi_x|^2.$$

Proof. This is an immediate consequence of Eq. (A.2):

$$p(x) = \mathrm{tr}[|e_x\rangle\langle e_x|\rho] = \langle e_x|\rho|e_x\rangle.$$

When $\rho = |\psi\rangle\langle\psi|$ with $\sum_{x\in\Omega} \psi_x|e_x\rangle$, then

$$p(x) = \langle e_x||\psi\rangle\langle\psi||e_x\rangle = |\langle\psi|e_x\rangle|^2 = |\psi_x|^2. \qquad \square$$

Finally, a slightly more general special case is where the measurement operators $\mu(x)$ are all projections, so $\mu(x) = P_x$. This is called a *projective measurement* or *projection-valued measure* (PVM). The condition that the measurement operators sum to the identity implies that the image of the different measurement operators must be orthogonal, and the images of the operators P_x must together span all of $\mathcal{H}$.

1.2.1 Qubit measurements

We return to our basic model: the qubit. What measurements are possible for a qubit? We will restrict to basis measurements for simplicity. You may verify the details of the following in Exercise 1.20. Up to phases (which are in any case not important) a choice of basis corresponds to an axis (i.e., a line through the origin) of the Bloch ball. This axis intersects the Bloch sphere in two pure states. For example, the z-axis gives a basis measurement in $|0\rangle, |1\rangle$, the x-axis gives measurement in the $|+\rangle, |-\rangle$ basis, etc. In general, given $\vec{r} = (x, y, z)$ with $\|\vec{r}\| = 1$, the state with $-\vec{r}$ is orthogonal, so

$$\mu_{\vec{r}}(0) = \rho(\vec{r}) = \frac{1}{2}\begin{pmatrix} 1+z & x-iy \\ x+iy & 1-z \end{pmatrix}, \qquad \mu_{\vec{r}}(1) = \rho(-\vec{r}) = \frac{1}{2}\begin{pmatrix} 1-z & -x+iy \\ -x-iy & 1+z \end{pmatrix} \tag{1.4}$$

defines a general qubit basis measurement. In Exercise 1.20 you will show that when performing this measurement on a quantum state with Bloch vector $\vec{s} = (x', y', z')$, then the probability of obtaining outcome 0 is given by

$$p(0) = \frac{1}{2} + \frac{1}{2}\vec{r} \cdot \vec{s} = \frac{1}{2} + \frac{1}{2}(xx' + yy' + zz').$$

Geometrically, $\vec{r} \cdot \vec{s}$ is the projection of the Bloch vector of the state onto the axis defining the measurement. This is consistent what we found in Example 1.13: we saw that if we measure in the standard Z-basis, we get a fixed outcome when we measure $|0\rangle$, but a completely random (uncertain) outcome when we measure $|+\rangle$. Similarly, if we measure $|0\rangle$ in the X-basis, we get a uniformly random outcome, whereas we get a fixed outcome when we measure $|+\rangle$.

In Exercise 1.19, you can show an *uncertainty relation* for qubits, which establishes a precise quantitative tradeoff between the uncertainty in the two measurement outcomes. In particular, there exists no quantum state for which both outcomes of an X and Z measurement are certain! This is a second fundamental difference between probability distributions and quantum states.

1.2.2 Measurements map quantum states to classical values

Our general definition of a 'measurement' and the corresponding Axiom 3 may seem to come out of thin air. To give it some motivation, we shall argue that it is a natural notion given our definition of a quantum state. What should be the most basic requirements a measurement must satisfy? It is clear that it should assign to any quantum state ρ a probability distribution p_ρ that describes the probabilities $p_\rho(x)$ of observing any outcome $x \in \Omega$. Secondly, consider the situation where have state ρ_1 with probability p_1 and state ρ_2 with probability p_2 (again, you may think of a device preparing, by making a random choice, either state ρ_1 or state ρ_2). Then it is reasonable to expect that the probability of obtaining outcome x is given by the mixture of the probabilities of obtaining outcome x for ρ_1 and ρ_2:

$$p_\rho(x) = p_1 p_{\rho_1}(x) + p_2 p_{\rho_2}(x) \quad \text{for } x \in \Omega,$$

or simply $p_\rho = p_1 p_{\rho_1} + p_2 p_{\rho_2}$. These two demands lead to our notion of measurement!

Lemma 1.15. *Suppose that we have a set of outcomes Ω, a Hilbert space $\mathcal{H}$, and for any state $\rho \in \mathrm{S}(\mathcal{H})$ a probability distribution $p_\rho \in \mathrm{P}(\Omega)$ such that the following holds:*

$$p_\rho = p_1 p_{\rho_1} + p_2 p_{\rho_2} \quad \textit{for any mixture } \rho = p_1\rho_1 + p_2\rho_2 \textit{ of states } \rho_1, \rho_2 \in \mathrm{S}(\mathcal{H})$$

Then there exists a measurement $\mu \colon \Omega \to \mathrm{PSD}(\mathcal{H})$ such that we have for all $x \in \Omega$

$$p_\rho(x) = \mathrm{tr}[\mu(x)\rho].$$

You can prove this in Exercise 1.17. Lemma 1.15 leaves open whether allowing *any* such measurement is physically reasonable. Later in the course we will provide evidence for this, by showing that any measurement can be constructed from a simpler set of operations (that you might already be familiar with from a previous course in quantum mechanics or quantum computing).

What happens to the quantum state after it has been measured? An important feature of quantum mechanics is that it is not possible to perform measurements on arbitrary states *without altering the state*. This is sometimes known as *collapse of the wave function*. In this formulation, it is typically assumed that in the case of a basis measurement, upon receiving outcome x, the state has collapsed to the post-measurement state $|x\rangle$.

However, for now we will take the perspective that the measurement is destructive and that the quantum state completely disappears after the measurement, leaving us only with the classical information about which outcome has occurred. In a later lecture, when we model quantum processes more generally, we will see how one can model general post-measurement states and prove that it is not possible to perform measurements without altering the state.

At this point, we summarize the objects we have introduced, and which symbols we conventionally use to denote them.

Concept	Notation
Quantum state	$\rho, \sigma, \ldots \in \mathrm{S}(\mathcal{H})$
Pure quantum state	$\|\phi\rangle, \|\psi\rangle, \ldots \in \mathcal{H}$
Maximally mixed state on $\mathcal{H} = \mathbb{C}^d$	$\tau = \frac{\mathbb{1}}{d}$
Measurement with outcomes $x \in \Omega$	$\mu, \nu, \ldots \in \mathrm{Meas}(\mathcal{H}, \Omega)$ $\mu = \{\mu(x)\}_{x\in\Omega}$ or $\{\mu_x\}_{x\in\Omega}$
Probability distributions	$p, q, \ldots \in \mathrm{P}(\Sigma)$ $p = \{p(x)\}_{x\in\Sigma}$ or $\{p_x\}_{x\in\Sigma}$

Outlook

The material in this lecture is covered in many textbooks and lecture notes. Standard modern introductions to the formalism of quantum mechanics from the perspective of information theory and computer science are [31,45,47]. A detailed description of the geometry of the set of quantum states, and applications to quantum information theory, can be found in [3]. An early exposition of the formalism of mixed quantum states is given by von Neumann in [43].

In this course we restrict ourselves to *finite dimensional* Hilbert spaces. For many models of the physics of quantum mechanical systems it is important to use infinite dimensional Hilbert spaces. For example, one may have a quantum particle that can have a *position* on a continuous space. At a fundamental level, models of particle physics (quantum field theories) always are based on infinite dimensional Hilbert spaces. One can indeed also develop quantum information theory for infinite dimensional systems. In the infinite dimensional setting, the most basic model is the *harmonic oscillator*. This is the starting point for *continuous variable* quantum information. A textbook with a discussion on quantum information in infinite dimensional systems is [22]. An operator algebraic perspective on quantum information, suitable for infinite dimensional systems, is given in [32]. While fundamental physics often requires infinite dimensional systems, for information processing purposes it is in many cases reasonable to restrict to a finite number of possible states, and thereby reduce to a finite dimensional Hilbert space. This not dissimilar to classical information and computer science: while electrical currents take continuous variables, one may do information processing and computation using a discrete set of current values (say 'on' or 'off'). Nevertheless, there are interesting phenomena in quantum information theory which are of a fundamentally infinite dimensional nature!

A fundamental aspect of quantum theory we have glossed over in this lecture is the *interpretation* of quantum mechanics. From its conception, quantum mechanics has been the subject of a debate on what quantum mechanics is supposed to mean precisely. What does it mean when we say that a quantum system is in some 'state'? What happens to the state when we measure? Why does measurement have a special status in quantum theory?

In the first half of the twentieth century, many physicists where unsatisfied with the formalism of quantum mechanics, and saw it as an 'effective' model which should really have some classical underlying description where there is a fixed classical description of states. However, not only have physical models using the set-up of the quantum mechanical formalism have been very successful in describing the world, there is actually strong experimental evidence that the physical world is *fundamentally quantum mechanical* (more about this in Lecture 3!). There are various 'competing' interpretations of quantum mechanics, which have different ideas on what the status of a quantum state is and what the role of measurements is. Some of the main interpretations are the Copenhagen interpretation [33], the many-worlds interpretation [13], pilot wave theory [4] and Bayesian interpretations [16]. While a fascinating debate, we will completely ignore these questions and just learn how to understand the quantum world *given the rules of the game* (i.e. the axioms defined in this lecture). This is known as the 'shut up and calculate' approach to quantum mechanics [30], see also [1].

1.3 Exercises

1.1 **Hermitian matrices:**

(a) If $M \in \mathrm{Lin}(\mathbb{C}^2)$, how can you compute the eigenvalues of M from $\det M$ and $\operatorname{tr} M$?

(b) Verify that the eigenvalues of the Pauli matrices X, Y, Z are ± 1 and compute the eigenvectors. *Hint: Page 18.*

(c) Consider the matrix $H = |0\rangle\langle 0| + |+\rangle\langle +|$, where $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Compute its eigenvectors and eigenvalues.

1.2 **Eigenvalue basics:**

(a) Suppose that $M = \sum a_i |\psi_i\rangle\langle\psi_i|$. Show that when the $|\psi_i\rangle$ are orthogonal then they are eigenvectors of M. Show that when the $|\psi_i\rangle$ are orthonormal then the numbers a_i are eigenvalues of M. Are these assumption necessary?

(b) Compute the eigenvalues and eigenvectors of the matrix

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Hint: You can avoid computing the determinant of a 4 by 4 matrix!

(c) Compute the eigenvalues and eigenvectors of the matrices

$$H_1 = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad H_2 = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \theta \in [0, 2\pi]$$

1.3 **Quantum states:** Consider the following operators in $\mathrm{Lin}(\mathbb{C}^2)$.

$$\rho_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \rho_2 = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \rho_3 = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \rho_4 = \frac{1}{2}\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

(a) Which of the ρ_i are density matrices? Which correspond to pure states?
(b) Write the ρ_i in bra-ket notation.
(c) Write a spectral decomposition for those ρ_i that are Hermitian.

1.4 **Trace versus inner product:** Let $M = |\psi\rangle\langle\psi|$, $N = |\phi\rangle\langle\phi|$ for $|\psi\rangle, |\phi\rangle \in \mathcal{H}$. Verify the following relation: $\mathrm{tr}[MN] = |\langle\psi|\phi\rangle|^2$.

1.5 **Mixed states and measurements:** Suppose that $\rho \in \mathrm{S}(\mathcal{H})$ is a mixture

$$\rho = \sum_i p_i \rho_i$$

where the p_i form a probability distribution and $\rho_i \in \mathrm{S}(\mathcal{H})$. Let μ be a measurement on $\mathcal{H}$ with outcomes $x \in \Omega$. Show that the following two procedures lead to the same outcome distribution of measurement outcomes:

(a) Measure the state ρ using μ.
(b) Sample i from the distribution p_i and measure the state ρ_i using μ.

1.6 **Mixed states as probabilistic mixtures:** For each of the following scenarios, write down the density matrix that results from the described procedure. Write down the density matrix both in bra-ket notation and in matrix form. All systems are qubits.

(a) Alice flips a fair coin. If the coin is heads, they prepare the state $|0\rangle$. If the coin lands tails, they prepare $|+\rangle$. You receive the state (but not the result of the coin toss).
(b) Alice measures the state $|0\rangle$ in the X-basis, with outcomes $+$ and $-$. Upon finding outcome $+$, they prepare the state $|0\rangle$, while upon finding $-$ they prepare $|1\rangle$. You receive the state (but not the measurement outcome).
(c) Alice source has the state $\frac{1}{\sqrt{5}}(|0\rangle + 2|1\rangle)$ and measures in the standard basis. If they obtain outcome 0, they prepare the state $|0\rangle$. If they obtain outcome 1, they prepare $|+\rangle$ with probability $\frac{1}{2}$ and $|-\rangle$ with probability $\frac{1}{2}$.

1.7 **Spectral theorem:** Using the spectral theorem for Hermitian matrices (Theorem A.1), show that any Hermitian matrix can be diagonalized by a unitary matrix. That is, show that for any $M \in \mathrm{Lin}(\mathbb{C}^d)$ with $M^\dagger = M$, there exists a unitary matrix $U \in \mathrm{U}(\mathbb{C}^d)$ and a diagonal matrix $\Lambda \in \mathrm{Lin}(\mathbb{C}^d)$ such that

$$M = U \Lambda U^\dagger.$$

1.8 **Positive numbers and positive matrices:**

(a) Give an example of a matrix such that each of the entries is positive, but the matrix is not positive.
(b) Give an example of positive matrices P, Q such that their product PQ is not positive.

1.9 **Criteria for positive definiteness:** Show a matrix P is positive *definite*, meaning $\langle\psi|P|\psi\rangle > 0$ for all $\psi \in \mathcal{H}$, if and only if it is Hermitian and has strictly positive eigenvalues.

1.10 **Convexity of $S(\mathcal{H})$:** This question is about the set of density matrices $S(\mathcal{H}) \subset \mathrm{Lin}(\mathcal{H})$. The goal is to prove Lemma 1.10.

(a) Prove that $S(\mathcal{H})$ is a convex subset of the vector space $\mathrm{Lin}(\mathcal{H})$.
(b) Show that any state which is *not* pure is not extremal.
(c) Show that every pure state is extremal. *Hint: Suppose that $\rho = |\psi\rangle\langle\psi|$ is pure and can be written as a convex combination, compute $1 = \mathrm{tr}[\rho] = \langle\psi|\rho|\psi\rangle$ and use the Cauchy-Schwarz inequality.*

These statements together prove Lemma 1.10.

1.11 **Functions of operators:**

(a) Show that if $P \in \mathrm{PSD}(\mathcal{H})$, there is a *unique* PSD operator $\sqrt{P}$ which squares to P.
(b) Let $F : \mathbb{C} \to \mathbb{C}$ be a polynomial

$$F(t) = \sum_{k=0}^{d} c_k t^k \ .$$

We can extend the definition of F to matrices $M \in \mathrm{Lin}(\mathcal{H})$, by setting

$$\tilde{F}(M) = \sum_{k=0}^{d} c_k M^k \ .$$

Show that, if M is diagonalisable with eigenvalues $\{\lambda_i\}_{i=1}^r$ then $\tilde{F}(M)$ is also diagonalisable with eigenvalues $\{F(\lambda_i)\}_{i=1}^r$. Argue that $\tilde{F}(M)$ is equal to $F(M)$ as defined in Eq. (A.4). [MW: The definition there is only for Hermitian M. Maybe we should state it for normal operators, to also cover unitaries.]
(c) Argue that if $M \in \mathrm{Lin}(\mathcal{H})$ is Hermitian, then $U = e^{i\theta M}$ is unitary for every $\theta \in \mathbb{R}$.

1.12 **Projections:** Let $P \in \mathrm{Lin}(\mathcal{H})$ be a linear map. Show that P is a projection (a Hermitian operator satisfying $P^2 = P$) if and only if P can be written as

$$P = \sum_i |\psi_i\rangle\langle\psi_i|$$

where the $|\psi_i\rangle$ are a basis for $\mathrm{im}(P)$.

1.13 **Extending an isometry to a unitary:** Suppose that $\mathcal{H}$ and $\mathcal{K}$ are Hilbert spaces, where $\mathcal{H}$ is a subspace of $\mathcal{K}$, and $V : \mathcal{H} \to \mathcal{K}$ is an isometry. Show that there exists a unitary $U \in \mathrm{U}(\mathcal{K})$ such that $U|\psi\rangle = V|\psi\rangle$ for all $|\psi\rangle \in \mathcal{H}$.

1.14 **Functionals of matrices:** Show that if $f : \mathrm{Lin}(\mathcal{H}) \to \mathbb{C}$ is a linear function, there must exist a unique $X \in \mathrm{Lin}(\mathcal{H})$ such that $f(Y) = \mathrm{tr}[XY]$. Next, show that f maps positive operators to $\mathbb{R}_{\geq 0}$ if and only if $X \geq 0$.

1.15 **Decompositions of operators:**

(a) Show that you can write any Hermitian $M \in \mathrm{Lin}(\mathcal{H})$ as $M = P - Q$ where $P, Q \in \mathrm{PSD}(\mathcal{H})$.
(b) Show that you can write any operator $M \in \mathrm{Lin}(\mathcal{H})$ as $M = N_1 + iN_2$ where N_1 and N_2 are Hermitian.

1.16 **Criteria for positive semidefiniteness:** Prove Lemma A.2. *Hint: To prove the implication (a) $\Rightarrow$ (b) you may find it useful to take the complex conjugate of the expression in (a), and consider the matrix $M = i(P - P^\dagger)$. What special property does M have?*

1.17 **The most general measurement:** Prove Lemma 1.15.

Hint: First use Exercise 1.15 to show that, for any fixed x, $\rho \mapsto p_\rho(x)$ extends to a linear map, and then use Exercise 1.14.

1.18 **Pure states as projection operators:** Prove Lemma 1.6.

1.19 **Uncertainty relation:** Given a two-outcome measurement $\mu \in \mathrm{Meas}(\mathcal{H}, \{0,1\})$ and a state $\rho \in \mathrm{S}(\mathcal{H})$, define the *bias* by

$$\beta(\rho) = \Big|\mathrm{tr}[\mu(0)\rho] - \mathrm{tr}[\mu(1)\rho]\Big|.$$

(a) Show that $\beta \in [0,1]$, that $\beta = 1$ iff the measurement outcome is certain, and that $\beta = 0$ iff the outcome is uniformly random (for the given measurement and state).

In class, we discussed how to measure a qubit in the standard (Z) basis $|0\rangle$,$|1\rangle$ and also in the X basis $|+\rangle$, $|-\rangle$. Let β_Z and β_X denote the bias for these two measurements.

(b) Compute $\beta_Z(\rho)$ and $\beta_X(\rho)$ in terms of the Bloch vector of the qubit state ρ.

(c) Show that $\beta_Z^2(\rho) + \beta_X^2(\rho) \leq 1$. Why is this called an *uncertainty relation*?

1.20 **Measurements and the Bloch sphere:** This question is concerned with measurements of a qubit states.

(a) If $\vec{r} = (x, y, z)$ with $\|\vec{r}\| = 1$, show that the pure quantum states corresponding to $\vec{r}$ and $-\vec{r}$ are orthogonal, and that they are eigenvectors with eigenvalues ±1 of the operator

$$xX + yY + zZ.$$

(b) Conclude that $\mu_{\vec{r}}$ as given in Eq. (1.4) defines a two-outcome measurement.

(c) Now, consider a state parametrized by $\vec{s} = (x', y', z')$ in the Bloch ball. Show that the probability of obtaining outcome 0 from the measurement $\mu_{\vec{r}}$ is given by

$$p(0) = \frac{1}{2} + \frac{1}{2}\vec{r} \cdot \vec{s} = \frac{1}{2} + \frac{1}{2}(xx' + yy' + zz').$$

Hint: Use that $\mathrm{tr}[MN] = 0$ *if M and N are different Pauli operators.*

(d) Consider the following set-up: You have access to a source producing an unknown state ρ, and you can do arbitrary qubit measurements. You can repeat many times, with different measurement settings; each time you receive the same state ρ. You would like to learn the state ρ, that is you would like to learn $\vec{s} = (x', y', z')$ such that $\rho \approx \rho(\vec{s})$. Suppose you measure N times along the z-axis (so using $\vec{r} = (0,0,1)$), obtaining N_0 times outcome 0 and N_1 times outcome N_1. What values do you expect for N_0/N and N_1/N for large N?

(e) Argue that a reasonable estimate for z' is given by

$$\frac{N_0 - N_1}{N}.$$

(f) Describe a procedure to estimate the unknown state ρ.

1.21 **Expectation values of observables:** The following question reviews measurements from a slightly different perspective from this course. If you have studied physics you may be familiar with this approach. An *observable* on a quantum system is by definition a Hermitian operator on the corresponding Hilbert space $\mathcal{H}$.

(a) Let $\mu \in \operatorname{Meas}(\mathcal{H}, \Omega)$ be a *projective* measurement with outcomes in the real numbers, i.e., a finite subset $\Omega \subseteq \mathbb{R}$. Show that the following operator is an observable:

$$\mathcal{O} = \sum_{x \in \Omega} x \, \mu(x) \tag{1.5}$$

In fact, this is always an eigendecomposition, but you need not prove this.

(b) Argue that, conversely, any observable can be written as in Eq. (1.5) for some suitable μ.

(c) Now suppose that the system is in state ρ and we perform the measurement μ. Show that the *expectation value* of the measurement outcome is given by $\operatorname{tr}[\rho\mathcal{O}]$.
For a pure state $\rho = |\psi\rangle\langle\psi|$, this can also be written as $\langle\psi|\mathcal{O}|\psi\rangle$. Do you recognize these formulas from your quantum mechanics class?

(d) Consider an arbitrary qubit observable $\mathcal{O} = tI + s_x X + s_y Y + s_z Z$. Compute its expectation value in a state with Bloch vector $\vec{r}$.

Lecture 2

Multiple quantum systems

Concept	Math translation
Joint system of Alice and Bob	Tensor product Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$.
Restricting to a subsystem	The partial trace $\rho_A = \operatorname{tr}_B[\rho_{AB}]$.
Every state can be realized as reduced state of a pure state.	Lemma 2.10: for any state ρ_A there exists a purification $\lvert\phi_{AR}\rangle$ such that ρ_A is its reduced state on A.
Entanglement for pure states	Theorem 2.13: Schmidt decomposition $$\lvert\phi_{AB}\rangle = \sum_i s_i \lvert e_i\rangle \lvert f_i\rangle.$$
Entanglement for mixed states	Definition 2.15: states which can *not* be written as a mixture of product states.

In the previous lecture we have seen how to describe a quantum system. From the perspective of information theory we may think of a device, preparing some quantum state, and of a measurement device. The natural next step is that our device may be a source producing a sequence of quantum states. Or, we may have multiple devices, perhaps in different laboratories, all generating quantum states. This begs the question: how do we describe *multiple (qu)bits*? In the classical case, this is rather intuitive: suppose we have n bits, then we can describe these as strings of bits of length n. The joint system assigns a probability to each string, for which we in total have 2^n possibilities. More generally, if we have classical random variables with outcome sets $\Sigma_1, \dots, \Sigma_n$, then their joint distribution is a probability distribution on the product set

$$\Sigma_1 \times \Sigma_2 \times \ldots \times \Sigma_n = \{(x_1, \dots, x_n) : x_j \in \Sigma_j\}.$$

The generalization of this to quantum systems is given by the *tensor product*. This is defined in detail in Appendix A.2. The easiest way to think about the tensor product of Hilbert spaces is that it has a product basis. If $\mathcal{H}_j$ has a basis $\lvert x_j\rangle$ labeled by $x_j \in \Sigma_j$, then their tensor product

$$\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_n$$

has a product basis

$$\lvert x_1, \dots, x_n\rangle = \lvert x_1\rangle \otimes \cdots \otimes \lvert x_n\rangle$$

labeled by tuples or strings $x = (x_1, \dots, x_n) \in \Sigma_1 \times \Sigma_2 \times \ldots \times \Sigma_n$. By counting the number of basis vectors we see that $\dim(\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_n) = \dim(\mathcal{H}_1) \cdots \dim(\mathcal{H}_n)$.

Axiom 4 (Multiple systems)**.** If we have n quantum systems, with Hilbert spaces $\mathcal{H}_1, \dots, \mathcal{H}_n$, then the joint system has associated Hilbert space $\mathcal{H}_1 \otimes \dots \otimes \mathcal{H}_n$.

Example 2.1. If we have two qubits, the joint Hilbert space $\mathbb{C}^2 \otimes \mathbb{C}^2$ is 4-dimensional and has the standard product basis

$$|00\rangle, |01\rangle, |10\rangle, |11\rangle$$

labeled by the four bitstrings of length two. We typically order the basis lexicographically (as we did here). Similarly, the Hilbert space of n qubits is $(\mathbb{C}^2)^{\otimes n}$ has the standard product basis

$$|x_1 \dots x_n\rangle \quad \text{for } x_1, \dots, x_n \in \{0, 1\}$$

labeled by the bitstrings of length n. This Hilbert space has dimension $\dim((\mathbb{C}^2)^{\otimes n}) = 2^n$. In general, the dimension of the joint Hilbert space grows exponentially in the number of systems.

To keep track of different quantum systems (which are also called quantum registers or variables), we will label them as $A, B, C, \dots$ and denote the associated Hilbert spaces as $\mathcal{H}_A, \mathcal{H}_B, \mathcal{H}_C, \dots$ and their dimensions as $|A|, |B|, |C|, \dots$ Often, these quantum systems come with distinguished standard bases, labeled by sets Σ_A, Σ_B, $\dots$. For quantum systems which are always in a classical state (for instance because we use them to keep track of measurement outcomes), we will often use the letters $X, Y, \dots$ to label the system to clarify the different interpretation of the systems. It will often be helpful in reasoning about such quantum systems to antropomorphize them, so we will refer to Alice, Bob, and Charlie as holding the respective quantum systems A, B, C in their quantum computers or laboratories. Formulated this way, Axiom 4 states that if Alice and Bob have quantum systems A and B with Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, respectively, their joint system AB has Hilbert space $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$. We will also write $\mathrm{Lin}(A) = \mathrm{Lin}(\mathcal{H}_A)$, $\mathrm{S}(A) = \mathrm{S}(\mathcal{H}_A)$, $\mathrm{PSD}(A) = \mathrm{PSD}(\mathcal{H}_A)$ etc. We use subscripts more generally to indicate the systems that mathematical objects relate to. For instance, we write $\rho_{AB} \in \mathrm{S}(AB)$ for a quantum state shared by Alice and Bob, $\mu_A \in \mathrm{Meas}(A, \Omega)$ for a measurement on Alice's quantum system, and $M_B \in \mathrm{Lin}(B)$ for an operator acting on Bob's Hilbert space, and we use $\Sigma_{AB} = \Sigma_A \times \Sigma_B$ to label the product basis of a joint system.

Concept	Notation
Agent	Alice, Bob, Charlie, $\dots$
Quantum system	$A, B, C, \dots$
Hilbert space	$\mathcal{H}_A, \mathcal{H}_B, \mathcal{H}_C, \dots$
Quantum state	$\rho_A, \rho_B, \rho_C, \rho_{AB}, \rho_{ABC}, \dots$
Alphabet of symbols	$\Sigma_A, \Sigma_B, \Sigma_C, \dots$
Basis	$\|a\rangle, \|b\rangle, \|c\rangle, \dots$ for $a \in \Sigma_A, b \in \Sigma_B, c \in \Sigma_C, \dots$
Dimensions	$\|A\|, \|B\|, \|C\|, \dots$ (i.e., $\|A\| := \dim(\mathcal{H}_A) = \|\Sigma_A\|$)
Classical systems	$X, Y, \dots$ (with alphabets Σ_X, Σ_Y, etc.)

What are the possible states of a joint system? An easy way to construct is by taking the tensor product of states of the subsystems. For concreteness, we define this in the case of two subsystems.

Definition 2.2. Let A and B be quantum systems. States of the form $\rho = \rho_A \otimes \rho_B$ for $\rho_A \in \mathrm{S}(A)$ and $\rho_B \in \mathrm{S}(B)$ are called *product states.* A state which is not a product state is called *correlated.*

For instance, if we have a pure state $|\psi_A\rangle \otimes |\phi_B\rangle$ then this is a product state. We will sometimes use the abbreviation

$$|\psi_A\rangle|\phi_B\rangle = |\psi_A\rangle \otimes |\phi_B\rangle,$$

which is quite common in the literature.

If there are more than two systems then we extend the definition in the obvious way by saying that a state $\rho_{A_1 \ldots A_n} \in \mathrm{S}(A_1 \ldots A_n)$ is a product state (between systems $A_1, \ldots, A_n$) if it can be written as $\rho_{A_1} \otimes \ldots \otimes \rho_{A_n}$ for $\rho_{A_i} \in \mathrm{S}(A_i)$.

We can also build joint states from joint probability distributions. Given a *joint probability distribution* $p_{XY} \in \mathrm{P}(XY)$, which associates a probability $p_{XY}(x, y)$ to each pair $(x, y) \in \Sigma_X \times \Sigma_Y$, the corresponding classical state of XY is

$$\rho_{XY} = \sum_{x,y} p_{XY}(x, y)|x, y\rangle\langle x, y| = \sum_{x,y} p_{XY}(x, y)|x\rangle\langle x| \otimes |y\rangle\langle y|,$$

and any classical joint state is of this form. The state ρ_{XY} is a product state if and only if the two random variables X and Y are independent under the distribution p_{XY}. See Exercise 2.4. Thus we can think of product states as the quantum generalization of independence in probability theory. Most quantum states are neither classical nor product states.

We will now give a few concrete example of quantum states on two-qubit systems. These examples will be useful to illustrate various concepts throughout this book, as archetypes of different types of correlations between quantum systems.

Example 2.3. Suppose Alice and Bob both have a qubit, so $\mathcal{H}_A = \mathcal{H}_B = \mathbb{C}^2$. Then their joint Hilbert space is $\mathcal{H}_{AB} = \mathbb{C}^2 \otimes \mathbb{C}^2$. They could for example share one of the basis states $|00\rangle$, $|01\rangle$, $|10\rangle$, or $|11\rangle$. These are pure product states. Another example would be if they both have a $|+\rangle$ state

$$\begin{aligned} |+\rangle_A \otimes |+\rangle_B &= \left(\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\right) \otimes \left(\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\right) \\ &= \frac{1}{2}\left(|00\rangle + |01\rangle + |10\rangle + |11\rangle\right). \end{aligned}$$

This is also a pure product state (but not a classical state).

A classical state that is not a product state is *maximally correlated state*, which corresponds to two bits which are both equal to `0` or both equal to `1`, with probability $\frac{1}{2}$ each:

$$\sigma_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|) = \frac{1}{2}(|0\rangle\langle 0| \otimes |0\rangle\langle 0| + |1\rangle\langle 1| \otimes |1\rangle\langle 1|),$$

We can write this out as a matrix in the lexicographically ordered basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$:

$$\sigma_{AB} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A famous state that is neither classical nor a product state is the *maximally entangled* state (which will come back to haunt us throughout the whole book!). It is given by

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

The density matrix is

$$\rho_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}| = \frac{1}{2}(|00\rangle\langle 00| + |00\rangle\langle 11| + |11\rangle\langle 00| + |11\rangle\langle 11|).$$

We can write this out as a matrix in the lexicographically ordered basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$:

$$\rho_{AB} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Recall that $|00\rangle = |0\rangle \otimes |0\rangle$, so $|00\rangle\langle 00| = (|0\rangle \otimes |0\rangle)(\langle 0| \otimes \langle 0|) = |0\rangle\langle 0| \otimes |0\rangle\langle 0|$ etc. This means that we can equivalently write the density matrix as

$$\rho_{AB} = \frac{1}{2}\left(|0\rangle\langle 0| \otimes |0\rangle\langle 0| + |0\rangle\langle 1| \otimes |0\rangle\langle 1| + |1\rangle\langle 0| \otimes |1\rangle\langle 0| + |1\rangle\langle 1| \otimes |1\rangle\langle 1|\right).$$

The maximally entangled state is *not* a product state! We will prove this later this lecture.

2.1 The partial trace

Suppose we have a quantum state ρ_{AB} on quantum systems A and B shared by Alice and Bob. We imagine that Alice and Bob have separated laboratories, so Alice has no access to the B system and can only manipulate the A system. Two closely related questions arise.

(a) Given that Alice only has access to system A, what does the set of measurements Alice can perform look like?

(b) The A system Alice has control over is *itself* a quantum system and there should be a description of a state ρ_A only on the subsystem A, i.e., without reference to B.

Let us first describe what measurements Alice (and Bob) can do locally. Given a measurement $\mu_A = \{\mu_A(x)\}_{x\in\Omega} \in \mathrm{Meas}(A,\Omega)$ that Alice would like to perform on her system, a natural candidate is to extend it to the following measurement on AB which 'does nothing' on the B system:

$$\mu_A \otimes \mathbb{1}_B := \{\mu_A(x) \otimes \mathbb{1}_B : x \in \Omega\}.$$

This is indeed a measurement on the whole system since

$$\sum_{x\in\Omega} \mu_A(x) \otimes \mathbb{1}_B = \left(\sum_{x\in\Omega} \mu_A(x)\right) \otimes \mathbb{1}_B = \mathbb{1}_A \otimes \mathbb{1}_B = \mathbb{1}_{AB}$$

and tensor products of positive operators are positive by Lemma A.6. This answers the first question (a). More generally, if Alice and Bob each separately perform measurements μ_A and μ_B with outcome sets Ω_1 and Ω_2, respectively, this corresponds to a measurement on AB given by

$$\mu_A \otimes \mu_B := \{\mu_A(x_1) \otimes \mu_B(x_2) : (x_1, x_2) \in \Omega_1 \times \Omega_2\}$$

In Exercise 2.7 you will verify that the outcome probabilities for Alice do not depend on the choice of measurement for Bob.

Let us now use this to reconstruct what is the appropriate description for Alice's state ρ_A if the overall system is in some joint state ρ_{AB}. We would clearly like to have

$$\mathrm{tr}[\mu_A(x)\rho_A] = \mathrm{tr}[(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}]$$

for any possible measurement operator $\mu_A(x)$. We compute this trace by choosing bases $|a\rangle$ of $\mathcal{H}_A$ and $|b\rangle$ of $\mathcal{H}_B$ and using the product basis $|ab\rangle = |a\rangle \otimes |b\rangle$ to compute the trace:

$$\begin{aligned}\mathrm{tr}[(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}] &= \sum_{a\in\mathcal{A}, b\in\mathcal{B}} \langle ab|(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}|ab\rangle \\ &= \sum_{a,b} \langle a|\mu_A(x)(\mathbb{1}_A \otimes \langle b|)\rho_{AB}(\mathbb{1}_A \otimes |b\rangle)|a\rangle \\ &= \sum_a \langle a|\mu_A(x)\left(\sum_b (\mathbb{1}_A \otimes \langle b|)\rho_{AB}(\mathbb{1}_A \otimes |b\rangle)\right)|a\rangle.\end{aligned}$$

Here, $\mathbb{1}_A \otimes |b\rangle$ is an operator from $\mathcal{H}_A$ to $\mathcal{H}_A \otimes \mathcal{H}_B$, and $\mathbb{1}_A \otimes \langle b|$ is its adjoint, mapping $\mathcal{H}_A \otimes \mathcal{H}_B$ to $\mathcal{H}_A$. See Remark A.7 for more details. We have used that $|ab\rangle = |a\rangle \otimes |b\rangle = (\mathbb{1}_A \otimes |b\rangle)|a\rangle$ (verify for yourself that this equation makes sense!). Thus, if we define

$$\rho_A := \sum_b (\mathbb{1}_A \otimes \langle b|)\rho_{AB}(\mathbb{1}_A \otimes |b\rangle) \tag{2.1}$$

then this operator satisfies exactly the desired relation:

$$\operatorname{tr}[(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}] = \operatorname{tr}[\mu_A(x)\rho_A]. \tag{2.2}$$

The expression in Eq. (2.1) looks similar to a trace, except we only 'trace out' the B system. This motivates the following definition:

Definition 2.4 (Partial trace and reduced state)**.** Suppose A and B are systems with Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ and choose a basis $|b\rangle$ for $\mathcal{H}_B$. Let $M_{AB} \in \operatorname{Lin}(AB)$. Then we define the *partial trace* over B of M_{AB} to be

$$\operatorname{tr}_B[M_{AB}] = \sum_b (\mathbb{1}_A \otimes \langle b|) M_{AB} (\mathbb{1}_A \otimes |b\rangle).$$

If $\rho_{AB} \in \mathrm{S}(AB)$, then we write $\rho_A = \operatorname{tr}_B[\rho_{AB}]$. We call $\rho_A \in \mathrm{S}(A)$ the *reduced state* of ρ_{AB} on A.

As you can verify in Exercise 2.8, the reduced state ρ_A is not only consistent with reproducing the correct measurement outcomes when measuring on A as in Eq. (2.2), it is also uniquely determined by this requirement and hence the only prescription which leads to the correct outcomes for *all* measurements. Thus we have answered the second question (b).

Let us make the partial trace more explicit. If we also choose a basis $|a\rangle$ for $\mathcal{H}_A$, then it follows directly from the definition that the matrix entries of the partial trace are given by

$$\langle a| \operatorname{tr}_B[M_{AB}]|a'\rangle = \langle a| \sum_b (\mathbb{1}_A \otimes \langle b|) M_{AB} (\mathbb{1}_A \otimes |b\rangle)|a'\rangle = \sum_b \langle ab|M_{AB}|a'b\rangle.$$

In other words, if we expand M_{AB} in the product basis as

$$M_{AB} = \sum_{a,a' \in \mathcal{A}} \sum_{b,b' \in \mathcal{B}} M_{ab,a'b'} |ab\rangle\langle a'b'|,$$

then the partial trace is given by

$$\operatorname{tr}_B[M_{AB}] = \sum_{a,a',b} M_{ab,a'b} |a\rangle\langle a'| = \sum_{a,a'} \left(\sum_b M_{ab,a'b} \right) |a\rangle\langle a'|. \tag{2.3}$$

For tensor product operators $M_{AB} = N_A \otimes O_B$, the partial trace is given by the natural formula

$$\operatorname{tr}_B[N_A \otimes O_B] = N_A \operatorname{tr}[O_B] = \operatorname{tr}[O_B]\, N_A. \tag{2.4}$$

This follows from Eq. (2.3), since in this case $M_{ab,a'b'} = N_{a,a'} O_{b,b'}$ and $\sum_b M_{ab,a',b} = M_{a,a'} \operatorname{tr}[O_B]$. Since every operator can be written as a linear combination of tensor product operators, this formula is sufficient to compute partial traces of arbitrary operators. Moreover, it shows that our notation for the reduced states is compatible with our notation for product states: if $\rho_{AB} = \rho_A \otimes \rho_B$ then ρ_A and ρ_B are the reduced states of A and B, respectively.

We have the following basic properties:

Lemma 2.5 (Properties of the partial trace)**.** *(a) The map* $\operatorname{tr}_B \colon \operatorname{Lin}(AB) \to \operatorname{Lin}(A)$ *is linear.*

(b) For any $N_A \in \operatorname{Lin}(A)$ *and* $M_{AB} \in \operatorname{Lin}(AB)$*, we have*

$$\operatorname{tr}[(N_A \otimes \mathbb{1}_B)M_{AB}] = \operatorname{tr}[N_A \operatorname{tr}_B[M_{AB}]]$$

(c) The partial trace does not depend on the choice of basis $\mathcal{B}$*.*

(d) For any $M_{AB} \in \operatorname{Lin}(AB)$ *we have* $\operatorname{tr}[\operatorname{tr}_B[M_{AB}]] = \operatorname{tr}[M_{AB}]$*.*

(e) If $P_{AB} \in \mathrm{PSD}(AB)$*, then* $\operatorname{tr}_B[P_{AB}] \in \mathrm{PSD}(A)$*.*

Proof. (a) The formula in the definition of $\operatorname{tr}_B[M_{AB}]$ is linear in M_{AB}.

(b) This is precisely the same calculation that led to Eq. (2.2), with M_{AB} in place of ρ_{AB} and with N_A in place of $\mu_A(x)$.

(c) From (b) we know that

$$\operatorname{tr}[(N_A \otimes \mathbb{1}_B)M_{AB}] = \operatorname{tr}[N_A \operatorname{tr}_B[M_{AB}]].$$

Since the left-hand side does not depend on a choice of basis, neither does the right-hand side. Since the equation holds for all $N_A \in \operatorname{Lin}(A)$, by Exercise 1.14 this completely determines $\operatorname{tr}_B[M_{AB}]$ which is therefore also independent of a choice of basis.

(d) This follows from (b) using $N_A = \mathbb{1}_A$.

(e) By Lemma A.2 it suffices to show that

$$\operatorname{tr}[Q_A \operatorname{tr}_B[P_{AB}]] \geq 0 \text{ for all } Q_A \in \operatorname{PSD}(A).$$

Now, if $Q_A \geq 0$, then $Q_A \otimes \mathbb{1}_B \geq 0$ by Lemma A.6 so applying first (b) and then Lemma A.2

$$\operatorname{tr}[Q_A \operatorname{tr}_B[P_{AB}]] = \operatorname{tr}[(Q_A \otimes \mathbb{1}_B)P_{AB}] \geq 0. \qquad \square$$

Note that (b) is a generalization of our observations on measurements on reduced systems. Moreover, (d) and (e) prove that if $\rho_{AB} \in \mathrm{S}(AB)$, then $\rho_A = \operatorname{tr}_B[\rho_{AB}] \in \mathrm{S}(A)$.

Example 2.6. Suppose Alice and Bob share a maximally entangled state $|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. What is Alice's reduced state? Using either Eq. (2.3) or Eq. (2.4), we see that the partial trace over B is given by

$$\begin{aligned}\rho_A = \operatorname{tr}_B[\rho_{AB}] &= \frac{1}{2}\operatorname{tr}_B[|00\rangle\langle 00| + |00\rangle\langle 11| + |11\rangle\langle 00| + |11\rangle\langle 11|] \\ &= \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|) = \frac{1}{2}\mathbb{1}_A\end{aligned}$$

For example, $|00\rangle\langle 00| = |0\rangle\langle 0| \otimes |0\rangle\langle 0|$ contributes $|0\rangle\langle 0|$, while $|00\rangle\langle 11| = |0\rangle\langle 1| \otimes |0\rangle\langle 1|$ maps to zero under the partial trace since we have an off-diagonal term on B. We see that the reduced state of the maximally entangles state is the *maximally mixed state* on Alice's system! However, this state *also* describes the situation where Alice has a classical bit, with equal probability equal to zero or one. If Alice can not communicate with Bob in any way, she can not see the difference between these two scenarios!

We learn an important lesson from this example: if we have a state which is not pure, this may either arise through some form of classical randomness (as a classical mixture of pure states) but it may also reflect ignorance of another quantum system while the global state is pure.

In conclusion, we see *three sources of mixed states* in our formalism of quantum mechanics:

Concept	Math translation
Probabilistic mixtures	If we receive ρ_x with probability $p(x)$, the state is described by $\rho = \sum_x p(x)\rho_x$.
Restricting to a subsystem	Even if the full state $|\psi_{AB}\rangle$ is a pure state, the reduced state ρ_A can be mixed.
Measurement	If we perform a measurement μ on ρ we get outcome x with probability $p(x) = \operatorname{tr}[\mu(x)\rho]$. This probability distribution can be described by the classical state $\sigma = \sum_x p(x)|x\rangle\langle x|$.

We will often need to manipulate reduced states and partial traces. Here are some more useful properties of the partial trace. The proof is Exercise 2.11.

Lemma 2.7 (More properties of the partial trace)**.** *Let* $M_{AB} \in \mathrm{Lin}(AB)$.

(a) For any $N_B \in \mathrm{Lin}(B)$ *we have* $\mathrm{tr}_B[(\mathbb{1}_A \otimes N_B)M_{AB}] = \mathrm{tr}_B[M_{AB}(\mathbb{1}_A \otimes N_B)]$.

(b) For unitary $U_B \in \mathrm{U}(B)$ *we have* $\mathrm{tr}_B[(\mathbb{1}_A \otimes U_B)M_{AB}(\mathbb{1}_A \otimes U_B^\dagger)] = \mathrm{tr}_B[M_{AB}]$.

(c) $\mathrm{tr}_B[M_A \otimes M_B] = \mathrm{tr}[M_B]M_A$ *for* $M_A \in \mathrm{Lin}(A)$, $M_B \in \mathrm{Lin}(B)$. *[MW: This is now stated in Eq. (2.4), so we can remove it soon.]*

(d) For $N_1 \in \mathrm{Lin}(A, C_1)$ *and* $N_2 \in \mathrm{Lin}(C_2, A)$ *we have*

$$\mathrm{tr}_B[(N_1 \otimes \mathbb{1}_B)M_{AB}(N_2 \otimes \mathbb{1}_B)] = N_1 \,\mathrm{tr}_B[M_{AB}]N_2.$$

(e) If we have quantum systems A, B, *and* C, *and an operator* $M_{ABC} \in \mathrm{Lin}(ABC)$ *then*

$$\mathrm{tr}_B[\mathrm{tr}_C[M_{ABC}]] = \mathrm{tr}_{BC}[M_{ABC}].$$

Marginal distributions

For classical states, the partial trace reduces to a familiar concept. Suppose we have a classical joint state

$$\rho_{XY} = \sum_{x,y} p_{XY}(x,y)|x,y\rangle\langle x,y| = \sum_{x,y} p_{XY}(x,y)|x\rangle\langle x| \otimes |y\rangle\langle y|,$$

where $p_{XY} \in \mathrm{P}(XY)$ is some joint probability distribution on $\Sigma_X \times \Sigma_Y$. If we compute the reduced density matrix on X by taking the partial trace over Y, we get

$$\rho_X = \mathrm{tr}_Y[\rho_{XY}] = \sum_x \left(\sum_y p_{XY}(x,y) \right) |x\rangle\langle x| = \sum_x p_X(x)|x\rangle\langle x|,$$

where $p_X \in \mathrm{P}(\Sigma_X)$ is the *marginal distribution* of random variable X, which is given by

$$p_X(x) = \sum_y p_{XY}(x,y)$$

for $x \in \Sigma_X$. The interpretation of this formula is clear: the total probability of outcome x is given by summing over all outcomes (x,y) for arbitrary y.

As mentioned earlier, the two random variables X and Y are independent, meaning that $p_{XY}(x,y) = p_X(x)p_Y(x)$, precisely if and only if ρ_{XY} is a product state, meaning $\rho_{XY} = \rho_X \otimes \rho_Y$.

Another important notion in classical probability theory is that of a *conditional probability.* In this case we ask what is the probability of outcome x *given* that we already know the outcome on Y is y. This probability is denoted by $p_{X|y}$ and is computed by

$$p_{X|Y=y}(x) = \frac{p_{XY}(x,y)}{p_Y(y)} = \frac{p_{XY}(x,y)}{\sum_x p_{XY}(x,y)}. \tag{2.5}$$

Example 2.8. We let X and Y be two bits (i.e., they have outcomes in $\{0,1\}$). We can describe a joint probability distribution as a table of values $p_{XY}(x,y)$, for instance

	$y=0$	$y=1$
$x=0$	$1/5$	$1/10$
$x=1$	$1/5$	$1/2$

Then the marginal distribution on X is computed by summing over the rows:

$$p_X(0) = p_{XY}(0,0) + p_{XY}(0,1) = 1/5 + 1/10 = 3/10$$
$$p_X(1) = p_{XY}(1,0) + p_{XY}(1,1) = 1/5 + 1/2 = 7/10.$$

Compute the marginal distribution p_Y yourself by summing over columns. Are X and Y independent?

For the conditional probability, we may for instance compute the conditional probability given $y=1$:

$$p_{X|Y=1}(0) = \frac{p_{XY}(0,1)}{p_Y(1)} = \frac{p_{XY}(0,1)}{p_{XY}(0,1)+p_{XY}(1,1)} = 1/6$$
$$p_{X|Y=1}(1) = \frac{p_{XY}(1,1)}{p_Y(1)} = \frac{p_{XY}(1,1)}{p_{XY}(0,1)+p_{XY}(1,1)} = 5/6.$$

In the standard product basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$ we can write the density matrix corresponding to p_{XY}

$$\rho_{XY} = \frac{1}{10}\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}$$

with reduced density matrices

$$\rho_X = \frac{1}{10}\begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix} \qquad \rho_Y = \frac{1}{5}\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

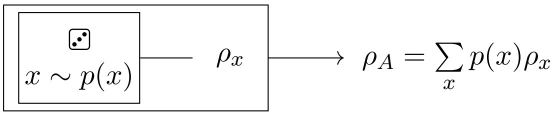
An important complication of quantum information is that there is no direct generalization of conditional probabilities for general quantum states. We will come back to this ominuous remark in [MW: ref].

2.2 Reference systems and purifications

If Alice possesses a quantum state, then it will be useful to be able to see this as the reduced state of a state on a larger system. For instance, if we have a device producing a state according to a classical process, it may be useful to separate the classical randomness and keep track in another reference system of the outcomes of this classical process. To be more precise, if we have a decomposition of a quantum state ρ_A as

$$\rho_A = \sum_{x \in \Omega} p(x) \rho_x \tag{2.6}$$

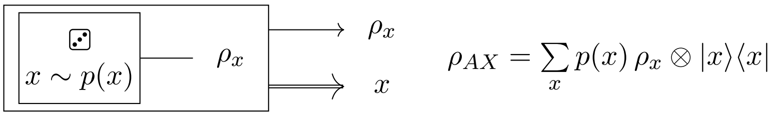
for some states $\rho_x \in \mathrm{S}(A)$ and a probability distribution p on a set Ω, we can interpret this as a probabilistic mixture of quantum states where we have the state ρ_x with probability $p(x)$. This illustrated by the following quantum information source, which samples x and outputs ρ_x (but not x): [MW: I find the figures a bit confusing. Also, there is some strange vertical space before the figure that I can't get rid of.]



We may also introduce a (classical) reference system $X = \mathbb{C}^\Omega$ and consider the joint state

$$\rho_{AX} = \sum_{x \in \Omega} p(x)\, \rho_x \otimes |x\rangle\langle x|_X. \tag{2.7}$$

This state models a similar situation, but now we receive the value of x as well:


$x \sim p(x)$
ρ_x
ρ_x
x
$\rho_{AX} = \sum_x p(x)\, \rho_x \otimes |x\rangle\langle x|$

We call a collection $\{p(x), \rho_x\}_x$ and the associated state in Eq. (2.7) an *ensemble* of quantum states. It is easy to verify that this state is such that taking the partial trace over X yields ρ_A. A final distinction is that once we *actually* receive a specific outcome x, then the state must be ρ_x (if this is confusing, think of the analogous situation in classical probability: once we have actually rolled a die and we see the outcome 3, then the die is in the state 3 with probability 1).

One may call the register X 'side information': if Alice does not have the register X, the state is given by ρ_A, but if she also has the side information, she may learn which specific state $|\psi_x\rangle$ she has. We can also use *quantum* side information. As a special case, it turns out to be extremely useful to consider quantum side information such that the total state is pure.

Definition 2.9. Given $\rho_A \in \mathrm{S}(\mathcal{H}_A)$, a *purification* of ρ_A is a pure state $|\phi_{AR}\rangle \in \mathcal{H}_A \otimes \mathcal{H}_R$ which is such that

$$\mathrm{tr}_R[|\phi_{AR}\rangle\langle\phi_{AR}|] = \rho_A.$$

The system R is called a *reference system* or *purifying system*. We will refer both to $|\phi_{AR}\rangle$ and $\rho_{AR} := |\phi_{AR}\rangle\langle\phi_{AR}|$ as a *purification* of ρ_A.

Such purifications always exist. This is important for two reasons:

(a) Conceptually, it means the formalism for mixed quantum states we introduced in Lecture 1 is not strictly more general than the formalism for pure quantum states: any mixed state can be understood as the state of a subsystem of a larger system that is in a pure state.

(b) In many situations it is also simply convenient to take a purification of a mixed state and reason about this pure state (so in this case the purification is just an artifical construction which need not reflect the 'true' quantum state).

Lemma 2.10. *Every $\rho_A \in \mathrm{S}(\mathcal{H}_A)$ has a purification. The dimension $|R|$ of the purifying system can be taken to be* $\operatorname{rank}(\rho_A)$*.*

Proof. Let $r = \operatorname{rank}(\rho_A)$ and let

$$\rho_A = \sum_{j=0}^{r-1} p_j |e_j\rangle\langle e_j|$$

be a spectral decomposition of ρ_A. Let $\mathcal{H}_R = \mathbb{C}^r$ and let

$$|\phi_{AR}\rangle = \sum_{j=0}^{r-1} \sqrt{p_j}\, |e_j\rangle \otimes |j\rangle.$$

Then we can easily verify using Eq. (2.4) that

$$\begin{aligned}
\operatorname{tr}_R[|\phi_{AR}\rangle\langle\phi_{AR}|] &= \operatorname{tr}_R\left[\sum_{j,k=0}^{r-1} \sqrt{p_j p_k}|e_j\rangle\langle e_k| \otimes |j\rangle\langle k|\right] \\
&= \sum_{j,k=0}^{r-1} \sqrt{p_j p_k}|e_j\rangle\langle e_k| \otimes \operatorname{tr}[|j\rangle\langle k|] \\
&= \sum_{j=0}^{r-1} p_j |e_j\rangle\langle e_j| = \rho_A.
\end{aligned}$$

□

In Exercise 2.9 you will give an alternative proof of Lemma 2.10, leading to the so-called *standard purification* of $\rho_A \in \mathrm{S}(A)$. This purification is given by the state

$$\sum_a (\sqrt{\rho_A} \otimes \mathbb{1}_R)|aa\rangle \in \mathcal{H}_A \otimes \mathcal{H}_R \tag{2.8}$$

where we have chosen a basis $|a\rangle$ for $\mathcal{H}_A$ and $\mathcal{H}_R$ is a copy of $\mathcal{H}_A$.

Remark 2.11. In the proof of Lemma 2.10 we used the spectral decomposition of ρ_A to find the purification. However, we may in fact start with *any* decomposition of the form

$$\rho_A = \sum_j p_j |\psi_j\rangle\langle\psi_j|.$$

Then

$$\sum_j \sqrt{p_j}|\psi_j\rangle \otimes |j\rangle$$

is a purification. As a concrete example, the qubit state

$$\rho_A = \frac{1}{3}(|0\rangle\langle 0| + |+\rangle\langle +| + |1\rangle\langle 1|)$$

has the following purification

$$|\phi_{AR}\rangle = \frac{1}{\sqrt{3}}(|00\rangle + |{+}1\rangle + |12\rangle),$$

where $\mathcal{H}_R = \mathbb{C}^3$.

If we have a purification $|\phi_{AR}\rangle$ of ρ_A it is easy to see that for any unitary $U_R \in \mathrm{U}(R)$ the pure state $(\mathbb{1}_A \otimes U_R)|\phi_{AR}\rangle$ is also a purification. More generally, if $V_{R\to S} \in \mathrm{Isom}(R,S)$ is an isometry from R to some other quantum system S, it is also true that $|\psi_{AS}\rangle = (\mathbb{1}_A \otimes V_{R\to S})|\phi_{AR}\rangle$ is again a purification of ρ_A, now with reference system S. This is in fact all the freedom we have in choosing purifications: up to isometries purifications are unique!

Lemma 2.12. *Suppose $|\phi_{AR}\rangle$ and $|\psi_{AS}\rangle$ are purifications of ρ_A. Without loss of generality, suppose that $|R| \leq |S|$. Then there exists an isometry $V_{R\to S} \in \mathrm{Isom}(R,S)$ such that*

$$(\mathbb{1}_A \otimes V_{R\to S})|\phi_{AR}\rangle = |\psi_{AS}\rangle.$$

In particular, when $S = R$ then the purification is unique up a unitary $U_R \in \mathrm{U}(R)$.

The proof of Lemma 2.12 relies on a very useful tool called the Schmidt decomposition, which we discuss in the next section. We will then give a proof sketch of Lemma 2.12 at the end of that section. You can fill in all details in Exercise 2.10.

2.3 The Schmidt decomposition

A basic fact from linear algebra, reviewed in Theorem A.4 is that every matrix has a *singular value decomposition.* [MW: Maybe state for matrices, since this is all we use?] That is, given $M \in \mathrm{Lin}(\mathcal{H},\mathcal{K})$, there exist bases $\{|e_j\rangle\}$ and $\{|g_j\rangle\}$ of $\mathcal{K}$ and $\mathcal{H}$, respectively, as well as positive numbers $s_1 \geq \ldots \geq s_r > 0$ (the *singular values* of M) such that

$$M = \sum_{j=1}^{r} s_j |e_j\rangle\langle g_j|.$$

The number of singular values equals $r = \mathrm{rank}(M)$

If we have a bipartite pure quantum state $|\psi_{AB}\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ we may reinterpret the state $|\psi_{AB}\rangle$ as a linear map M in $\mathrm{Lin}(\mathcal{H}_A^*, \mathcal{H}_B)$ and apply the singular value decomposition to M. This leads directly to the *Schmidt decomposition* of pure bipartite quantum states.

Theorem 2.13 (Schmidt decomposition). *Suppose $|\psi_{AB}\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ is a pure quantum state. Then there are bases $\{|e_j\rangle\}_{j=1}^{|A|}$ and $\{|f_j\rangle\}_{j=1}^{|B|}$ of $\mathcal{H}_A$ and $\mathcal{H}_B$, and positive numbers $s_1 \geq s_2 \ldots \geq s_r > 0$, where $r \leq \min(|A|,|B|)$, such that $\sum_j s_j^2 = 1$ and*

$$|\psi_{AB}\rangle = \sum_{j=1}^{r} s_j \, |e_j\rangle \otimes |f_j\rangle.$$

The numbers $s_1, \ldots, s_r$ are called the Schmidt coefficients *and r is called the* Schmidt rank.

Proof. The idea is that we interpret $|\psi_{AB}\rangle$ as a $|A| \times |B|$-matrix and apply Theorem A.4. To make this completely explicit we choose arbitrary bases $|a\rangle$ and $|b\rangle$ for $\mathcal{H}_A$ and $\mathcal{H}_B$ and let M be the $|A| \times |B|$ matrix defined by

$$M_{ab} = \langle ab|\psi_{AB}\rangle.$$

We now apply Theorem A.4 to find that

$$M = \sum_{j=1}^{r} s_j |e_j\rangle\langle g_j|$$

so

$$\begin{aligned} M_{ab} = \langle a|M|b\rangle &= \sum_{j=1}^r s_j \langle a|e_j\rangle\langle g_j|b\rangle \\ &= \sum_{j=1}^r s_j \langle a|e_j\rangle\overline{\langle b|g_j\rangle} \\ &= \sum_{j=1}^r s_j \langle a|e_j\rangle\langle b|f_j\rangle, \end{aligned}$$

where we denote by $|f_j\rangle$ the vectors whose entries with respect to the basis $|b\rangle$ are the complex conjugate of the entries of $|g_j\rangle$ (i.e., $\langle b|f_j\rangle = \overline{\langle b|g_j\rangle} = \langle g_j|b\rangle$ for all b). Note that $\{|f_j\rangle\}$ is also an orthonormal basis. We now use that these are the coefficients of our quantum state:

$$\begin{aligned} |\psi_{AB}\rangle &= \sum_{a,b} M_{ab}\,|a\rangle \otimes |b\rangle \\ &= \sum_{j=1}^r \sum_{a,b} s_j\,|a\rangle\langle a|e_j\rangle \otimes |b\rangle\langle b|f_j\rangle \\ &= \sum_{j=1}^r \sum_{a,b} s_j\,|e_j\rangle \otimes |f_j\rangle. \end{aligned}$$

The fact that the state is normalized implies that $\sum_j s_j^2 = 1$. □

To see why this is useful, let us compute the reduced density matrix of $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$ on the A and B systems. To compute the reduced density matrix ρ_A we compute the partial trace over B using Eq. (2.4) (or using the basis $\{|f_i\rangle\}$):

$$\begin{aligned} \rho_A &= \mathrm{tr}_B[|\psi_{AB}\rangle\langle\psi_{AB}|] \\ &= \mathrm{tr}_B\left[\sum_{j,k=1}^r s_j s_k |e_j\rangle\langle e_k| \otimes |f_j\rangle\langle f_k|\right] \\ &= \sum_{j=1}^r s_j^2\,|e_j\rangle\langle e_j|. \end{aligned}$$

We see that the reduced density matrix is diagonal in the basis $\{|e_i\rangle\}$ and that its nonzero eigenvalues equal the squared singular values. This is very similar to the computation in Lemma 2.10; the state $|\phi_{AR}\rangle$ constructed there was already written as a Schmidt decomposition! If we take the partial trace over A we have (by an analogous calculation)

$$\rho_B = \sum_{j=1}^r s_j^2\,|f_j\rangle\langle f_j|.$$

We collect this important fact as a lemma.

Lemma 2.14. *If $|\psi_{AB}\rangle$ is a pure state with Schmidt decomposition*

$$|\psi_{AB}\rangle = \sum_{i=1}^{r} s_i \, |e_i\rangle \otimes |f_i\rangle,$$

then the reduced density matrices of $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$ are given by

$$\rho_A = \sum_{i=1}^{r} s_i^2 |e_i\rangle\langle e_i| \quad \text{and} \quad \rho_B = \sum_{i=1}^{r} s_i^2 |f_i\rangle\langle f_i|.$$

In particular, the Schmidt rank and the Schmidt coefficients are uniquely determined are the rank and the nonzero eigenvalues of the reduced states, respectively.

An important conclusion is that if we have a pure bipartite state, the reduced density matrices on both subsystems have the same nonzero eigenvalues!

Lemma 2.14 can be used to prove Lemma 2.12. Suppose $|\phi_{AR}\rangle$ and $|\psi_{AS}\rangle$ are two purifications of ρ_A. Their Schmidt decompositions must have the form

$$|\psi_{AR}\rangle = \sum_{j=1}^{r} s_j \, |e_j\rangle \otimes |f_j\rangle \quad \text{and} \quad |\psi_{AS}\rangle = \sum_{j=1}^{r} s_j \, |e_j'\rangle \otimes |f_j'\rangle$$

as the Schmidt rank and coefficients are uniquely determined by ρ_A. Suppose for simplicity that the Schmidt coefficients are all distinct. Then the eigenspaces of ρ_A are one-dimensional and it follows that the eigenvectors $|e_j\rangle$ and $|e_j'\rangle$ are the same up to a phase. By absorbing the phase into the definition of $|f_j'\rangle$, we can in fact assume that $|e_j\rangle = |e_j'\rangle$. It follows that $(I_A \otimes V_{R\to S})|\psi_{AR}\rangle = |\psi_{AS}\rangle$ for any isometry $V_{R\to S}$ that extends $|f_j\rangle \mapsto |f_j'\rangle$. The general case where the Schmidt coefficients need not all be distinct is discussed in Exercise 2.10.

Another useful consequence of the Schmidt decomposition is that a pure bipartite state $|\psi_{AB}\rangle$ with density matrix ρ_{AB} is a product state if and only if the reduced density matrix ρ_A is pure (or equivalently, ρ_B is pure). We will prove this observation in Lemma 2.16 in the following section, where we will learn more about states which are *not* product states!

2.4 Entanglement

Suppose we have a probability distribution with density matrix ρ_{XY} on classical systems X and Y. The two random variables are independent if they form a product state $\rho_{XY} = \rho_X \otimes \rho_Y$. The terminology is fitting, as it means that the outcome on X does not influence the outcome on Y and vice versa. If X and Y are not independent they are *correlated*. For instance, we already saw the maximally correlated state on two qubits

$$\rho_{XY} = \frac{1}{2}\left(|00\rangle\langle 00| + |11\rangle\langle 11|\right).$$

This state is such that the reduced density matrices on X and Y are maximally mixed. As soon as we learn the outcome of X however, we know that Y must have the same outcome. We will now discuss a diffrent kind of 'non-classical' correlations, so-called *entanglement*. The existence of entanglement creates fundamental differences between classical and quantum information theory.

A classical state on systems X and Y is of the form

$$\rho_{XY} = \sum_{x,y} p(x,y) |x\rangle\langle x| \otimes |y\rangle\langle y|$$

In particular, we see that it is a convex combination of the product states $|x\rangle\langle x| \otimes |y\rangle\langle y|$. The next definition captures a wider class of the states where the correlations between A and B are of a classical nature.

Definition 2.15. A state $\rho_{AB} \in \mathrm{S}(AB)$ is *separable* if there exists a collection of density matrices $\rho_{A,x} \in \mathrm{S}(A)$, $\rho_{B,x} \in \mathrm{S}(B)$ for $x \in \Omega$ and a probability distribution $p \in \mathrm{P}(\Omega)$ for some set Ω such that

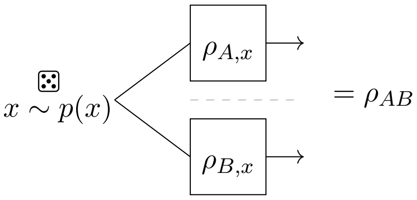
$$\rho_{AB} = \sum_{x \in \Omega} p(x) \rho_{A,x} \otimes \rho_{B,x}.$$

A state is called *entangled* if it is not separable.

Clearly, any classical state is separable. Entangled states are therefore a class of non-classical states. In particular, if a state ρ_{AB} is entangled, there is no choice of basis for A and B such that the state is classical in that basis. More generally, the class of separable states between Alice and Bob is the class of states they can prepare in the following way:

(a) Alice and Bob generate some shared classical random variable with an outcome $x \in \Omega$.

(b) Based on their random outcome x Alice prepares state $\rho_{A,x}$ and Bob prepares $\rho_{B,x}$.



From this interpretation we see that separable states form a natural class of states where the correlations between Alice and Bob are of a classical rather than quantum nature.

Let us investigate the notions of entanglement and separability for pure states. In this case, the condition simplifies significantly:

Lemma 2.16. *A pure state $|\psi_{AB}\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ is separable if and only if it is is a product state $|\psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$ for $|\psi_A\rangle \in \mathcal{H}_A$ and $|\psi_B\rangle \in \mathcal{H}_B$. In particular, a pure state ρ_{AB} is entangled if and only if the reduced density matrix ρ_A (or ρ_B) is not pure.*

Proof. If $|\psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$ then clearly

$$\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}| = |\psi_A\rangle\langle\psi_A| \otimes |\psi_B\rangle\langle\psi_B|$$

is a product state and hence separable. For the converse, by Lemma 1.10 pure states are extremal in the set of states, and therefore if a pure state $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$ is a convex combination of product states, it must be a product state itself. If

$$|\psi_{AB}\rangle\langle\psi_{AB}| = \rho_{AB} = \rho_A \otimes \rho_B$$

then we must have $1 = \mathrm{rank}(\rho_{AB}) = \mathrm{rank}(\rho_A)\,\mathrm{rank}(\rho_B)$ (see Lemma A.6) so ρ_A and ρ_B must have rank one and hence $\rho_A = |\psi_A\rangle\langle\psi_A|$ and $\rho_B = |\psi_B\rangle\langle\psi_B|$ so $|\psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$. □

As an example of an entangled state we saw the *maximally entangled state* of a pair of qubits in Examples 2.3 and 2.6. We now give a general definition of maximally entangled states.

Definition 2.17. We say that a state $\rho_{AB} \in \mathrm{S}(AB)$ is *maximally entangled state* if ρ_{AB} is pure and both reduced states are maximally mixed, that is, $\rho_A = \frac{\mathbb{1}}{|A|}$ and $\rho_B = \frac{\mathbb{1}}{|B|}$.

Lemma 2.16 implies that the maximally entangled states are indeed entangled, since their reduced states are maximally mixed – in stark contrast to the unentangled pure states, which only have a single nonzero Schmidt coefficient. This gives a first explanation for the terminology "maximally entangled".

By the Schmidt decomposition (Lemma 2.14), a pure state is maximally entangled if and only if its Schmidt rank is d and its Schmidt coefficients are all equal to $1/\sqrt{d}$, where $d = |A| = |B|$. In particular, maximally entangled states exist if and only if $|A| = |B|$, and they take the form

$$|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^{d} |e_j\rangle \otimes |f_j\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B \tag{2.9}$$

where $\{|e_j\rangle\}_{j=1}^d$ and $\{|f_j\rangle\}_{j=1}^d$ are bases of $\mathcal{H}_A$ and $\mathcal{H}_B$, with $d = |A| = |B|$. For example, when $\mathcal{H}_A = \mathcal{H}_B = \mathbb{C}^d$ and we use the standard basis for both, we get

$$|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{d}} \sum_{x=0}^{d-1} |xx\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d \tag{2.10}$$

Some useful properties are the following, which you will prove in Exercise 2.12

Lemma 2.18. *Let $|\Phi^+_{AB}\rangle$ be the maximally entangled state in Eq. (2.9) and let $\rho_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|$.*

(a) The reduced density matrices are maximally mixed: $\rho_A = \frac{1}{d}\mathbb{1}_A$.

(b) For any $d \times d$ matrix M, we have $(M_A \otimes \mathbb{1}_B)|\Phi^+_{AB}\rangle = (\mathbb{1}_A \otimes M_B^{\mathsf{T}})|\Phi^+_{AB}\rangle$.

(c) For any two $d\times d$-matrices M, N we have $\langle\Phi^+_{AB}|M_A \otimes N_B|\Phi^+_{AB}\rangle = \frac{1}{d}\operatorname{tr}\left[M^{\mathsf{T}} N\right] = \frac{1}{d}\operatorname{tr}\left[M N^{\mathsf{T}}\right]$.

When $\mathcal{H}_A = \mathcal{H}_B = \mathbb{C}^d$ and we use the standard basis then the notation is clear. In general, the operators M_A, M_B^{T}, etc. are defined with respect to the same bases as in Eq. (2.9). For example, $M_A = \sum_{a,a'} M_{a,a'}|e_a\rangle\langle e_{a'}|$, while $N_B = \sum_{b,b'} N_{b,b'}|f_b\rangle\langle f_{b'}|$.

The statement of (b) is known as the *transpose trick*. Again, both the maximally entangled state in Eq. (2.9) and the transpose in Lemma 2.18 depend on the choice of bases. However, there is some redundancy. For instance, if U is any unitary with real entries, so $U = \overline{U}$ and $U^{\mathsf{T}} = U^\dagger$, by the transpose trick we have $(U \otimes U)|\Phi^+_{AB}\rangle = |\Phi^+_{AB}\rangle$. For example, for a pair of qubits we have

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

by applying the unitary

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

Outlook

Entanglement for mixed states

The Schmidt decomposition gives essentially a complete understanding of entanglement of bipartite pure states. In particular, given a pure bipartite quantum state it is easy to decide whether the state is entangled: one only needs to compute the Schmidt decomposition and verify whether the state is a product state.

The situation is markedly different for bipartite mixed states. The definition of entanglement is somewhat implicit: to determine whether a state is entangled we have to prove that there is *no* decomposition into a convex combination of product states. It turns out that determining whether a given bipartite quantum state is entangled is *NP-hard* [18]. This means that any known algorithm which always succeeds in determining separability is exponential in the dimension of the Hilbert spaces of Alice and Bob. This is a classical problem: the input is a description of the state ρ_{AB} as a matrix. Furthermore, the size of this matrix is itself exponential in the number of qubits Alice and Bob share!

This means that in practice any approach to quantify mixed state entanglement is either computationally inefficient, or it must 'fail' on certain instances. That is, there are measures which are easy to compute but which do not give a conclusive answer. There are a number of useful ways to 'witness' that a state is entangled. The most well-known is the negativity, or PPT criterion [35]. Here, one applies a *transpose* operation to the density matrix ρ_{AB}, but only to the Alice's system to get the partial transpose $\Gamma_A(\rho_{AB})$. If the resulting matrix is no longer positive, ρ_{AB} must have been entangled. If $\Gamma_A(\rho_{AB}) \geq 0$, the result is inconclusive: ρ_{AB} could be either entangled or separable. You can explore this in Exercise 2.16.

A different approach to quantifying entanglement is the *extendibility criterion* [14]. Here the idea is that if a state ρ_{AB} is separable, there must exist a state $\sigma_{AB_1B_2}$ where B_1 and B_2 are copies of the B system such that $\sigma_{AB_1} = \rho_{AB}$ and $\sigma_{AB_2} = \rho_{AB}$. If you can prove that no such extension of ρ_{AB} exists, the state must be entangled. You can also ask whether there are extensions to more than two systems; i.e., whether there exists $\sigma_{AB_1\ldots B_n}$ such that $\sigma_{AB_i} = \rho_{AB}$ for all $i = 1, \ldots, n$. This leads to a complete criterion for entanglement: a state is separable if and only if it has extensions $\sigma_{AB_1\ldots B_n}$ for all n.

If you want to learn more about mixed state entanglement, a good review is [24].

Multipartite entanglement

Another extension is to study entanglement for states on more than two parties. For example, one can study pure states $|\psi_{ABC}\rangle$ for three parties Alice, Bob and Charlie. This is complicated and not very well understood; in contrast to the bipartite case it is not even quite clear what the right definitions to study are! For example: what is the right notion of a 'product state'? One can either take states $|\phi_A\rangle|\phi_B\rangle|\phi_C\rangle$ which are a product over the three parties, or states like $|\phi_{AB}\rangle|\phi_C\rangle$ which are a bipartite product state between some partition of the parties.

An approach to study multipartite entanglement that is not very satisfying from an information theoretic point of view, but has the advantage of being amenable to mathematical analysis, is to study transformations between states under arbitrary product transformations. That is, we consider transformations of the form

$$|\psi_{ABC}\rangle = (M_A \otimes M_B \otimes M_C)|\phi_{ABC}\rangle \tag{2.11}$$

where M_A, M_B and M_C are arbitrary linear maps on the A, B and C systems. Such transformations can be realized using *SLOCC* protocols which use local quantum operations (LO), classical communication (CC), but also *postselection* (S for stochastic). Postselection on a measurement outcome means that one repeats a protocol, until one obtains the desired outcome; this typically makes such protocols infeasible in practice due to the number of required repetitions. We will discuss LOCC (so without postselection) in Lecture 5.

Given this notion of SLOCC one can now say that two states $|\phi_{ABC}\rangle$ have differing multipartite entanglement if there is *no* SLOCC transformation between them. Under this notion of multipartite entanglement, the situation for three qubits is completely understood [15]. In that case, one can have a bipartite entangled state between two of the parties. Beyond that, there are only two different multipartite entangled states: the GHZ-state and the W-state:

$$|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{3}}(|000\rangle + |111\rangle) \qquad \text{and} \qquad |W\rangle = \frac{1}{\sqrt{3}}(|100\rangle + |010\rangle + |001\rangle).$$

In general (so if we have more parties, or the parties have arbitrary dimensions) there is no complete classification, but there are mathematical tools to study SLOCC transformations [44].

An important application of quantum information theory is to study the structure of the type of quantum systems one encounters in real physical systems. These will consist of many quantum particles (say electrons in a strongly correlated system). A mathematical tool for representing such quantum states are *tensor networks*. They impose a certain entanglement structure to the particle, where entanglement is spatially local, and they are closely related to the SLOCC transformations defined in Eq. (2.11). Reviews of the theory of tensor networks can be found in [9, 49], see [7] for an explanation of the relation to multipartite entanglement.

The quantum marginal problem

In this lecture we saw that given a state ρ_{ABC}, we can compute reduced density matrices such as ρ_{AB} and ρ_{BC}. One could also ask a reverse question: given states ρ_{AB} and ρ_{BC}, does there exist a global state σ_{ABC} with the reduced density matrices $\sigma_{AB} = \rho_{AB}$ and $\sigma_{BC} = \rho_{BC}$. This is known as the *quantum marginal problem.*

As a concrete example: does there exist a state σ_{ABC} on three qubits such that the reduced states σ_{AB} and σ_{BC} are both maximally entangled? The answer is no! You can show this in Exercise 2.17.

One can also investigate the same question with larger collections of parties. Solving the marginal problem would be very powerful, as it would allow one to reduce computational problems involving many particles (such as appear in quantum chemistry or condensed matter physics) to local problems as was already observed by Coulson in [11]. He argued that if one could characterize the possible collections of marginals, one could use this to efficiently compute energies of large interacting quantum mechanical systems. However, the quantum marginal problem is computationally hard [27, 28]. Indeed, it is QMA-complete, meaning that we do not expect it to be solvable efficiently on a quantum computer.

2.5 Exercises

2.1 **Bipartite states, bra-ket notation and partial traces:** If we have a tensor product $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ and we would like to write a vector $|v\rangle \in \mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ as an element of $\mathbb{C}^{d_1 d_2}$ we order the product standard basis $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ *lexicographically*, as we also did in Example 2.1. [1] For

[1] This order is similar to the way you order a dictionary. Formally: $(i_1, i_2) < (j_1, j_2)$ for $i_1, j_1 \in \{0, \ldots, d_1 - 1\}$ and $i_2, j_2 \in \{0, \ldots, d_2 - 1\}$ if $i_1 < j_1$ or $i_1 = j_2$ and $i_2 < j_2$.

instance, under this ordering we may identify $|01\rangle \in \mathbb{C}^2 \otimes \mathbb{C}^2$ with

$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{C}^4,$$

because $|01\rangle$ is the second basis vector in the lexicographically ordered list $|00\rangle, |01\rangle, |10\rangle, |11\rangle$.

(a) Let $|\psi_{AB}\rangle \in \mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^2 \otimes \mathbb{C}^3$ be the vector given by

$$|\psi_{AB}\rangle = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

in the lexicographically ordered product basis. Write $|\psi_{AB}\rangle$ in bra-ket notation in the product basis $|ab\rangle$ for $\mathcal{H}_{AB}$.

(b) Let $|\phi_{A'B'}\rangle \in \mathcal{H}_{A'B'} = \mathcal{H}_{A'} \otimes \mathcal{H}_{B'} = \mathbb{C}^3 \otimes \mathbb{C}^2$ be the vector given by

$$|\phi_{A'B'}\rangle = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

in the lexicographically ordered product basis. Write $|\phi_{A'B'}\rangle$ in bra-ket notation in the computational basis for $\mathcal{H}_{A'B'}$.

(c) Continuing the last exercise, let $V_{A'B' \to AB}$ denote the isometry $\mathcal{H}_{A'B'} \to \mathcal{H}_{AB}$ that swaps the two subsystems (mapping $|xy\rangle \mapsto |yx\rangle$ for $x \in \{0,1,2\}$ and $y \in \{0,1\}$). Write down $V_{A'B' \to AB}|\phi_{A'B'}\rangle \in \mathcal{H}_{AB}$ as a 6-dimensional vector as well as in bra-ket notation.

(d) Let $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$ and $\sigma_{A'B'} = |\phi_{A'B'}\rangle\langle\phi_{A'B'}|$. Compute the reduced density matrices ρ_A, ρ_B, $\sigma_{A'}$ and $\sigma_{B'}$.

2.2 **Maximally entangled states:** Let A and B be qubit systems.

(a) Let $|\Psi^-_{AB}\rangle = \frac{1}{\sqrt{2}}(|{+}{-}\rangle - |{-}{+}\rangle)$. Show that $|\Psi^-_{AB}\rangle = \frac{1}{\sqrt{2}}(|10\rangle - |01\rangle)$.

(b) Let $|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ be a maximally entangled state. Show that for any $U_A \in \mathrm{U}(A)$ and $U_B \in \mathrm{U}(B)$, the reduced states of $|\phi_{AB}\rangle = (U_A \otimes U_B)|\Phi^+_{AB}\rangle$ on both A and on B are maximally mixed.

2.3 **Marginal distributions:** Let X and Y be bits with joint probability distribution p_{XY} given by

$$p_{XY}(0,0) = p_{XY}(1,1) = \frac{1}{4}, \quad p_{XY}(0,1) = \frac{1}{8} \quad p_{XY}(1,0) = \frac{3}{8}.$$

(a) Compute the probability distributions p_X and p_Y.
(b) Compute the marginal distribution $p_{X|Y=0}$.
(c) Are X and Y independent? Prove your claim.

2.4 **Independence and product states:** Show that if p_{XY} is a probability distribution then X and Y are independent (recall this means $p_{XY}(x,y) = p_X(x)p_Y(y)$ for all x, y) if and only if the corresponding density matrix ρ_{XY} is a product state, i.e. $\rho_{XY} = \rho_X \otimes \rho_Y$.

2.5 **Tensor products of operators:** This exercise studies tensor products of linear operators: if $M \in \mathrm{Lin}(\mathcal{H})$, $N \in \mathrm{Lin}(\mathcal{K})$ then $M \otimes N \in \mathrm{Lin}(\mathcal{H} \otimes \mathcal{K})$.

(a) Prove Lemma A.6.
(b) Show that tensor products also preserve other properties: if $M \in \mathrm{Lin}(\mathcal{H})$ and $N \in \mathrm{Lin}(\mathcal{K})$ have one of the properties {Hermitian, projection, unitary} then the tensor product $M \otimes N$ has that property as well.

2.6 **Not product states:** Show that the maximally entangled state and the maximally correlated state on two qubits are both not product states.

2.7 **Product measurement:** For measurements $\mu_A \in \mathrm{Meas}(A, \Omega_1)$ and $\mu_B \in \mathrm{Meas}(B, \Omega_2)$ on quantum systems A and B, the *product measurement* $\mu_A \otimes \mu_B \in \mathrm{Meas}(AB, \Omega_1 \times \Omega_2)$ is defined by, for $x_1 \in \Omega_1$ and $x_2 \in \Omega_2$,

$$(\mu_A \otimes \mu_B)(x_1, x_2) := \mu_A(x_1) \otimes \mu_B(x_2).$$

It describes the situation that we perform both measurements, one on each subsystem.

(a) Show that this formula indeed defines a measurement.
(b) Show that if we measure any state $\rho_{AB} \in \mathrm{S}(AB)$ using the product measurement, then the marginal probability of Alice's outcome $x_1 \in \Omega_1$ is the same as if Bob did not make any measurement at all. That is, show that it is given by $\mathrm{tr}[\mu_A(x_1)\rho_A]$ for

Now suppose that Alice and Bob share a maximally entangled state $|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$.

(c) Suppose that Alice and Bob measure in the standard basis. What is the probability distribution of the outcomes?
(d) suppose that Alice and Bob instead measure in the X-basis $|+\rangle$, $|-\rangle$. What is the probability distribution of the outcomes?
(e) In both cases, what does the marginal distribution of the outcomes look like for Alice and Bob? Relate this to their reduced states. *Note that while the measurement outcomes are correlated, performing local measurements on a maximally entangled state does not allow Alice and Bob to communicate information.*

2.8 **Uniqueness of partial trace:** Suppose that $\rho_{AB} \in \mathrm{S}(AB)$. Let $\sigma_A \in \mathrm{S}(A)$ be such that for an arbitrary measurement $\mu_A = \{\mu_A(x)\}_{x\in\Omega}$ on A, we have

$$\mathrm{tr}[\mu_A(x)\sigma_A] = \mathrm{tr}[(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}].$$

Argue that $\sigma_A = \rho_A = \mathrm{tr}_B[\rho_{AB}]$. *Hint: argue that this condition implies* $\mathrm{tr}[M_A\rho_A] = \mathrm{tr}[M_A\sigma_A]$ *for all* $M_A \in \mathrm{Lin}(A)$.

2.9 **The standard purification:** In this exercise you will prove Lemma 2.10 again. Suppose that $\rho_A \in \mathrm{S}(A)$. Let $|a\rangle$ be a basis for $\mathcal{H}_A$ and let $\mathcal{H}_R = \mathcal{H}_A$. Show that

$$\sum_a (\sqrt{\rho_A} \otimes \mathbb{1}_R)|aa\rangle \in \mathcal{H}_A \otimes \mathcal{H}_R$$

is a purification of ρ_A. This purification is called the *standard purification*.

2.10 **Uniqueness of purifications:** The goal of this exercise is to prove Lemma 2.12. Suppose that $\rho_A \in \mathrm{S}(A)$, and suppose that $|\psi_{AR}\rangle$ and $|\phi_{AS}\rangle$ are purifications on systems R, S with $\dim(\mathcal{H}_R) \leq \dim(\mathcal{H}_S)$. Let

$$|\psi_{AR}\rangle = \sum_{i=1}^{r} s_i |e_i\rangle|f_i\rangle \quad \text{and} \quad |\phi_{AS}\rangle = \sum_{i=1}^{q} t_i |g_i\rangle|h_i\rangle$$

be Schmidt decompositions.

(a) Show that $r = q$ and $s_i = t_i$ for $i = 1, \ldots, r$ (recall that $s_1 \geq s_2 \geq \cdots \geq s_r$ and $t_1 \geq t_2 \geq \cdots \geq t_q$ in the Schmidt decomposition).

(b) Suppose that the Schmidt spectrum is *nondegenerate*, meaning that $s_1 \neq s_2 \neq \ldots \neq s_r$. Show that we must have $|e_i\rangle\langle e_i| = |g_i\rangle\langle g_i|$ for $i = 1, \ldots, r$. Use this to prove Lemma 2.12 in the special case where the spectrum is nondegenerate and moreover $\dim(\mathcal{H}_R) = r$.

(c) In the general case, consider the map $\tilde{V} \in \mathrm{Lin}(R, S)$ given by

$$\tilde{V} = \sum_{i,j=1}^{r} \langle e_j|g_i\rangle |h_i\rangle\langle f_j|.$$

Show that $(\mathbb{1}_A \otimes \tilde{V})|\psi_{AR}\rangle = |\phi_{AS}\rangle$.

(d) Prove Lemma 2.12. *Hint: note that the map $\tilde{V}$ need not be an isometry if $r < \dim(\mathcal{H}_R)$.*

2.11 **Properties of the partial trace:** Prove Lemma 2.7.

2.12 **Properties of maximally entangled states:** Prove Lemma 2.18.

2.13 **Properties of bipartite systems:** Decide whether each of the following statements is true or false, and provide either a proof or a counterexample.

(a) $\operatorname{tr}_A X_{AB} = \operatorname{tr}_B X_{AB} = 0 \Rightarrow X_{AB} = 0$.

(b) $\rho_A \otimes \sigma_B \in \mathrm{S}(AB)$ if and only if $\rho_A \in \mathrm{S}(A)$ and $\sigma_B \in \mathrm{S}(B)$.

(c) $X_A \otimes Y_B = Y_A \otimes X_B$ if and only if $X \propto Y$.

2.14 **Separability:** Show that a state $\rho_{AB} \in \mathrm{S}(AB)$ is separable if and only if there exists a collection $P_{A,x} \in \mathrm{PSD}(A)$ and $P_{B,x} \in \mathrm{PSD}(B)$ such that

$$\rho_{AB} = \sum_{x \in \Omega} P_{A,x} \otimes P_{B,x}$$

(so as opposed to Definition 2.15, the operators need not have normalized trace).

2.15 **Purity of quantum states:** The *purity* of a quantum state ρ is defined as $P(\rho) = \operatorname{tr} \rho^2$.

(a) For $\rho \in S(A)$, $\dim \mathcal{H}_A = d$, show that

$$\frac{1}{d} \leq P(\rho) \leq 1 \ .$$

When is equality achieved for each of these bounds?

(b) Let $\rho_{AB} \in S(A \otimes B)$ be a pure state with marginal states ρ_A and ρ_B. Show that $P(\rho_A) = P(\rho_B)$.

2.16 **The PPT criterion:** The *partial transpose map* $\Gamma_A : \mathrm{Lin}(\mathbb{C}^2 \otimes \mathbb{C}^2)$ is defined as the linear extension of

$$\Gamma_A(X \otimes Y) = X^{\mathsf{T}} \otimes Y \ .$$

In other words, we take the transpose of the operator on the first system and leave the second unchanged.

(a) Suppose $\rho_{AB} \in \mathrm{S}(AB)$ is expanded in a basis as

$$\rho_{AB} = \sum_{a,a'} \sum_{b,b'} \rho_{aa',bb'} |a\rangle\langle a'| \otimes |b\rangle\langle b'|.$$

Write down an expansion of $\Gamma_A(\rho_{AB})$ in the same basis.

(b) Let A and B be qubit systems, and let ρ_{AB} be the maximally entangled state. Compute $\Gamma_A(\rho_{AB})$ and write it down as a 4×4 matrix in the standard basis.

A state $\rho \in S(AB)$ satisfies the *positive partial transpose* (PPT) criterion if $\Gamma_A(\rho) \geq 0$.

(c) Show that the maximally entangled state is *not* PPT.

(d) Show that if $P_A \in \mathrm{PSD}(A)$, then $P_A^{\mathsf{T}} \in \mathrm{PSD}(A)$.

(e) Prove that if ρ is a separable state, then ρ is PPT. Conclude that a state which is not PPT must be entangled.

(f) Consider the two-qubit states

$$|\Psi^-_{AB}\rangle = \frac{1}{\sqrt{2}}\big(|01\rangle - |10\rangle\big)$$
$$|\phi_{AB}\rangle = \frac{1}{2}\big(\omega|00\rangle + |01\rangle - i|10\rangle + \omega|11\rangle\big) \ , \quad \text{where } \omega = e^{i\pi/4}.$$

Are these states PPT? Are they entangled or separable? *Hint: as a shortcut for the second state, write*

$$|\phi_{AB}\rangle = \frac{1}{2}(|0\rangle(\omega|0\rangle + |1\rangle) + |1\rangle(-i|0\rangle + \omega|1\rangle)).$$

How can you now use your result for $|\Psi^-_{AB}\rangle$*?*

(g) For $\nu \in [0,1]$, define the following family of two qubit states $\rho_\nu \in S(\mathbb{C}^2 \otimes \mathbb{C}^2)$:

$$\rho_\nu = \nu|\Psi^-_{AB}\rangle\langle\Psi^-_{AB}| + \frac{1-\nu}{4}\mathbb{1}_A \otimes \mathbb{1}_B \ .$$

These are called *Werner states.* Compute the values of ν for which ρ_ν is PPT.

Comment: On two qubits, the PPT criterion is a necessary and sufficient condition for a state to be entangled! For general states on $\mathbb{C}^d \otimes \mathbb{C}^d$, for $d \geq 3$, the condition is sufficient but not necessary.

2.17 **Marginal problem for maximally entangled states:**

(a) Suppose that $\rho_{AB} \in \mathrm{S}(AB)$. Show that if ρ_A is pure, then $\rho_{AB} = \rho_A \otimes \rho_B$. *Hint: consider a purification of* ρ_{AB}*.*

(b) Suppose that $\rho_{ABC} \in \mathrm{S}(ABC)$ and suppose that ρ_{AB} is a pure state. Show that $\rho_{BC} = \rho_B \otimes \rho_C$.

(c) Now let $\rho_{ABC} \in \mathrm{S}(ABC)$ be such that both ρ_{AB} and ρ_{AC} are both pure states. Show that $\rho_{ABC} = \rho_A \otimes \rho_B \otimes \rho_C$.

(d) Conclude that there can be no state ρ_{ABC} on three qubits such that ρ_{AB} is maximally entangled and ρ_{BC} is maximally entangled.

Lecture 3

Correlations, entanglement and games

Concept	Math translation
Correlations between two systems	Bell game: players get questions x, y and answer a, b according to a probability $p(a, b\|x, y)$ where Alice only knows x and Bob only knows y.
Classical correlations	Strategies for the Bell game that only use shared randomness.
Quantum correlations	Quantum strategies for a Bell game, where Alice and Bob measure their part of an entangled state. The measurement settings depend on the question. $$p(a,b\|x,y) = \mathrm{tr}\Big[(\mu_A^{(x)}(a) \otimes \mu_B^{(y)}(b))\rho_{AB}\Big].$$
Separation between quantum and classical correlations	The classical and quantum value of a game $\omega(G)$ and $\omega^*(G)$. Lemma 3.4, Theorem 3.5 and Theorem 3.6 show that for the CHSH game $$\omega(G) = \frac{3}{4} < \cos^2(\frac{\pi}{8}) = \omega^*(G).$$

Suppose we have a probability distribution with density matrix ρ_{XY} on classical systems X and Y. In the last lecture we introduced the notion of independence, which is equivalent to the state being a product state $\rho_{XY} = \rho_X \otimes \rho_Y$. The terminology is fitting, as it means that the outcome on X does not influence the outcome on Y and vice versa. If X and Y are not independent they are *correlated*. For instance, we already saw the maximally correlated state on two qubits

$$\rho_{XY} = \frac{1}{2}\left(|00\rangle\langle 00| + |11\rangle\langle 11|\right).$$

This state is such that the reduced density matrices on X and Y are maximally mixed. As soon as we learn the outcome of X however, we know that Y must have the same outcome. In this lecture we will investigate what kind of correlations can occur in quantum mechanics. We will see that there is a way to distinguish between 'non-classical' correlations that arise from *entanglement* rather than from classical correlations.

Classical versus quantum correlations

Is there a fundamental difference between classical and quantum correlations? In the end, as classical agents we only have access to statistics of quantum states through measurements. In principle, we have not yet excluded the possibility that there exists an alternative description of a quantum system which is purely classical and which exactly reproduces the measurement statistics. One could wonder whether quantum mechanics is a very effective model for making experimental predictions, but underneath it is some more elaborate model, with some *hidden variables* which is of a classical nature but gives rise to the same predictions as the quantum mechanical model.

To make this very concrete, consider two maximally entangled qubits between Alice and Bob $|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Suppose that Alice and Bob are spatially separated, and Alice measures her qubit in the standard basis. She will find outcome 0 or 1 with equal probability. Then, if *afterwards* Bob also measures his qubit in the standard basis, he will find the same outcome as Alice found. However, if Alice and Bob would share a *maximally correlated* state $\rho_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$, the very same phenomenon happens! If Alice measures in the standard basis, then with equal probability she will find 0 or 1, and if Bob measures afterwards he will find the same value. Here, we have the mental model that Alice and Bob either both receive the state 0 or both the state 1, but that they simply did not know yet which one it was. In other words, before the measurement there was a *hidden variable* which determined the state of the system. We are left with the question: can there be a classical hidden variable model for quantum mechanics? In other words, is there some way in which the whole framework of quantum theory we introduced in the previous two lectures can be reduced to classical probability theory? The answer is a resounding no! It turns out, as first shown by Bell, that there are certain correlations which can only be explained by quantum theories and not by classical models. To be more precise, this deals with *local hidden variable* models. The locality condition refers to the situation that there are systems which are such that when they are sufficiently spatially separated, they can not communicate. The assumption that distant systems satisfy such a locality constraint is reasonable, and is central to notions of causality especially in relativistic theories.

3.1 Bell games

A nice way to understand such correlations is by formulating them in terms of a certain type of 'game'. The set-up will be that we have a number of players, each of which receives a question from a referee. Each player then sends an answer to the referee. The players then win if they satisfy some winning condition. A crucial aspect is that the players *are not allowed to communicate*. Such games are called *Bell games*.[1] However, the players are allowed to share some state that has been prepared beforehand. This could either be a classical state (so the players have *shared randomness* as a resource) or a quantum state (so the players may use *entanglement* as a resource). We will see an example where this actually makes a difference.

Let us define more formally what we mean by a Bell game. We define a Bell game for two players, Alice and Bob. The generalization to more players is obvious, you will see an example in Exercise 3.3.

[1] They are also often called *nonlocal games*, we avoid this term because it is confusing terminology with respect to the relation with (non)local hidden variable theories.

Definition 3.1. A Bell game G for two players Alice and Bob consists of the following data:

(a) Two sets of *questions* $\mathcal{X}$ and $\mathcal{Y}$ the referee can pose to respectively Alice and Bob.

(b) Two sets of *answers* $\mathcal{A}$ and $\mathcal{B}$ respectively Alice and Bob can give to the referee.

(c) A probability distribution $p(x, y)$ according to which the referee asks questions $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

(d) A winning condition W which is a function $\mathcal{X} \times \mathcal{Y} \times \mathcal{A} \times \mathcal{B} \to \{0, 1\}$. If the players receive questions x, y and answer a, b, they win if $W(x, y, a, b) = 1$.

Alice and Bob play the game according to some *strategy*. They are allowed to coordinate beforehand but during the game they are not allowed to communicate. The goal of Alice and Bob is to maximize their probability of winning the game.

We consider three possible 'types' of strategies:

(a) *Deterministic strategies:* Here the answer is a deterministic function of the question, so there exist functions $f : \mathcal{X} \to \mathcal{A}$ and $g : \mathcal{Y} \to \mathcal{B}$ such that $a = f(x)$ and $b = g(y)$.

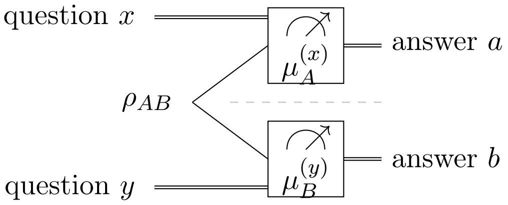
(b) *Randomized strategies:* Here the players may randomly pick their strategies, possibly based on some *shared* randomness. That is, Alice and Bob share the outcome of some random $\lambda \in \Lambda$ with probability $p_\Lambda(\lambda)$, and if they see outcome λ and questions x and y Alice answers a with probability $p_A(a|x, \lambda)$ and Bob answers b with probability $p_B(b|y, \lambda)$. Together, this means that Alice and Bob answer a and b when posed questions x and y with probability

$$p(a, b|x, y) = \sum_{\lambda \in \Lambda} p_\Lambda(\lambda) p_A(a|x, \lambda) p_B(b|y, \lambda)$$

(c) *Quantum strategies:* In this case, Alice and Bob are allowed to share a quantum state ρ_{AB}, and their answers are the result of a measurement on their system. That is, for each $x \in X$ there is a measurement $\mu_A^{(x)} = \{\mu_A^{(x)}(a) : a \in A\}$ on Alice's system and Alice answers the outcome of the measurement. Similarly, Bob has a measurement $\mu_B^{(y)} = \{\mu_B^{(y)}(b) : b \in B\}$ for each question y and answers the measurement outcome b. This means that Alice and Bob answer a and b when posed questions x and y with probability

$$p(a, b|x, y) = \operatorname{tr}\Big[(\mu_A^{(x)}(a) \otimes \mu_B^{(y)}(b))\rho_{AB}\Big].$$

The procedure may be visualised as



for a quantum strategy using state ρ. The dashed line indicates that Alice and Bob are not allowed to communicate. An overview of the notation we use:

Concept	Notation
Bell game	G
Agents	Referee, Alice and Bob
Questions	$x \in \mathcal{X}$ to Alice, $y \in \mathcal{Y}$ to Bob
Answers	$a \in \mathcal{A}$ from Alice, $b \in \mathcal{B}$ from Bob
Probability of answers	$p(a,b\|x,y)$ given questions x,y
Quantum strategy	$p(a,b\|x,y) = \operatorname{tr}[\mu_A^{(x)}(a) \otimes \mu_B^{(y)}(b)\rho_{AB}]$

A strategy, whether deterministic, random or quantum, gives a probability distribution $p(a,b|x,y)$ for each pair of questions x and y. Together with the probability distribution $p(x,y)$ according to which the questions are sampled, this gives the following expression for the *winning probability* of a strategy:

$$p_{\text{win}} = \sum_{x,y,a,b} p(x,y)p(a,b|x,y)W(x,y,a,b)$$

as the function W makes sure that only the probabilities where Alice and Bob win the game contribute to the sum.

The randomized strategies are the best you can do with a local hidden variable theory: the hidden variable is the $\lambda \in \Lambda$, and the assumption of locality is reflected by the fact that Alice and Bob are not allowed to communicate and their answer only depends on λ and the question they receive. One can show that shared randomness does not improve the winning probability of a game.

Lemma 3.2. *Suppose that there exists a randomized strategy with shared randomness Λ, winning a Bell game with probability p_{win}. Then there also exists a deterministic strategy winning the game with probability at least p_{win}.*

The proof is Exercise 3.5. The idea of the proof is that a randomized strategy is simply the (weighted) average over an ensemble of deterministic strategies. Therefore, the winning probability is the (weighted) average of the winning probabilities of the deterministic strategies, and therefore at least one the deterministic winning probabilities must be at least as large as the winning probability of the randomized strategy.

We now introduce the concept of the *(quantum or classical) value of a game*, which is the optimal winning probability for the game. In light of Lemma 3.2, when studying classical strategies we may restrict to deterministic strategies. Another observation is that we may assume without loss of generality that for a quantum strategy the quantum state is pure, by purifying the state. That is, given a strategy with state ρ_{AB} and measurement operators $\mu_A^{(x)}(a)$ and $\mu_{B,b}^{(y)}$, we let $|\psi_{ABR}\rangle$ be a purification of ρ_{AB}, we assign the reference system R to Alice, so her system is $\tilde{A} = AR$ and let her perform measurements $\tilde{\mu}_{\tilde{A},a}^{(x)} = \mu_A^{(x)}(a) \otimes \mathbb{1}_R$. Moreover, by a similar argument, we may assume that the measurements Alice and Bob apply are all *projective* measurements where the $\mu_A^{(x)}(a)$ and $\mu_B^{(y)}(b)$ are all projection operators. We will not give the argument for this fact at this point, but you may show later in Exercise 5.2 how this follows from a general principle. For this reason, for the rest of this lecture we will only consider quantum strategies suing pure states and projective measurements.

Definition 3.3. The classical value $\omega(G)$ of a game is the optimal winning probability using deterministic or randomized strategies. The quantum value $\omega^*(G)$ of a game is the optimal winning probability using quantum strategies.

CHSH game

To make the concept of a Bell game concrete we will study the important example of the *Clauser-Horne-Shimony-Holt (CHSH) game*. In this case, the game G has questions and answers which are both bits, so $\mathcal{X} = \mathcal{Y} = \mathcal{A} = \mathcal{B} = \{0,1\}$. The referee asks all questions with equal likelihood. The winning condition is specified by the following table:

x	y	winning condition
0	0	$a = b$
0	1	$a = b$
1	0	$a = b$
1	1	$a \neq b$

Another way to formulate the winning condition is that

$$x \,\mathrm{AND}\, y = a \,\mathrm{XOR}\, b$$

where XOR is the exclusive or (so $a \,\mathrm{XOR}\, b = 1$ if $a = 0, b = 1$ or $a = 1, b = 0$ and 0 otherwise) or

$$x \cdot y = a + b \mod 2.$$

Let us first investigate what a classical strategy can do. By Lemma 3.2 we may restrict to deterministic strategies. Alice and Bob can not win the game in all instances. One can prove this by trying out all deterministic strategies, or by the following argument: suppose that there exists a deterministic strategy with functions $f, g : \{0,1\} \to \{0,1\}$ such that $a = f(x)$ and $b = g(y)$ wins in all instances. Then the winning condition implies that

$$\sum_{x,y\in\{0,1\}} f(x) + g(y) \mod 2 \;=\; \sum_{x,y\in\{0,1\}} x \cdot y = 1$$

and on the other hand

$$\sum_{x,y\in\{0,1\}} (f(x) + g(y)) = 2 \sum_{x\in\{0,1\}} f(x) + 2 \sum_{y\in\{0,1\}} g(y)$$

is even which leads to a contradiction. So, in at least one of the four options for pairs of questions (x, y), Alice and Bob they will lose. So, $\omega(G) \leq 3/4$. On the other hand, if Alice and Bob both always answer $a = b = 0$, then they win in the first three cases so they win with probability $3/4$. We have proven:

Lemma 3.4. *The classical value of the CHSH game G is given by*

$$\omega(G) = \frac{3}{4}.$$

We proceed to study the quantum value of G.

Given a 2-outcome measurement μ with outcomes $0, 1$, it will be useful to define an operator

$$O = \mu(0) - \mu(1).$$

If you are a physicist, you may think of this as an *observable*, which takes values ±1 (associated to the measurement outcomes $0, 1$). However, for our purposes here these operators are just for convenient bookkeeping. Recall that we have the Bloch sphere picture of qubit measurements, as per Eq. (1.4). If we measure a qubit in the basis given by the antipodal point $\vec{r}, -\vec{r} \in \mathbb{R}^3$ for $\vec{r} = (x, y, z)$ on the Bloch sphere, then we get an observable

$$\begin{aligned} O(\vec{r}) = \mu(0) - \mu(1) &= \frac{1}{2}\begin{pmatrix} 1+z & x-iy \\ x+iy & 1-z \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1-z & -x+iy \\ -x-iy & 1+z \end{pmatrix} \\ &= \begin{pmatrix} z & x-iy \\ x+iy & -z \end{pmatrix}. \end{aligned} \tag{3.1}$$

Suppose we are given a quantum strategy, that is a quantum state $\rho_{AB} = |\psi\rangle\langle\psi|$ and for $x, y \in \{0, 1\}$ we are given measurements

$$\{\mu_A^{(x)}(a) : a \in \{0,1\}\} \quad \text{and} \quad \{\mu_B^{(y)}(b) : b \in \{0,1\}\}.$$

We then define the associated observables

$$O_A^{(x)} = \mu_A^{(x)}(0) - \mu_A^{(x)}(1) \quad \text{and} \quad O_B^{(y)} = \mu_B^{(y)}(0) - \mu_B^{(0)}(1).$$

The observables depend on the questions x, y. We then claim that

$$\alpha = \frac{1}{4}\langle\psi| O_A^{(0)} \otimes O_B^{(0)} + O_A^{(0)} \otimes O_B^{(1)} + O_A^{(1)} \otimes O_B^{(0)} - O_A^{(1)} \otimes O_B^{(1)} |\psi\rangle \tag{3.2}$$

equals the winning probability *minus* the losing probability of the quantum strategy. That is,

$$\alpha = p - (1-p) = 2p - 1.$$

To prove this claim, note that by definition of $O_A^{(x)}$ and $O_B^{(y)}$

$$\begin{aligned} \langle\psi| O_A^{(x)} \otimes O_B^{(y)} |\psi\rangle = {} & \langle\psi|\mu_A^{(x)}(0) \otimes \mu_B^{(y)}(0)|\psi\rangle + \langle\psi|\mu_A^{(x)}(1) \otimes \mu_B^{(y)}(1)|\psi\rangle \\ & - \langle\psi|\mu_A^{(x)}(0) \otimes \mu_B^{(y)}(1)|\psi\rangle - \langle\psi|\mu_A^{(x)}(1) \otimes \mu_B^{(y)}(0)|\psi\rangle \end{aligned}$$

For $xy \in \{00, 01, 10\}$ the first two terms correspond to Alice and Bob winning using this strategy (since they give the same answer), whereas in the last two cases they lose (since their answers are different). For $xy = 11$ this is the other way around, confirming Eq. (3.2).

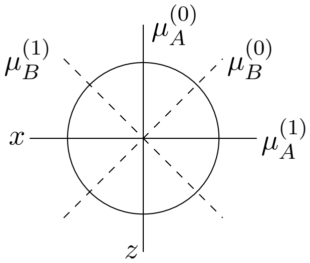
We now pick a smart strategy. Alice measures in the standard Z-basis $|0\rangle, |1\rangle$ for $x = 0$ or in the X-basis $|+\rangle, |-\rangle$ for $x = 1$. This corresponds to

$$O_A^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad O_A^{(1)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.3}$$

Slightly more complicatedly, we let Bob measure according to the basis $\vec{r} = \frac{1}{\sqrt{2}}(1, 0, 1)$ for $y = 0$ and $\vec{r} = \frac{1}{\sqrt{2}}(-1, 0, 1)$ for $y = 1$. In Exercise 3.1 you will verify that this corresponds to basis measurements in bases $\cos(\frac{\pi}{8})|0\rangle + \sin(\frac{\pi}{8})|1\rangle, -\sin(\frac{\pi}{8})|0\rangle + \cos(\frac{\pi}{8})|1\rangle$ for $y = 0$ or in the basis $\cos(-\frac{\pi}{8})|0\rangle + \sin(-\frac{\pi}{8})|1\rangle, -\sin(\frac{\pi}{8})|0\rangle + \cos(-\frac{\pi}{8})|1\rangle$ for $y = 1$. As observables, from Eq. (3.1) we get

$$O_B^{(0)} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad O_B^{(1)} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix}. \tag{3.4}$$

In Exercise 3.1 you will verify that this corresponds to basis measurements in bases $\cos(\frac{\pi}{8})|0\rangle + \sin(\frac{\pi}{8})|1\rangle, -\sin(\frac{\pi}{8})|0\rangle + \cos(\frac{\pi}{8})|1\rangle$ for $y = 0$ or in the basis $\cos(-\frac{\pi}{8})|0\rangle + \sin(-\frac{\pi}{8})|1\rangle, -\sin(\frac{\pi}{8})|0\rangle + \cos(-\frac{\pi}{8})|1\rangle$ for $y = 1$. If we draw the x, z-plane in the Bloch sphere, this corresponds to measuring along the following axes:



Finally, as quantum state we pick the maximally entangled state between Alice and Bob $|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Now, let us evaluate α from Eq. (3.2).

Using Lemma 2.18 it is a straightforward exercise, which we suggest you perform yourself in Exercise 3.2, to show that for $\vec{r}, \vec{s}$ in the x, z-plane

$$\langle\psi|O_A(\vec{r}) \otimes O_B(\vec{s})|\psi\rangle = \vec{r} \cdot \vec{s}.$$

It is easy to verify (which you should do yourself in Exercise 3.2) that Eq. (3.2) gives

$$\alpha = \frac{1}{\sqrt{2}}.$$

We therefore get a winning probability

$$p = \frac{1}{2}(\alpha + 1) = \frac{1}{2} + \frac{1}{2\sqrt{2}} = \cos^2(\frac{\pi}{8}) \approx 0.85\ldots$$

which is larger than the classical winning probability! We have proven:

Theorem 3.5. *The quantum value of the CHSH game G is at least*

$$\omega^*(G) \geq \frac{1}{2} + \frac{1}{2\sqrt{2}}$$

which is strictly larger than the classical value $\omega(G)$.

The Nobel Prize of 2022 was awarded to Aspect, Clauser and Zeilinger 'for experiments with entangled photons, establishing the violation of Bell inequalities and pioneering quantum information science' [34]. Clauser (the C in CHSH) proposed the experiments as encoded in the Bell game explained in this lecture [10]. Violations of the maximal classical winning probability were observed by experiments led by Aspect and Zeilinger [2, 46]. Many refinements have been made to these experiments to close as many as possible 'loopholes', see e.g. [21]. This means that we have very strong experimental evidence that Nature can not be described by a local hidden variable model! We must accept some form of quantum theory, or use nonlocal models which have significant conceptual drawbacks.[2]

Tsirelson bound

Above we saw a particular quantum strategy that beats all classical strategies. You may wonder whether we can do even *better* than the above strategy! The answer is that we can not, which is the content of the *Tsirelson bound.*

Theorem 3.6. *Let G be the CHSH game. Then*

$$\omega^*(G) \leq \frac{1}{2} + \frac{1}{2\sqrt{2}}.$$

Since we already showed by an explicit strategy that $\omega^*(G) \geq \frac{1}{2} + \frac{1}{2\sqrt{2}}$ we have equality! For the proof we recall the *Cauchy-Schwarz inequality*, which states that for any Hilbert space $\mathcal{H}$ and $\phi, \psi \in \mathcal{H}$ we have $|\langle\phi|\psi\rangle|^2 \leq \langle\phi|\phi\rangle\langle\psi|\psi\rangle$.

Proof. The key to this result is Eq. (3.2), which we derived for an *arbitrary* quantum strategy. Denote by $|\psi_{AB}\rangle$ the shared quantum state, which we may assume to be pure. We may also assume the measurement to be projective (as mentioned we will see later that this is always possible). If we construct the operators $O_A^{(x)}$ and $O_B^{(y)}$ from a projective two-outcome measurement, we see that $O_A^{(x)}$ and $O_B^{(y)}$ are Hermitian operators with eigenvalues in $\{-1, 1\}$. As a consequence $(O_A^{(x)})^2 = \mathbb{1}_A$ and $(O_B^{(y)})^2 = \mathbb{1}_B$. We define the operator

$$M_{AB} = O_A^{(0)} \otimes O_B^{(0)} + O_A^{(0)} \otimes O_B^{(1)} + O_A^{(1)} \otimes O_B^{(0)} - O_A^{(1)} \otimes O_B^{(1)}$$

so comparing with Eq. (3.2)

$$4\alpha = \langle\psi_{AB}|M_{AB}|\psi_{AB}\rangle$$

The Cauchy-Schwarz inequality gives

$$\begin{aligned}|\langle\psi_{AB}|M_{AB}|\psi_{AB}\rangle| &\leq \sqrt{\langle\psi_{AB}|M_{AB}M_{AB}^\dagger|\psi_{AB}\rangle}\sqrt{\langle\psi_{AB}|\psi_{AB}\rangle} \\ &= \sqrt{\langle\psi_{AB}|M_{AB}^2|\psi_{AB}\rangle}\end{aligned}$$

since $M_{AB} = M_{AB}^\dagger$ and $|\psi_{AB}\rangle$ is normalized. We now expand and rewrite

$$\begin{aligned}M_{AB}^2 &= \left(O_A^{(0)} \otimes (O_B^{(0)} + O_B^{(1)}) + O_A^{(1)} \otimes (O_B^{(0)} - O_B^{(1)})\right)^2 \\ &= (O_A^{(0)})^2 \otimes (O_B^{(0)} + O_B^{(1)})^2 + (O_A^{(1)})^2 \otimes (O_B^{(0)} - O_B^{(1)})^2 \\ &\quad + O_A^{(0)}O_A^{(1)} \otimes (O_B^{(0)} + O_B^{(1)})(O_B^{(0)} - O_B^{(1)}) + O_A^{(1)}O_A^{(0)} \otimes (O_B^{(0)} - O_B^{(1)})(O_B^{(0)} + O_B^{(1)}) \\ &= \mathbb{1}_A^2 \otimes ((O_B^{(0)} + O_B^{(1)})^2 + (O_B^{(0)} - O_B^{(1)})^2) \\ &\quad + O_A^{(0)}O_A^{(1)} \otimes (O_B^{(0)} + O_B^{(1)})(O_B^{(0)} - O_B^{(1)}) + O_A^{(1)}O_A^{(0)} \otimes (O_B^{(0)} - O_B^{(1)})(O_B^{(0)} + O_B^{(1)})\end{aligned}$$

[2] An example of a nonlocal classical model for quantum mechanics is *pilot-wave theory* as developed by Bell.

using that $(O_A^{(x)})^2 = \mathbb{1}_A$. In Exercise 3.4 you will show that using $(O_A^{(x)})^2 = \mathbb{1}_A$ and $(O_B^{(y)})^2 = \mathbb{1}_B$ this simplifies to

$$M_{AB}^2 = 4\mathbb{1}_A \otimes \mathbb{1}_B + (O_A^{(0)}O_A^{(1)} - O_A^{(1)}O_A^{(0)}) \otimes (O_B^{(0)}O_B^{(1)} - O_B^{(1)}O_B^{(0)}).$$

We are almost there! Note that, again by Cauchy-Schwarz,

$$|\langle\psi_{AB}|O_A^{(x)} \otimes O_B^{(y)}|\psi_{AB}\rangle| \leq \sqrt{\langle\psi_{AB}|(O_A^{(x)} \otimes O_B^{(y)})^2|\psi_{AB}\rangle} = 1$$

and hence

$$\langle\psi_{AB}|M_{AB}^2|\psi_{AB}\rangle \leq 8$$

Therefore, $4\alpha \leq \sqrt{8} = 2\sqrt{2}$ and $\omega^*(G) = \frac{1}{2} + \frac{1}{2}\alpha \leq \frac{1}{2} + \frac{1}{2\sqrt{2}}$. □

Outlook

Self-testing quantum states

By Theorem 3.6, we have found the optimal quantum strategy for the CHSH game! Another question you may have at this point is whether the one strategy we found is perhaps the *unique* strategy. This is clearly not the case! If we keep the same state but rotate our measurement operators by fixed unitaries U_A and U_B

$$\mu_A^{(x)}(a) \to U_A \mu_A^{(x)}(a) U_A^\dagger \quad \text{and} \quad \mu_B^{(y)}(b) = U_B \mu_B^{(y)}(b) U_B^\dagger$$

and we update

$$|\psi_{AB}\rangle \to (U_A \otimes U_B)|\psi_{AB}\rangle$$

then this clearly does not change the result! However, it turns out that *up to* similar trivial modifications (one can also use an isometry to a larger space) the strategy is unique! This fact is known as *self-testing*. The proof essentially goes by carefully studying the estimates made in the proof of Theorem 3.6 and proving that they can only be inequalities if $|\psi_{AB}\rangle$ and the measurements satisfy certain specific algebraic relations. It is moreover *robust* in the sense that if a strategy wins the game with probability close to $\omega^*(G)$, the strategy must also be 'close' in an appropriate sense to the exact strategy. This provides a black-box way to verify entanglement: suppose Alice and Bob claim to be able to produce maximally entangled pairs of qubits, and an outside referee wishes to verify this without having access to the quantum systems of Alice and Bob. The referee can verify their ability to generate entanglement simply by playing the CHSH game with them many times, and checking that they are able to win with probability close to the quantum value of the game. The only requirement is that Alice and Bob are not allowed to communicate after having received the questions from the referee. This is a useful tool in quantum cryptography, and is a basic building block in verification procedures of quantum computers. See [41] for a review of self-testing.

Correlations beyond quantum theory

We have seen that quantum theory allows for the existence of types of correlations that do not exist classically. Do there also exist correlations which go 'beyond' quantum theory? This is a question that we can sensibly ask in the context of Bell games. The key condition in our set-up of Bell games is that Alice and Bob are not allowed to communicate. This reflects the assumption, coming from the theory of relativity, that if they are sufficiently separated in space, they are not able to communicate due to the causal structure of space-time (i.e. because there is no faster-than-light communication). One might also investigate 'beyond quantum' correlations where the only condition is that the correlations that Alice and Bob share do not enable them to communicate information. That is, these are all possible correlations that are in principle consistent with the locality imposed by the causal structure of space-time.

Definition 3.7. A *non-signalling* strategy for a game G is a strategy where Alice and Bob answer according to probability distributions $p_{AB}(a,b|x,y) \in \Pr(\mathcal{A}\times\mathcal{B})$ given questions $x \in \mathcal{X}, y \in \mathcal{Y}$ which is such that

$$p_B(b|x,y) = p_B(b|x',y)$$

for all $x, x' \in \mathcal{X}$ and all $y \in \mathcal{Y}$

$$p_A(a|x,y) = p_A(a|x,y')$$

for all $x \in \mathcal{X}$ and all $y, y' \in \mathcal{Y}$.

The intuition behind this definition is that the answer Alice gives does not depend on the question that Bob received, so a reveals no information about y, and similarly b reveals no information about the question x. In Exercise 2.7 you already showed that quantum strategies are non-signalling. We can now easily construct a non-signalling strategy that always wins the CHSH game:

$$p(a,b|x,y) = \begin{cases} \frac{1}{2} & \text{if } x\cdot y = a+b \mod 2 \\ 0 & \text{else} \end{cases}$$

This shows that there are correlations *beyond quantum theory* which are compatible with causality (although there is no physical evidence for a theory beyond quantum mechanics supporting such correlations).

3.2 Exercises

3.1 **Basis measurements in the CHSH game:** On the single qubit Hilbert space $\mathcal{H} = \mathbb{C}^2$, define the states

$$|\psi_0(\theta)\rangle = \cos\theta|0\rangle + \sin\theta|1\rangle\ , \quad |\psi_1(\theta)\rangle = -\sin\theta|0\rangle + \cos\theta|1\rangle\ .$$

(a) Show that $\{|\psi_i\rangle\}_{i=0,1}$ is a basis for $\mathcal{H}$. What are the corresponding points on the Bloch sphere?

(b) Show that the observable O corresponding to a measurement in this basis takes the form

$$O = |\psi_0(\theta)\rangle\langle\psi_0(\theta)| - |\psi_1(\theta)\rangle\langle\psi_1(\theta)| = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}\ .$$

Verify that $O = Z$ for $\theta = 0$, $O = X$ for $\theta = \frac{\pi}{4}$. Verify that setting $\theta = \pm\pi/8$ yields the choice of observables in Eq. (3.4).

3.2 **The bias of the CHSH game:**

(a) Given $\vec{r} = (x, y, z), \vec{s} = (x', y', z')$ on the Bloch sphere, show that

$$\langle\psi|O_A(\vec{r}) \otimes O_B(\vec{s})|\psi\rangle = xx' - yy' + zz'$$

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

(b) Confirm that $\alpha = \frac{1}{\sqrt{2}}$ by computing the expression in Eq. (3.2) with the choices for $O_A^{(x)}$ and $O_B^{(y)}$ in Eq. (3.3) and Eq. (3.4).

(c) The goal of the following is to prove Tsirelson's bound in Theorem 3.6 for a restricted category of strategies. Namely, suppose that Alice and Bob share a maximally entangled qubit state $|\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |00\rangle)$, Alice measures along $\vec{r}_x$ and Bob measures along $\vec{s}_y$ for questions x, y and a choice of vectors on the Bloch sphere. Assume that $\vec{r}_x$ and $\vec{s}_y$ are all in the x, z-plane. Show that the bias of the corresponding strategy is given by

$$\alpha = \frac{1}{4}\left(\vec{r}_0 \cdot \vec{s}_0 + \vec{r}_0 \cdot \vec{s}_1 + \vec{r}_1 \cdot \vec{s}_0 - \vec{r}_1 \cdot \vec{s}_1\right).$$

(d) Show that

$$\alpha \leq \frac{1}{4}\left(\|\vec{s}_0 + \vec{s}_1\| + \|\vec{s}_0 - \vec{s}_1\|\right).$$

When do we have equality? *Hint: you can use the Cauchy-Schwarz inequality. For Bloch vectors $\vec{r}, \vec{s} \in \mathbb{R}^3$ it states that $|\vec{r} \cdot \vec{s}| \leq \|\vec{r}\|\|\vec{s}\|$. We have equality if and only if $\vec{r}$ is proportional to $\vec{s}$.*

(e) Next, show that

$$\|\vec{s}_0 + \vec{s}_1\| + \|\vec{s}_0 - \vec{s}_1\| \leq 2\sqrt{2}$$

When do we have equality? *Hint: show that $\|\vec{s}_0 + \vec{s}_1\| + \|\vec{s}_0 - \vec{s}_1\| = \sqrt{\gamma + \delta} + \sqrt{\gamma - \delta}$ for the real numbers $\gamma = \|\vec{s}_0\|^2 + \|\vec{s}_1\|^2$ and $\delta = 2s_0 \cdot s_1$. Square the result to find that it is minimal for $\delta = 0$.*

(f) Conclude that $\alpha \leq \frac{1}{\sqrt{2}}$. When is the strategy optimal (i.e. when do we have equality)?

3.3 **The GHZ game:** Let $\mathcal{H}_{ABC} = \mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2$ be a Hilbert space of three qubits shared between three parties. Define the *Greenberger-Horne-Zeilinger* (GHZ) state to be

$$|\text{GHZ}\rangle := \frac{1}{\sqrt{2}}\big(|000\rangle + |111\rangle\big) .$$

(a) Show that the reduced state on A is given by

$$\text{tr}_{BC}\left[|\text{GHZ}\rangle\langle\text{GHZ}|\right] = \frac{1}{2}\mathbb{1}_A .$$

So, the reduced state on A is the maximally mixed state τ_A.

(b) Define the operators

$$\begin{aligned} M_0 &= X \otimes X \otimes X \ , \\ M_1 &= X \otimes Y \otimes Y \ , \\ M_2 &= Y \otimes X \otimes Y \ , \\ M_3 &= Y \otimes Y \otimes X \ , \end{aligned}$$

where X and Y are the Pauli matrices. Show that $|\text{GHZ}\rangle$ is an eigenstate of each of these operators, so

$$M_i|\text{GHZ}\rangle = \lambda_i|\text{GHZ}\rangle \ , \quad i = 0, 1, 2, 3 \ ,$$

for some λ_i which you should determine.

The GHZ game is a *three-player* Bell game in which the referee gives a question x, y, z in $\{0,1\}$ to each of Alice, Bob, and Charlie respectively. The referee either gives 0 to all three players, or gives 0 to one of them and 1 to the other two. The players then output bits in $\{0,1\}$ which we call a, b, and c.

The winning condition for the GHZ game is that

$$x \,\mathrm{OR}\, y \,\mathrm{OR}\, z = a + b + c \mod 2 \; .$$

This is summarised in the following table ($\oplus$ denotes addition modulo 2).

x	y	z	winning condition
0	0	0	$a \oplus b \oplus c = 0$
0	1	1	$a \oplus b \oplus c = 1$
1	0	1	$a \oplus b \oplus c = 1$
1	1	0	$a \oplus b \oplus c = 1$

(c) Suppose that Alice, Bob, and Charlie use a classical deterministic strategy, so $a = f(x)$, $b = g(y)$, $c = h(z)$.
By considering the sum

$$\sum_{\text{questions}} (f(x) + g(y) + h(z)) \mod 2 \; ,$$

or otherwise, show that the maximum winning probability is $\omega(G) = 3/4$.

(d) Now suppose that Alice, Bob, and Charlie share the 3-party state $|\mathrm{GHZ}\rangle$. Each of them adopts the following strategy:
If 0 is received, measure X. If 1 is received, measure Y. If the obtained output is $+1$ then give answer 0, else give the answer 1.
Show that, using this quantum strategy, the players can win the game with certainty – that is, $\omega^*(G) = 1$.

3.4 **Tsirelson inequality:** Verify the claim in the proof of Theorem 3.6

3.5 **Shared randomness:** Prove Lemma 3.2. You may assume that Λ is a finite set.

3.6 **Pauli operators:** Let $A, B \in \mathrm{Lin}(\mathbb{C}^2)$ be Hermitian matrices satisfying $A^2 = B^2 = \mathbb{1}$, and $AB = -BA$.

(a) Let $C = -iAB$. Show that $C^2 = \mathbb{1}$.

(b) Show that there exists a unitary U such that

$$UAU^\dagger = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = Z \; , \quad UBU^\dagger = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = X \; , \quad UCU^\dagger = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = Y \; .$$

Comment: this is the type of fact required to prove self-testing of the CHSH game! The idea is that if a strategy has optimal winning probability, this imposes $O_A^{(0)} O_A^{(1)} + O_A^{(1)} O_A^{(0)} = 0$ which then implies that up to a change of basis $O_A^{(0)}$ $O_A^{(1)}$ are the matrices Z and X.

3.7 **The magic square game:** Consider the following "magic square":

$$\begin{array}{ccccccc}
\square & \times & \square & \times & \square & = & +1 \\
\times & & \times & & \times & & \\
\square & \times & \square & \times & \square & = & +1 \\
\times & & \times & & \times & & \\
\square & \times & \square & \times & \square & = & +1 \\
\shortparallel & & \shortparallel & & \shortparallel & & \\
-1 & & -1 & & -1 & &
\end{array}$$

Each of the boxes can be filled in with -1 or $+1$. A *solution* to this magic square is a filling such that each of the rows multiplies to $+1$, and each of the columns multiplies to -1.

Referee Robin, an avid fan of Bell games, challenges Alice and Bob to the following task. Alice and Bob are given questions x, y respectively from $\Omega_X = \Omega_Y = \{1, 2, 3\}$. Alice must then respond with a filling for row x, and Bob must respond with a filling for column y. In other words, they give Robin answers $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ respectively from $\{-1, 1\}^3$.

Alice and Bob *win* the game if three conditions are satisfied:

- The product of the elements of Alice's answer is $+1$, $\prod_{i=1}^{3} a_i = +1$.
- The product of the elements of Bob's answer is -1, $\prod_{j=1}^{3} b_j = -1$.
- They agree on the overlapping element, $a_y = b_x$.

In other words, they win if their answers could form part of a solution to the magic square.

(a) Convince yourself that $\omega(G) = 1$ if and only if there exists a solution to the above magic square.

(b) Prove that there is no solution to the above magic square.

Alice and Bob furiously attempt to devise a quantum strategy to beat Robin. They produce the following construction, known as the *Mermin-Peres magic square.*

$\mathbb{1} \otimes Z$	$Z \otimes \mathbb{1}$	$Z \otimes Z$
$X \otimes \mathbb{1}$	$\mathbb{1} \otimes X$	$X \otimes X$
$-X \otimes Z$	$-Z \otimes X$	$Y \otimes Y$

Each element of this square is a tensor product of Pauli operators on $\mathbb{C}^2 \otimes \mathbb{C}^2$.

(c) Show that the operators of a given row or column of the Mermin-Peres square commute with each other. Deduce that they can be simultaneously diagonalised.

(d) Show that the product of all the operators on a given row is always $+1$, and that the product of all the operators on a given column is always -1.

(e) Suppose Alice and Bob now share the 4-qubit entangled state

$$|\psi\rangle = \frac{1}{\sqrt{2}}\big(|0\rangle_{A_1} \otimes |0\rangle_{B_1} + |1\rangle_{A_1} \otimes |1\rangle_{B_1}\big) \otimes \frac{1}{\sqrt{2}}\big(|0\rangle_{A_2} \otimes |0\rangle_{B_2} + |1\rangle_{A_2} \otimes |1\rangle_{B_2}\big) ,$$

where Alice has access to the A_1 and A_2 systems, and Bob has access to the B_1 and B_2 systems.

Now, upon receiving their questions from Robin, Alice and Bob each measure their two qubits with the operators from the corresponding row or column in the Mermin-Peres magic square to determine their outputs. For example, if Alice receives $x = 3$, she measures $-X \otimes Z$, $-Z \otimes X$, and $Y \otimes Y$, and answers (a_1, a_2, a_3) according to the outcomes.

Show that with this quantum strategy, Alice and Bob will win the magic square game with probability $\omega^*(G) = 1$.

Lecture 4

Classical and quantum processing

[MW: Notation: $\Phi \to \mathcal{M}, N, \ldots, \mathcal{T}, \ldots, \mathcal{E}, \ldots$.]

Concept	Math translation
Quantum operations map quantum states to quantum states.	Quantum channel Φ = completely positive trace preserving map, such that $\Phi \otimes \mathcal{I}$ maps quantum states to quantum states.
When is a superoperator a quantum channel?	Theorem 4.21: characterization of CPTP maps.
One only needs to check positivity for maximally entangled state.	Φ is completely positive if and only if the Choi matrix $J(\Phi)$ is positive.
Every quantum operation arises from applying isometries and discarding a subsystem.	Φ is a quantum channel if and only if it has a Stinespring representation $$\Phi[\rho] = \operatorname{tr}_E[V \rho V^\dagger].$$
A quantum operation applies a linear map with some probability.	Φ is a quantum channel if and only if it has a Kraus representation $$\Phi[\rho] = \sum_i X_i \rho X_i^\dagger.$$

We have introduced the formalism of quantum states. We have also seen a first application: entanglement gives rise to correlations that cannot be explained by a classical model.

So far, our abstract model of quantum mechanics is able to describe the state of a system, and how to perform measurements of the system. This is a *static* picture. One ingredient is still missing: what *dynamics* are possible in quantum systems? In other words, given a quantum system, what is the class of operations we can apply to it? This is crucial to *quantum information processing*.

We will start by describing two important special cases and then make a general proposal for a general class of quantum operations called *quantum channels*. We will then prove a structure theorem that gives a classification of quantum channels and suggests different representations of such channels.

Classical channels

For our first example we go back to classical states. What are possible dynamics on classical probability distributions? Given X, Y with outcome sets Ω_X, Ω_Y, and suppose that Y is the result from applying some operation to X. If we start with a fixed $x \in \Omega_X$ we get outcome $y \in \Omega_Y$ with probability $q(y|x)$. If we now start with an arbitrary distribution p_X on X, we see that we get outcome y with probability

$$p_Y(y) = \sum_x q(y|x) p_X(x)$$

since we have outcome x with probability $p(x)$, and given x we obtain y with probability $q(y|x)$. This leads to the following:

Definition 4.1. A *classical channel* from X to Y is a map

$$Q : \mathrm{P}(X) \to \mathrm{P}(Y)$$

where $p_Y = Q(p_X)$ for $p_X \in \mathrm{P}(X)$ is given by

$$p_Y(y) = \sum_x q(y|x) p_X(x)$$

for some collection of $q(y|x) \in \mathbb{R}_{\geq 0}$ such that

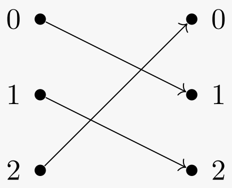
$$\sum_y q(y|x) = 1 \quad \text{for all } x$$

The terminology *channel* is natural in the context of information theory, where one can have in mind the example of sending information from X to Y over some medium (radio waves, electric cable,…).

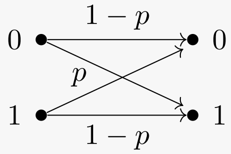
Example 4.2. Any function $f : \Omega_X \to \Omega_Y$ induces a channel by letting

$$q(y|x) = \begin{cases} 1 & \text{if } f(x) = y \\ 0 & \text{otherwise.} \end{cases}$$

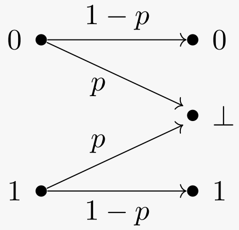
For example, the below diagram visualizes the classical channel where $\Omega_X = \Omega_Y = \{0, 1, 2\}$ and $f(i) = i + 1 \mod 3$.



Example 4.3. Let X and Y be bits. The *binary symmetric channel* is the channel which with some probability p flips the value of the bit



which can be thought of as a noisy communication channel. For instance, you can think of a cable which transmits a bit over some spatial distance, and with probability $1-p$ the bit arrives correctly, while with probability p it gets corrupted. A closely related example is the case where a bit gets corrupted with probability p but whenever this happens we *know* that this has happened (for instance in the cable example there does not arrive a message at all). This situation is described by the *binary erasure channel* in which case the system Y has an extra outcome symbol $\perp$ denoting a corruption:



Unitaries and isometries

For our second basic example we will look at *pure* states. We consider a natural class of dynamics which map pure states to pure states. Let us assume that we map from a quantum system A to a system B and that

(a) This operation is linear, so it preserves superpositions, and hence it has to be given by $|\psi\rangle \mapsto V|\psi\rangle$ for some $V \in \mathrm{Lin}(A, B)$

(b) It must send states to states, so $V|\psi\rangle$ must be normalized.

The second condition implies that the map V should be an isometry, and if $|A| = |B|$ it will be unitary.

If you have taken a course on quantum mechanics or quantum computing it will be familiar to you that dynamics of (pure) quantum states are given by unitary maps. Formulated as an action on density matrices we have

$$|\psi\rangle \mapsto U|\psi\rangle$$
$$\rho = |\psi\rangle\langle\psi| \mapsto U|\psi\rangle\langle\psi|U^\dagger.$$

Remark 4.4. We make a small side tour to comment on how this relates to how physicists model dynamics in quantum physics. In standard quantum mechanics one has continuous time and unitary dynamics happens by continuous time evolution for some time t. The quantum system evolves according to a Hamiltonian $H(t)$, which is a (possibly time-dependent) self-adjoint operator $H(t) \in \mathrm{Lin}(\mathcal{H})$. The quantum state of the system $|\psi_t\rangle$ at time t evolves as

$$\frac{\mathrm{d}|\psi_t\rangle}{\mathrm{d}t} = -iH(t)|\psi_t\rangle.$$

This is the *Schrödinger equation*. Given the initial state $|\psi_0\rangle$, this is a differential equation that can be solved to determine the state at all times $t > 0$. This solution is such that there exists a continuous family of unitary maps U_t such that $|\psi_t\rangle = U_t|\psi_0\rangle$ for any choice of initial state $|\psi_0\rangle$. Conversely, for any given unitary map $U \in \mathrm{U}(\mathcal{H})$ one can find a Hamiltonian H such that time evolving along H for time $t = 1$ one obtains U (you can think about this in Exercise 4.13). The details of this are unimportant to us right now, the only thing to take home is that *in principle* quantum mechanics allows for any unitary map to be realized by a physical system. In practice, it may be really hard to engineer a quantum system such that its Hamiltonian gives rise to a desired unitary dynamics, but this will not concern us in this course!

4.1 Operations on states

The two special cases of operations preserving classical states and operations preserving pure states have the key property that they *send states to states*. Moreover, in the case of mixed classical states it was natural to impose *linearity*. We will now argue that these two properties determine the full class of operations on (arbitrary) quantum states. So, what we are looking for is a class of maps sending states to states. Let $\Phi_{A\to B}$ a map from $\mathrm{S}(A)$ to $\mathrm{S}(B)$. In order to represent a physical process such a map must preserve mixtures of states: if we have a system which is in state ρ_i with probability p_i (so $\rho = \sum_i p_i\rho_i$) then applying dynamics to it should have the same effect as applying the dynamics to each of the ρ_i separately, and weighing by probability p_i:

$$\Phi_{A\to B}\left(\sum_i p_i\rho_i\right) = \sum_i p_i\Phi_{A\to B}(\rho_i). \tag{4.1}$$

This means (see Exercise 4.6) that we can extend $\Phi_{A\to B}$ to a linear map!

What may be a little confusing at first encounter is that the set of quantum states is a set of linear operators itself. To make the distinction between operators as linear maps representing quantum states and maps that should be seen as dynamics, we will call a linear map between spaces of linear maps a *superoperator*.

Definition 4.5. If A and B are quantum systems with Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, a superoperator from A to B is a linear map

$$\Phi_{A\to B} : \mathrm{Lin}(A) \to \mathrm{Lin}(B).$$

If $A = B$ we write $\Phi_{A\to A} = \Phi_A$.

Example 4.6. Here are a three basic but essential examples of superoperators.

(a) The easiest example of a superoperator is the *identity superoperator*. If we have a quantum system A, then we let $\mathcal{I}_A$ denote the superoperator defined by $\mathcal{I}_A(M_A) = M_A$ for all $M_A \in \mathrm{Lin}(A)$.

(b) If we have two quantum systems A and B, then taking the partial trace over B defines a superoperator, mapping

$$\mathrm{tr}_B : \mathrm{Lin}(AB) \to \mathrm{Lin}(A).$$

(c) If $V \in \mathrm{Isom}(A, B)$ is an isometry, then

$$M_A \in \mathrm{Lin}(A) \mapsto V M_A V^\dagger \in \mathrm{Lin}(B)$$

defines a superoperator.

We can also take compositions and tensor products of superoperators. If we have superoperators $\Phi_{A\to B}$ and $\Psi_{B\to C}$ then $\Psi_{B\to C} \circ \Phi_{A\to B}$ is a superoperator from A to C. If $\Phi_{A\to B}$ and $\Psi_{C\to D}$ then we may define a tensor product superoperator $\Phi_{A\to B} \otimes \Psi_{C\to D}$ from AC to BD by linear extension of

$$\Phi_{A\to B} \otimes \Psi_{C\to D}(M_A \otimes M_C) = \Phi_{A\to B}(M_A) \otimes \Psi_{C\to D}(M_C)$$

for $M_A \in \mathrm{Lin}(A)$ and $M_C \in \mathrm{Lin}(C)$.

Example 4.7. Taking the trace of an operator is also a superoperator tr (where we identify $\mathbb{C} \cong \mathrm{Lin}(\mathbb{C})$). Then, if we have quantum systems A and B we may identify the partial trace over B with

$$\mathrm{tr}_B = \mathcal{I}_A \otimes \mathrm{tr}\,.$$

Which superoperators can represent dynamics on a quantum system? Clearly, a minimal condition is that it must send quantum states to quantum states. In particular (by linearity) this implies that it must send positive operators to positive operators, which leads to the following definition:

Definition 4.8. A superoperator $\Phi_{A\to B}$ is a *positive map* (or *positivity preserving map*) if it maps every positive operator $P_A \in \mathrm{PSD}(A)$ to a positive operator $\Phi_{A\to B}(P_A) \in \mathrm{PSD}(B)$.

The terminology 'positive map' can be a little confusing: note that a positive map is *not* the same as the condition that $\Phi \in \mathrm{PSD}(\mathrm{Lin}(\mathcal{H}))$, when interpreting Φ as a linear map and giving a Hilbert space structure to $\mathrm{Lin}(\mathcal{H})$. The three examples we gave in Example 4.6 are all positive maps. This is trivial for the identity superoperator. For the partial trace we already saw this fact before, in Lemma 2.5. Finally, if we have any $V \in \mathrm{Lin}(A, B)$, the superoperator

$$M_A \in \mathrm{Lin}(A) \mapsto V M_A V^\dagger \in \mathrm{Lin}(B)$$

is a positive map by Corollary A.3.

It is not hard to come up with examples of superoperators which are not positive maps (do so yourself!). A basic observation is

Lemma 4.9. *If $\Phi_{A\to B}$ and $\Psi_{B\to C}$ are positive maps, the composition $\Psi_{B\to C}\circ\Phi_{A\to B}$ is a positive map.*

However, and perhaps surprisingly, *the condition of positivity is not enough!* To see why this is the case, suppose that $\Phi_{A\to A'}$ and $\Psi_{B\to B'}$ represent some quantum dynamics, then their tensor product must as well, as this should describe the dynamics on the joint system AB. It turns out that being a positive map is *not* preserved under taking tensor products of superoperators. To see this, we consider the following example, which is also explored in Exercise 2.16. Let T_A be the superoperator on a qubit system A defined by $T_A(M_A) = M_A^{\mathsf{T}}$ for $M_A \in \mathrm{Lin}(A)$. Then it is easy to see that if $\rho_A \in \mathrm{S}(A)$ we also have $\rho_A^{\mathsf{T}} \in \mathrm{S}(A)$ so T_A is a positive map. However, let B be another qubit system, and let $|\Phi^+_{AB}\rangle$ be the maximally entangled state and $\rho_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|$. If we apply T_A on the A system and the identity superoperator on the B system we find

$$\begin{aligned}(T_A\otimes\mathcal{I}_B)(\rho_{AB}) &= \frac{1}{2}(T_A\otimes\mathcal{I}_B)(|00\rangle\langle 00| + |11\rangle\langle 11| + |00\rangle\langle 11| + |11\rangle\langle 00|)\\ &= \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11| + |10\rangle\langle 01| + |01\rangle\langle 10|)\\ &= \frac{1}{2}\begin{pmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{pmatrix}\end{aligned}$$

As one can see directly from the determinant, the resulting operator is not positive! This suggests the following definition:

Definition 4.10. A superoperator $\Phi_{A\to B}$ is a *completely positive* (CP) map if for any reference system R the superoperator $\Phi_{A\to B}\otimes\mathcal{I}_R$ is positive. We denote the set of completely positive maps from A to B by $\mathrm{CP}(A,B)$ and abbreviate $\mathrm{CP}(A,A) = \mathrm{CP}(A)$.

By our above discussion this is a *necessary* requirement for a valid quantum operation. It turns out that this set is closed under tensor products as well as under composition.

Lemma 4.11. *(a) Suppose $\Phi_{A\to B}\in\mathrm{CP}(A,B)$ and $\Psi\in\mathrm{CP}(B,C)$, then $\Psi_{B\to C}\circ\Phi_{A\to B}\in\mathrm{CP}(A,C)$.*

(b) Suppose $\Phi_{A\to A'}\in\mathrm{CP}(A,A')$ and $\Psi_{B\to B'}\in\mathrm{CP}(B,B')$, then

$$\Phi_{A\to A'}\otimes\Psi_{B\to B'}\in\mathrm{CP}(AB,A'B')$$

Proof. By Lemma 4.9 (a) follows from

$$(\Psi_{B\to C}\circ\Phi_{A\to B})\otimes\mathcal{I}_R = (\Psi_{B\to C}\otimes\mathcal{I}_R)\circ(\Phi_{A\to B}\otimes\mathcal{I}_R).$$

For (b) we note that it follows from the definition that $\Phi_{A\to A'}\otimes\mathcal{I}_B$ is CP as well as $\mathcal{I}_{A'}\otimes\Psi_{B\to B'}$, and therefore by (a)

$$\Phi_{A\to A'}\otimes\Psi_{B\to B'} = (\mathcal{I}_{A'}\otimes\Psi_{B\to B'})\circ(\Phi_{A\to A'}\otimes\mathcal{I}_B)$$

is CP as well. □

To make sure that a superoperator $\Phi_{A\to B}$ maps quantum states to quantum states we additionally need that it maps operators with $\mathrm{tr}[M_A] = 1$ to an operator with $\mathrm{tr}[\Phi_{A\to B}(M_A)] = 1$. By linearity this implies that we must demand the following condition:

Definition 4.12. A superoperator $\Phi_{A\to B} : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$ is called *trace preserving* (TP) if

$$\mathrm{tr}[\Phi_{A\to B}(M_A)] = \mathrm{tr}[M_A]$$

for all $M_A \in \mathrm{Lin}(A)$.

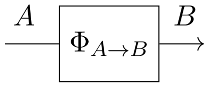
If we have a superoperator which is both completely positive and trace preserving (CPTP) we will also call such a superoperator a *quantum channel*. We denote the set of quantum channels, or CPTP maps, from A to B by

$$\mathrm{C}(A,B) = \{\Phi_{A\to B} \in \mathrm{CP}(A,B) \cap \mathrm{TP}(A,B)\}.$$

The three examples in Example 4.6 are in fact all quantum channels! For the identity superoperator (which we from now on will also call the *identity channel*) this is clear. For the partial trace, we note that if we tensor with the identity channel we again have a partial trace, so it is still positivity preserving. Finally, if we have the superoperator $\Phi^V_{A\to B}$ which is given by application of an isometry $V \in \mathrm{Isom}(A,B)$, then upon tensoring with the identity channel on a reference system R this yields the superoperator $\Phi^{V\otimes\mathbb{1}_R}_{AR\to BR}$ for $V \otimes \mathbb{1}_R \in \mathrm{Isom}(AR,BR)$ which is again positive. We now come to our final axiom of quantum theory:

Axiom 5. The set of possible operations from a quantum system A to a quantum system B is given by the set of channels $\mathrm{C}(A,B)$.

We will graphically denote a quantum channel by a box acting on a quantum system as follows:



Example 4.13. Here is another important example: the *depolarizing channel* with noise parameter p. Consider a quantum system A, then $\mathcal{D}_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A)$ is the superoperator given by

$$\mathcal{D}_p(M_A) = (1-p)M_A + p\,\mathrm{tr}[M_A]\tau_A$$

where τ_A is the maximally mixed state $\tau_A = \frac{1}{|A|}\mathbb{1}_A$. You can roughly think of this as a quantum analog of the binary symmetric channel in Example 4.3. It models the situation where the state is unchanged with probability $1-p$ and with probability p it gets lost and is replaced by a uniformly random state. You will show in Exercise 4.8 that this indeed is a quantum channel.

There are many more examples! This table collects a few. You can prove they are legitimate quantum channels in Exercise 4.8 and Exercise 4.11

Quantum operation	Math translation as quantum channel
Lose all information with some probability p.	The *depolarizing channel* $D_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A)$ $$M_A \mapsto (1-p)M_A + p\,\mathrm{tr}[M_A]\tau_A.$$
Lose coherent information, dampening of the off-diagonal terms in the density matrix.	The *dephasing channel* $\mathcal{P}_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A)$ $$M_A \mapsto (1-p)M_A + p\sum_{a\in\mathcal{A}} \langle a\|M_A\|a\rangle \|a\rangle\langle a\|.$$
Get an error with probability p, and check whether it occured.	The *erasure channel* $\mathcal{E}_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A')$ where $\mathcal{H}_{A'} = \mathcal{H}_A \oplus \mathrm{span}\{\|\perp\rangle\}$ $$M_A \mapsto (1-p)M_A + p\,\mathrm{tr}[M_A]\|\perp\rangle\langle\perp\| \ .$$
Discard and replace by a fixed state.	The *replacement channel* $\mathcal{R}_\rho : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$, for $\rho_B \in S(B)$, given by $$M_A \mapsto \mathrm{tr}[M_A]\rho_B \ .$$
Apply a random unitary.	If we have $U_i \in \mathrm{U}(A)$ with probability p_i for $i = 1, \ldots, r$ $$M_A \mapsto \sum_{i=1}^{r} p_i U_i M_A U_i^\dagger.$$

4.2 Characterization of quantum channels

The definition of a quantum channel is rather cumbersome: we have to verify that for *any* reference system the superoperator $\Phi_{A\to B} \otimes \mathcal{I}_R$ is positive. Fortunately, we can describe the class of quantum channels in various more concrete ways, which will be a powerful tool both for concrete examples and for theoretical arguments.

The Choi operator

If we have a positive superoperator $\Phi_{A\to B}$, then if we want to test whether $\Phi_{A\to B} \otimes \mathcal{I}_R$ is positive we need to test it on states which are entangled between A and R (see Exercise 4.2). In particular, we could choose a system $R = A'$ which is a copy of A and check that it maps the maximally entangled state to a quantum state. The Choi operator is the result of applying $\Phi_{A\to B} \otimes \mathcal{I}_{A'}$ to an unnormalized maximally entangled state

Definition 4.14. Given a superoperator $\Phi_{A\to B}$, let A' be a quantum system with $\mathcal{H}_{A'} = \mathcal{H}_A$. Let $\mathcal{A}$ be an orthonormal basis for $\mathcal{H}_A = \mathcal{H}_{A'}$, then the Choi operator $J(\Phi) \in \mathrm{Lin}(BA)$ is defined as

$$J(\Phi) = \sum_{a,a' \in \mathcal{A}} \Phi_{A\to B}(|a\rangle\langle a'|) \otimes |a\rangle\langle a'|.$$

Note that if we let $|\Phi^+_{AA'}\rangle = \frac{1}{\sqrt{|A|}} \sum_a |aa\rangle$, then

$$J(\Phi) = |A|(\Phi_{A\to B} \otimes \mathcal{I}_{A'})(|\Phi^+_{AA'}\rangle\langle\Phi^+_{AA'}|).$$

This implies that if $\Phi_{A\to B}$ is a quantum channel, its Choi operator must be positive. We will see later this section that this is also a sufficient condition!

A nice feature of the Choi operator is that it completely determines the superoperator. This is easy to see: by linearity it suffices to know how $\Phi_{A\to B}$ acts on the matrices $|a\rangle\langle a'|$ (since these form a basis for $\mathrm{Lin}(A)$) and this information can be inferred from $J(\Phi)$.

The following shows how to recover $\Phi_{A\to B}$ from $J(\Phi)$ (for an arbitrary superoperator, not just for a quantum channel).

Lemma 4.15 (Choi isomorphism)**.** *Suppose $J(\Phi)$ is the Choi operator of a superoperator $\Phi_{A\to B}$. Then for $M_A \in \mathrm{Lin}(A)$*

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_{A'}[(\mathbb{1}_B \otimes M_A^\mathsf{T})J(\Phi)]$$

where the transpose is computed with respect to the choice of basis in the definition of the Choi operator.

The proof is Exercise 4.3.

Characterization of complete positivity

We now come to a characterization of completely positive maps.

Theorem 4.16 (Characterization of CP maps)**.** *Suppose $\Phi_{A\to B}$ is a superoperator. Then following statements are equivalent:*

(a) *$\Phi_{A\to B}$ is completely positive.*

(b) *The Choi operator $J(\Phi)$ is positive.*

(c) *There exists a collection of operators $\{X_i \in \mathrm{Lin}(A,B)\}_{i=1}^r$ such that*

$$\Phi_{A\to B}(M_A) = \sum_{i=1}^r X_i M_A X_i^\dagger$$

This is known as a Kraus representation.

(d) *There exists a quantum system E and an operator $V \in \mathrm{Lin}(A, BE)$ such that*

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_E[V M_A V^\dagger].$$

This is known as a Stinespring representation.

Proof. As we already observed, (a) implies (b). For the implication (b) $\Rightarrow$ (c), suppose that $J(\Phi) \in \mathrm{Lin}(BA')$ is positive. Then there exists a decomposition

$$J(\Phi) = \sum_{i=1}^{r} |v_i\rangle\langle v_i|$$

where $v_i \in \mathcal{H}_B \otimes \mathcal{H}_{A'}$ need not be normalized. Write

$$|v_i\rangle = \sum_{a,b} v_{i,ba} |b\rangle|a\rangle$$

and let

$$X_i = \sum_{a,b} v_{i,ba} |b\rangle\langle a| = \sum_{a,b} \langle ba|v_i\rangle |b\rangle\langle a|.$$

Then, by Lemma 4.15 for $M_A \in \mathrm{Lin}(A)$

$$\begin{aligned}
\Phi_{A\to B}(M_A) &= \sum_i \mathrm{tr}_A\left[(\mathbb{1}_B \otimes M_A^{\mathsf{T}})|v_i\rangle\langle v_i|\right] \\
&= \sum_i \sum_{a,b} \sum_{a',b'} v_{i,ba}\overline{v_{i,b'a'}}\, \mathrm{tr}_A\left[(\mathbb{1}_B \otimes M_A^{\mathsf{T}})|ba\rangle\langle b'a'|\right] \\
&= \sum_i \sum_{a,b} \sum_{a',b'} v_{i,ba}\overline{v_{i,b'a'}}\, \mathrm{tr}\left[M_A^{\mathsf{T}}|a\rangle\langle a'|\right] |b\rangle\langle b'| \\
&= \sum_i \sum_{a,b} \sum_{a',b'} v_{i,ba}\overline{v_{i,b'a'}}\, \underbrace{\langle a'|M_A^{\mathsf{T}}|a\rangle}_{=\langle a|M_A|a'\rangle} |b\rangle\langle b'|
\end{aligned}$$

and hence

$$\begin{aligned}
\Phi_{A\to B}(M_A) &= \sum_i \sum_{a,b} \sum_{a',b'} v_{i,ba} |b\rangle\langle a|M_A|a'\rangle\langle b'| \overline{v_{i,b'a'}} \\
&= \sum_i X_i M_A X_i^\dagger.
\end{aligned}$$

For the implication (c) $\Rightarrow$ (d) we let $E = \mathbb{C}^r$ with basis $\{|i\rangle\}_{i=1}^r$ and we let

$$V = \sum_{i=1}^{r} X_i \otimes |i\rangle$$

then it is clear that

$$\mathrm{tr}_E[V M_A V^\dagger] = \sum_{i=1}^{r} X_i M_A X_i^\dagger.$$

Finally, the partial trace and $M_A \mapsto V M_A V^\dagger$ are completely positive, so by Lemma 4.11 (d) implies (a). □

A first observation is that the Choi operator is positive if $\Phi_{A\to B} \otimes \mathcal{I}_A$ is positive, so we have

Corollary 4.17. *A superoperator $\Phi_{A\to B}$ is a completely positive map if and only if $\Phi_{A\to B} \otimes \mathcal{I}_A$ is a positive map.*

So, while the original definition allows for an arbitrary reference system R, which could have arbitrarily large dimension, we find that it suffices to show positivity when the reference system R is a copy of A.

Remark 4.18. The Kraus and Stinespring representations are in general not unique. In the proof of Theorem 4.16 we constructed them from the Choi operator, but often one can find a Kraus or Stinespring representation directly from the description of the channel. In Exercise 4.14 and Exercise 4.15 you will investigate the freedom of choice there is for the Kraus and Stinespring representations for $\Phi_{A\to B} \in \mathrm{C}(A, B)$.

Example 4.19. Let us compute the Choi operator, and Kraus and Stinespring representations for a concrete example. We take the *completely dephasing channel* $\mathcal{P} = \mathcal{P}_1$ on a quantum system A, defined by

$$M_A \mapsto \sum_{a\in\mathcal{A}} \langle a|M_A|a\rangle |a\rangle\langle a|$$

given a basis $\mathcal{A}$, typically the standard basis for $\mathcal{H}_A = \mathbb{C}^{|A|}$. In other words, it sets all off-diagonal terms in the density matrix to zero. The Choi operator is given by

$$\begin{aligned} J(\mathcal{P}) &= \sum_{a,a'} \mathcal{P}(|a\rangle\langle a'|) \otimes |a\rangle\langle a'| \\ &= \sum_{a} |a\rangle\langle a| \otimes |a\rangle\langle a|. \end{aligned}$$

Apart from normalization, this is a maximally correlated state $\frac{1}{|A|}\sum_a |a\rangle\langle a| \otimes |a\rangle\langle a|$.

From the definition of the channel we see that

$$\mathcal{P}(M_A) = \sum_a |a\rangle\langle a|M_A|a\rangle\langle a|$$

so the operators $X_a = |a\rangle\langle a|$ for $a \in \mathcal{A}$ gives a Kraus representation. Finally, from the construction in Theorem 4.16, as a Stinespring representation we may take E to be a copy of A and

$$V = \sum_a |aa\rangle\langle a|.$$

Characterization of quantum channels

Theorem 4.16 characterizes when a superoperator $\Phi_{A\to B}$ is a completely positive map. We now identify the appropriate conditions for when $\Phi_{A\to B}$ is a quantum channel (i.e. when it is also trace preserving).

We note the following basic fact, which follows from Exercise 1.14:

Lemma 4.20. *Suppose $N_A \in \mathrm{Lin}(A)$. Then $N_A = \mathbb{1}_A$ if and only if for all $M_A \in \mathrm{Lin}(A)$ we have $\mathrm{tr}[N_A M_A] = \mathrm{tr}[M_A]$.*

We use this to determine the conditions under which a completely positive map is also trace-preserving and hence a quantum channel.

Theorem 4.21 (Characterization of quantum channels)**.** *Suppose* $\Phi_{A\to B} \in \mathrm{CP}(A, B)$*. Then following statements are equivalent:*

(a) $\Phi_{A\to B} \in \mathrm{C}(A, B)$ *(i.e., as well as being completely positive it is also trace preserving).*

(b) The Choi operator $J(\Phi)$ *is such that*

$$\mathrm{tr}_B[J(\Phi)] = \mathbb{1}_{A'}.$$

(c) If

$$\Phi_{A\to B}(M_A) = \sum_{i=1}^{r} X_i M_A X_i^\dagger$$

is a Kraus representation, then

$$\sum_{i=1}^{r} X_i^\dagger X_i = \mathbb{1}_A$$

(d) If

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_E[V M_A V^\dagger]$$

is a Stinespring representation, then V *is an isometry.*

In (c) and (d), if the statement holds for one *particular Kraus/Stinespring representation, it holds for* all *Kraus/Stinespring representations.*

Proof. For the equivalence of (a) and (b) we use Lemma 4.15 to see that

$$\mathrm{tr}[\Phi_{A\to B}(M_A)] = \mathrm{tr}[\mathrm{tr}_A[(\mathbb{1}_B \otimes M_A^\mathsf{T})J(\Phi)]] = \mathrm{tr}[(\mathbb{1}_B \otimes M_A^\mathsf{T})J(\Phi)] = \mathrm{tr}[M_A^\mathsf{T}\,\mathrm{tr}_B[J(\Phi)]]$$

so $\Phi_{A\to B}$ is trace-preserving if and only if for all $M_A \in \mathrm{Lin}(A)$

$$\mathrm{tr}[M_A^\mathsf{T}] = \mathrm{tr}[M_A] = \mathrm{tr}[M_A^\mathsf{T}\,\mathrm{tr}_B[J(\Phi)]]$$

and by Lemma 4.20 this is equivalent with $\mathrm{tr}_B[J(\Phi)] = \mathbb{1}_A$

For the equivalence of (a) and (c) we observe that if we have a Kraus representation, then for $M_A \in \mathrm{Lin}(A)$

$$\begin{aligned}\mathrm{tr}[\Phi_{A\to B}(M_A)] &= \sum_{i=1}^{r} \mathrm{tr}[X_i M_A X_i^\dagger] \\ &= \sum_{i=1}^{r} \mathrm{tr}[X_i^\dagger X_i M_A] \\ &= \mathrm{tr}[\left(\sum_{i=1}^{r} X_i^\dagger X_i\right) M_A]\end{aligned}$$

so $\Phi_{A\to B}$ is trace-preserving if and only if we have

$$\operatorname{tr}[M_A] = \operatorname{tr}[\left(\sum_{i=1}^{r} X_i^\dagger X_i\right) M_A].$$

By Lemma 4.20 this holds for all $M_A \in \operatorname{Lin}(A)$ if and only if

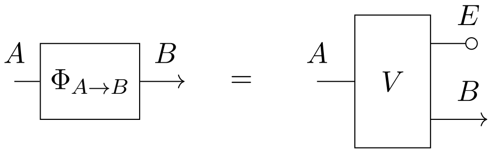
$$\sum_{i=1}^{r} X_i^\dagger X_i = \mathbb{1}_A.$$

For the equivalence of (a) and (d) we similarly find that if we have a Stinespring representation, $\Phi_{A\to B}$ is trace-preserving if and only if for all $M_A \in \operatorname{Lin}(A)$

$$\operatorname{tr}[M_A] = \operatorname{tr}[\operatorname{tr}_E[V M_A V^\dagger]] = \operatorname{tr}[V M_A V^\dagger] = \operatorname{tr}[V^\dagger V M_A].$$

By Lemma 4.20 this is the case if and only if $V^\dagger V = \mathbb{1}_A$ so if and only if V is an isometry. □

The Stinespring extension may be visualized as follows:



where the dot indicates we take a partial trace. We will comment a bit more on its meaning next lecture!

We summarize the concepts and notations we have introduced in this lecture.

Concept	Notation
Superoperator	$\Phi_{A\to B}, \Psi_{A\to B} \in \operatorname{Lin}(\operatorname{Lin}(A), \operatorname{Lin}(B))$
Classical channel	$Q : P(X) \to \P(Y)$ $p_Y(y) = \sum_x q(y\|x) p_X(x).$
Quantum channel	$\Phi_{A\to B}, \Psi_{A\to B} \in \mathrm{C}(A, B)$
Identity channel	$\mathcal{I}_A \in \mathrm{C}(A)$
Choi operator	$J(\Phi) \in \operatorname{Lin}(AB)$ for a superoperator $\Phi_{A\to B}$.
Kraus representation	$\Phi_{A\to B}(M_A) = \sum_i X_i M_A X_i^\dagger$ for $X_i \in \operatorname{Lin}(A, B)$ with $\sum_i X_i^\dagger X_i = \mathbb{1}_A$.
Stinespring representation	$\Phi_{A\to B}(M_A) = \operatorname{tr}_E[V M_A V^\dagger]$ for $V \in \operatorname{Isom}(A, BE)$.

Outlook

The exposition in this lecture is based on [45], which has an extensive discussion of various classes of quantum channels and their structure. The same material is covered in many textbooks, for example [31, 47].

Continuous time

Our ‘model’ of a quantum operation was that of an operation that happens one single step. This is in contrast to the Schrödinger equation which describes *continuous* time evolution. It is clear that there should be notions of continuous time noisy quantum channels, modelling the continuous interaction with a bath, or the continuous presence of some noise process. Consider a quantum channel Φ_A. We would like to see Φ_A as the result of applying some infinitesimal operation many times. As a first step, we could ask whether there exist *any* channels $\Psi_A^{(1)}$ and $\Psi_A^{(2)}$, which are both different applying a unitary conjugation, such that Φ_A is the composition

$$\Phi_A = \Psi_A^{(2)} \circ \Psi_A^{(1)}.$$

If such channels exist we call Φ_A *divisible*. There exist channels which are *not* divisible. We demand that the $\Psi_A^{(i)}$ are not a unitary conjugation, since we can clearly always write $\Phi_A(M_A) = U_A(U_A^\dagger \Phi_A(M_A) U_A) U_A^\dagger$ which is not an interesting decomposition. Next, we could also ask for the existence, for each $n \in \mathbb{N}$ of a channel $\Phi_A^{(n)}$ such that

$$\Phi_A = \underbrace{\Phi_A^{(n)} \circ \cdots \circ \Phi_A^{(n)}}_{n \text{ times}}.$$

This means that we can divide the channel into arbitrarily many ‘short time’ channels. In this case the channel is called *infinitely divisible*. Finally, there is the notion of a *Markovian* channel, which is a family of channels $\Phi_A^{(t)}$ for $t \geq 0$ which continuous in t and is such that for any s, t

$$\Phi_A^{(s+t)} = \Phi_A^{(s)} \circ \Phi_A^{(t)}$$

In this case, it turns out that $\rho_A(t) = \Phi_A^{(t)}(\rho_A)$ is the solution of a differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_A(t) = \mathcal{L}\rho_A(t)$$

where $\mathcal{L}$ is the generator of the dynamics, so formally $\rho(t) = e^{\mathcal{L}t}\rho_A$. The generator $\mathcal{L}$ is known as the *Lindbladian* and can be given a standard form

$$\mathcal{L}\rho = i[\rho, H] + \sum_j \left(L_j \rho L_j^\dagger - \{L_j^\dagger L_j, \rho\} \right)$$

where $[\cdot,\cdot]$ and $\{\cdot,\cdot\}$ are the commutator and anticommutator, H is a Hamiltonian and the L_j are arbitrary linear operators. The commutator with H corresponds to the usual Schrödinger equation and models the unitary part of the evolution, the operators L_j model noise processes. This approach is very useful for modelling physical systems which are undergoing continuous noise processes. The Markovian assumption means that the environment does not have a ‘memory’ of the quantum state on the physical system and is often reasonable if the environment is much larger and behaves thermally.

If you would like to learn more, see [48] for the notion of divisible channels. A textbook on open quantum systems with detailed derivations for Lindblad dynamics is [5].

Quantum channels and entanglement

The notion of a completely positive map is intimately related to how a superoperator $\Phi_{A\to B}$ tensored with the identity channel $\mathcal{I}_R$ acts on *entangled* states between A and R. Certain questions about quantum channels are hard to answer; this is particularly true for questions concerning the capacity to send information over such channels. It can be useful to restrict to classes of channels where one restricts what can happen to entanglement under the channel $\Phi_{A\to B} \otimes \mathcal{I}_R$. Three examples of relevant classes of quantum channels, for which analysis can be easier, are:

- *Entanglement-breaking channels:* a channel $\Phi_{A\to B}$ is called entanglement-breaking [23] if it 'breaks entanglement' between A and any reference system, in the sense that for any reference system R and state $\rho_{AR} \in \mathrm{S}(AR)$ the state

  $$\sigma_{BR} = (\Phi_{A\to B} \otimes \mathcal{I}_R)(\rho_{AR})$$

  is separable. This is equivalent to the Choi operator $J(\Phi)$ being separable.
- *PPT channels:* This is a similar notion, but now imposing the weaker condition that for any reference system R and state $\rho_{AR} \in \mathrm{S}(AR)$ the state

  $$\sigma_{BR} = (\Phi_{A\to B} \otimes \mathcal{I}_R)(\rho_{AR})$$

  is PPT, i.e. applying the partial transpose on B gives a positive operator ($\Gamma_B(\sigma_{BR}) \geq 0$ in the notation of Exercise 2.16). Equivalently, the channel $\Phi_{A\to B}$ is PPT if it remains completely positive when composing with the transpose on B, so the superoperator

  $$M_A \mapsto \Phi_{A\to B}(M_A)^{\mathsf{T}}$$

  is a quantum channel. An interesting conjecture is that if Φ_A is PPT, then $\Phi_A \circ \Phi_A$ is entanglement breaking [8].
- *Degradable channels:* a channel is degradable if it does not 'leak all information to the environment'. Given a Stinespring representation of $\Phi_{A\to B}$

  $$\Phi_{A\to B}(M_A) = \mathrm{tr}_E[V\rho_A V^\dagger]$$

  we can define the *complementary channel* which is the channel mapping to the environment

  $$\Phi^c_{A\to B}(M_A) = \mathrm{tr}_B[V\rho_A V^\dagger].$$

  The channel is degradable if one can obtain the complementary channel from the channel itself. This means that there is a channel $\Psi_{B\to E}$ such that

  $$\Phi^c_{A\to B} = \Psi_{B\to E} \circ \Phi_{A\to B}.$$

  A channel is *anti-degradable* if the complementary channel is degradable (so in this case one can recover the channel outcome from the complementary channel). For example, entanglement-breaking channels are degradable.

See [47] for more information and applications of these classes of channels.

4.3 Exercises

4.1 **Classical channels:** Show that classical channels are always CPTP.

4.2 **Positive maps on separable states:** Show that if $\Phi_{A\to B}$ is positive, then for any reference system R and any *separable* state $\rho \in \mathrm{S}(AR)$ we have

$$(\Phi_{A\to B} \otimes \mathcal{I}_R)(\rho_{AR}) \geq 0.$$

4.3 **The Choi isomorphism:** Prove Lemma 4.15.

4.4 **Convex combinations:** Show that the set of quantum channels between systems A and B is convex, that is, if $\Phi_{A\to B}, \Psi_{A\to B} \in \mathrm{C}(A,B)$ and $p \in [0,1]$, then the superoperator defined by

$$M_A \mapsto p\,\Phi_{A\to B}(M_A) + (1-p)\Psi_{A\to B}(M_A)$$

defines a quantum channel.

4.5 **Writing down channels:** For the following scenarios, write down the quantum channel that models it.

(a) You have a qubit system. With probability $\frac{1}{2}$ you do nothing; with probability $\frac{1}{2}$ you discard the state and prepare the state $|0\rangle$.
(b) You have two qubit systems, A and B. You replace the state on the B-system with $|0\rangle$ and apply a Pauli Z operator to the A-system.
(c) You start with system A, add a register E in state $|0\rangle$. You apply a global unitary U on A and E and then discard the system E.

4.6 **Extending to a linear map:** Show that given a map $\Phi_{A\to B} : \mathrm{S}(A) \to \mathrm{S}(B)$ satisfying Eq. (4.1) extends uniquely to a linear map $\Phi_{A\to B} : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$. *Hint: use Exercise 1.15.*

4.7 **Measurement as a quantum channel:**

(a) Given a measurement $\mu_A = \{\mu_{A,x}\}_{x\in\Omega_X}$, on a quantum system A with outcomes stored in a classical register X, the *measurement channel* is

$$\Phi^{\mu}_{A\to X}(M_A) = \sum_x \mathrm{tr}[\mu_{A,x} M_A]\, |x\rangle\langle x| \ .$$

Prove that this superoperator is a quantum channel.
(b) Suppose that $\Phi_{A\to X}$ is a quantum channel such that $\Phi_{A\to X}(\rho_A)$ is classical for any $\rho_A \in \mathrm{S}(A)$. Such a channel could be called a *quantum-to-classical* channel, as it maps a quantum system to a classical system. Show that there exists a measurement $\mu_A \in \mathrm{Meas}(A,X)$ such that $\Phi_{A\to X}$ corresponds to the measurement channel with measurement μ_A, so the set of quantum-to-classical channels to X coincides with the set of measurements with outcomes in X.

4.8 **Examples of quantum channels:** Let $p \in [0,1]$. Prove that the following superoperators are quantum channels.

(a) The *depolarising channel* $\mathcal{D}_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A)$, given by

$$M_A \mapsto (1-p)M_A + p\,\mathrm{tr}[M_A]\frac{1}{d_A}\mathbb{1} \ .$$

Compute the Choi matrix of $\mathcal{D}_p$.
(b) The *dephasing channel* $\mathcal{P}_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A)$ given by

$$M_A \mapsto (1-p)M_A + p\sum_{a\in\mathcal{A}} \langle a|M_A|a\rangle |a\rangle\langle a|.$$

where $\mathcal{A}$ is a basis (typically the standard basis) for A. Give a Kraus representation for $\mathcal{P}_p$.

(c) The *erasure channel* $\mathcal{E}_p : \mathrm{Lin}(A) \to \mathrm{Lin}(A')$ where $\mathcal{H}_{A'} = \mathcal{H}_A \oplus \mathrm{span}\{|\perp\rangle\}$, given by

$$M_A \mapsto (1-p)M_A + p\,\mathrm{tr}[M_A]|\perp\rangle\langle\perp| \ .$$

(d) The *replacement channel* $\mathcal{R}_\rho : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$, for $\rho \in S(B)$, given by

$$M_A \mapsto \mathrm{tr}[M_A]\rho \ .$$

Compute a Stinespring representation for $\mathcal{R}_\rho$. *Hint: if you prepare a purification of ρ on BR and discard both system A and the purifying system R, then this has the same effect as the replacement channel.*

4.9 **Different representations:** Compute the Choi operator and give Kraus and Stinespring representations of the following channels.

(a) The partial trace $\mathrm{tr}_B : \mathrm{Lin}(AB) \to \mathrm{Lin}(A)$.
(b) The depolarising channel from exercise 4.8.
(c) The replacement channel from exercise 4.8.
(d) The swap channel Swap : $\mathrm{Lin}(A \otimes A) \to \mathrm{Lin}(A \otimes A)$ given by $\rho \otimes \sigma \mapsto \sigma \otimes \rho$.

4.10 **Quantum channels on matrices:** Let $\rho \in S(\mathbb{C}^2)$ be a single qubit state, written in the computational basis as

$$\rho = \begin{pmatrix} a & b \\ b^* & c \end{pmatrix} \ .$$

(a) Compute the action of the depolarising channel $\mathcal{D}_p$ from Exercise 4.8 on this matrix.
(b) Compute the action of the dephasing channel $\mathcal{P}_p$ from Exercise 4.8 on this matrix.
(c) Show that for $p > 0$,

$$\mathcal{D}_p^n(\rho) \to \frac{1}{2}\mathbb{1} = \mathcal{D}_1(\rho) \ , \quad \mathcal{P}_p^n(\rho) \to \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix} = \mathcal{P}_1(\rho) \quad \text{as } n \to \infty.$$

4.11 **More examples of channels:**

(a) Consider the superoperator $\Phi_{A\to AB}$ defined by

$$M_A \mapsto M_A \otimes \rho_B$$

for some fixed $\rho_B \in \mathrm{S}(B)$. Show that this defines a quantum channel and compute its Choi matrix.

(b) Consider a measurement $\mu_{A,x}$ and a collection of states $\rho_{B,x} \in \mathrm{S}(B)$. The superoperator $\Phi_{A\to B}$ is defined by

$$\Phi_{A\to B}(M_A) = \sum_x \mathrm{tr}[\mu_{A,x} M_A]\rho_{B,x}.$$

Show that this defines a quantum channel and give a Kraus representation.

(c) Let $U_1, \ldots, U_r \in \mathrm{U}(A)$ and let p_i for $i = 1, \ldots, r$ be a probability distribution. Let Φ_A be the superoperator defined by

$$\Phi_A(M_A) = \sum_{i=1}^r p_i U_i M_A U_i^\dagger.$$

Show that this defines a quantum channel and give a Stinespring representation.

(d) For each of the above channels, give an operational interpretation.

4.12 **Constructions for completely positive maps:** This question provides some alternative constructions for the proof of Theorem 4.16.

(a) Suppose we are given $\Phi_{A\to B} : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$ in its Stinespring representation

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_E[VM_AV^\dagger] ,$$

for $V \in \mathrm{Lin}(A, BE)$. Use V to directly construct Kraus operators $X_i \in \mathrm{Lin}(A, B)$ for $i = 1, \ldots, r$ such that

$$\Phi_{A\to B} = \sum_{i=1}^{r} X_i M_A X_i^\dagger .$$

(b) Suppose instead we are given $\Phi_{A\to B}$ in its Kraus representation as above. $J(\Phi_{A\to B})$ in terms of the X_i, and show that it is positive.

(c) Now, assuming only that $J(\Phi_{A\to B}) \geq 0$, show that $\Phi_{A\to B}$ is completely positive. In other words, show that for any auxiliary system C and $M_{AC} \in \mathrm{Lin}(AC)$,

$$M_{AC} \geq 0 \Rightarrow \big[\Phi_{A\to B} \otimes \mathcal{I}_C\big](M_{AC}) \geq 0 .$$

4.13 **The Schrödinger equation:** The Schrödinger equation for the time-evolution of pure states is

$$\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle = -iH(t)|\psi(t)\rangle ,$$

where $H(t) \in \mathrm{Lin}(A)$ satisfying $H^\dagger(t) = H(t)$.

(a) Show that

$$|\psi(t)\rangle = U_t|\psi(0)\rangle ,$$

for some unitary $U_t \in \mathrm{Lin}(A)$.

(b) Write down the effect of time-evolution by t on a general quantum state $\rho \in S(A)$, and deduce that time-evolution is CPTP.

4.14 **Uniqueness of Kraus representations:** Let $\Phi_{A\to B} \in \mathrm{C}(A, B)$ be a channel with two different Kraus representations

$$\Phi_{A\to B}(\rho) = \sum_{i=1}^{r} X_i\rho X_i^\dagger = \sum_{j=1}^{s} Y_j\rho Y_j^\dagger ,$$

where $X_i, Y_j \in \mathrm{Lin}(A, B)$. Note that without loss of generality we may assume that $r = s$ (otherwise just add zero operators to the shorter representation).

(a) Let $\{|i_A\rangle\}$ be a basis for A, and define the following vectors in $A \otimes A$:

$$|x_i\rangle := \sum_j |j_A\rangle \otimes X_i|j_A\rangle ,$$

$$|y_i\rangle := \sum_j |j_A\rangle \otimes Y_i|j_A\rangle .$$

Show that

$$\sum_{i=1}^{r} |x_i\rangle\langle x_i| = \sum_{i=1}^{r} |y_i\rangle\langle y_i| := K ,$$

and that $K \geq 0$.

(b) Since K is positive it can be decomposed as

$$K = \sum_{k=1}^{r} \lambda_k |k_{AA}\rangle\langle k_{AA}| \ ,$$

for some $\lambda_k \geq 0$ and a basis $\{|k_{AA}\rangle\}$ for $A \otimes A$. Show that

$$|x_i\rangle = \sum_{k=1}^{r} v_{ik} \sqrt{\lambda_k} |k_{AA}\rangle \ , \quad |y_i\rangle = \sum_{k=1}^{r} w_{ik} \sqrt{\lambda_k} |k_{AA}\rangle \ ,$$

for some $r \times r$ unitary matrices (v_{ik}) and (w_{ik}).

(c) Deduce that the two Kraus representations are related by a unitary transformation. That is,

$$X_i = \sum_{j=1}^{r} u_{ij} Y_j \ ,$$

where u_{ij} are the elements of an $r \times r$ unitary matrix.

4.15 **Uniqueness of Stinespring representations:** Let $\Phi_{A\to B} \in \mathrm{C}(A, B)$ be a channel with two different Stinespring representations

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_E[V M_A V^\dagger] = \mathrm{tr}_F[W M_A W^\dagger] \ ,$$

for $V \in I(A, BE)$ and $W \in I(A, BF)$ isometries. By extending the smaller system, assume without loss of generality that $\dim E = \dim F$. Show that

$$V = (\mathbb{1} \otimes U) W \ ,$$

where U is a unitary matrix mapping from F to E. *Hint: use Exercise 4.14.*

Lecture 5

Protocols as quantum channels

Concept	Math translation
Quantum channels as physical processes.	The Stinespring representation: every quantum channel can be realized as unitary evolution, when also including a system modelling the environment.
Measurements are quantum channels.	Measurement $\mu_A = \{\mu_A(x) : x \in \mathcal{X}\}$ corresponds to channel $$M_A \mapsto \sum_x \operatorname{tr}[M_A \mu_A(x)] \, \lvert x \rangle\langle x \rvert.$$
Quantum protocols involving local quantum operations and classical communication (LOCC)	LOCC channels are compositions of one-way LOCC channels as in Definition 5.5 and Definition 5.6.
Superdense coding	A quantum channel, which consumes an entangled pair and uses one qubit of communication to send two classical bits.
Teleportation	A quantum channel, which consumes an entangled pair and uses two classical bits of communication to send a qubit.

In this lecture we will discuss three different aspects of quantum theory which are modelled by quantum channels. Firstly, quantum channels model physical processes. Unitary evolution (as given by the Schrödinger equation) models the evolution of a *closed* quantum system, that does not interact with an external environment. *Open* quantum systems are quantum systems which have interactions with a (non-controlled) environment. In general, if we have a *noisy* system, this is due to an interaction with an environment. In this lecture we will first discuss two consequences of our characterization theorem for quantum channels: the Stinespring extension shows that any quantum channel can be understood as extending to an environment system, on which we perform unitary dynamics.

A second application of quantum channels is that they incorporate *measurements*, as channels which map to classical systems. Our characterization of quantum channels will elucidate how general measurements can be realized by projective measurements.

Finally, quantum channels are the way we model general information processing *protocols*. We will define an important class of quantum protocols (Local Operations and Classical Communication, LOCC in short). After that we conclude by describing two important protocols, *superdense coding* and *teleportation*, dealing with the question of how to send classical bits using quantum bits and vice versa. This will be our first encounter with proper information theory in these lectures!

5.1 Quantum channels are physical

In the previous lecture we 'derived' quantum channels by imposing reasonable conditions: we demanded that channels sends states to states, and that this property is stable under taking tensor products. A question that may arise is that while these seem like *necessary* conditions it is perhaps not clear that they are also *sufficient*. Indeed, it could be that there is some other condition we missed so far, which could cause certain quantum channels to not correspond to a legitimate quantum dynamics. So, we could be worried that our Axiom 5 is too loose and that the set of possible dynamics between systems A and B is some strict subset of $\mathrm{C}(A, B)$. Fortunately, the fact that every channel has a Stinespring dilation shows there are no such additional conditions. To see this we slightly reformulate the Stinespring extension. Suppose that $\Phi_{A\to B}$ is a quantum channel, with Stinespring extension $V \in \mathrm{Isom}(A, BE)$ so

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_E[V M_A V^\dagger].$$

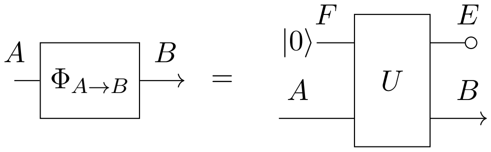
Then we may choose E' and F with $\mathcal{H}_E \subseteq \mathcal{H}_{E'}$ such that $\dim(AF) = \dim(BE')$ and we may further extend V to a unitary $U \in \mathrm{U}(AF, BE')$ in such a way that

$$U|\psi_A\rangle|0_F\rangle = V|\psi_A\rangle \tag{5.1}$$

for some arbitrary fixed state $|0\rangle$ on the extending system F and for all $\psi_A \in \mathcal{H}_A$. The fact that you can construct such an extension is Exercise 5.1. Given such a unitary extension, it follows directly from Eq. (5.1) that

$$\Phi_{A\to B}(M_A) = \mathrm{tr}_{E'}[U(M_A \otimes |0_F\rangle\langle 0_F|)U^\dagger]. \tag{5.2}$$

In diagrammatic notation: [MW: different category]



The interpretation of this is that any quantum channel can be realized with the following three steps:

(a) Prepare a fixed pure state $|0\rangle\langle 0|$ in an additional system F.

(b) Apply a unitary map to AF.

(c) Discard the subsystem E'.

These three operations are all physically reasonable operations, and therefore, *in principle* any quantum channel represent a physical process and Axiom 5 is no stronger than assuming that we can prepare pure states, apply unitaries and discard subsystems. In physics terminology, the above means that we can realize any quantum channel by coupling our system to an environment, time-evolve along a global Hamiltonian, and then restrict to the relevant subsystem.

Remark 5.1. Now that we have seen that quantum channels are in the above sense not more general than unitary evolution and the possibility to restrict to subsystems, you may wonder why we bothered to introduce quantum channels in the first place. Similarly, mixed states always have purifications, so *in principle* we could do all quantum (information) theory using only pure states and (projective) measurements. However, there are good reasons to use the formalism of mixed states and quantum channels. In many situations we do not have access (physically) to the purification of a state, and similarly, we do not know the actual interaction of a system with its environment, but we just know how the channel acts on our system of interest. For instance, there may be some really complicated interaction with the environment, but the effective result on our system is by good approximation a depolarizing channel. Besides this, in many cases we will want to do probabilistic operations on our states, and while we in theory could purify, this would be an artificial construction and obscure the interpretation of the process.

Quantum channels as noise models

We just saw that quantum channels can be modelled by an interaction with an environment. Such dynamics are also known as *open* or *noisy* dynamics. For instance, if one would like to build a quantum computer, ideally one can create a completely closed system, which performs exactly the desired unitary gates. However, in practice there will be some interaction with the environment which induces noise in the quantum computing device. One could model the noise in the Stinespring picture by explicitly taking the environment into account. It is often difficult to make accurate models for the (large and uncontrolled) environment, so it is often more useful to describe noise processes as quantum channels, for example using a Kraus representation.

A popular noise model is depolarizing noise, as defined in Example 4.13

$$\mathcal{D}_p(M_A) = (1-p)M_A + p\,\mathrm{tr}[M_A]\tau_A$$

which models that with probability $1-p$ no error happens and with probability $1-p$ the state gets replaced by a maximally mixed state. For a qubit, you can check that a Stinespring representation is given by $\{\sqrt{1-p}\mathbb{1}, \sqrt{\frac{p}{4}}X, \sqrt{\frac{4}{4}}Y, \sqrt{\frac{p}{4}}Z, \sqrt{\frac{p}{4}}\mathbb{1}\}$, so

$$\mathcal{D}_p(M_A) = (1-\frac{3p}{4})M_A + \frac{p}{4}\left(XM_AX + YM_AY + ZM_AZ\right). \tag{5.3}$$

This means that this *also* models the scenario where with probability p a random Pauli operator out of $\{\mathbb{1}, X, Y, Z\}$ is applied to the qubit!

Qubit channels

To gain some more feeling for quantum channels, we will now visualize some qubit channels as operations on the Bloch ball, parametrizing qubit states. To begin with we would like to know what happens to the Bloch ball when we apply a unitary, so we map

$$\rho \mapsto U\rho U^\dagger$$

It is a fact that every qubit unitary can be written as

$$U = e^{i\phi}e^{iH}$$

for a phase $\phi \in [0, 2\pi]$ and a Hermitian qubit operator H with $\mathrm{tr}[H] = 0$. Global phases do not matter, so we can ignore ϕ. Since the Pauli operators $\{X, Y, Z\}$ for a real basis for Hermitian traceless operators, we may write

$$H = \theta(xX + yY + zZ)$$

for $\vec{r} = (x, y, z)$ on the Bloch sphere. In Exercise 5.5 you will show that

$$\begin{aligned} U = U(\vec{r}, \theta) &= \exp(i\theta(xX + yY + zZ)) \\ &= \cos(\theta)\mathbb{1} + i\sin(\theta)(xX + yY + zZ) \end{aligned}$$

and that this corresponds to a *rotation of the Bloch ball* around the axis given by $\vec{r}$, with angle 2θ. In conclusion, there is a one-to-one correspondence between unitaries on a qubit and rotations of the Bloch sphere.

For another example we take the *depolarizing channel*

$$\mathcal{D}_p(M_A) = (1-p)M_A + p\,\mathrm{tr}[M_A]\frac{\mathbb{1}}{2}.$$

Given $\rho(\vec{r})$ it gets mapped as

$$\frac{1}{2}\begin{pmatrix} 1+z & x-iy \\ x+iy & 1-z \end{pmatrix} \mapsto \frac{1}{2}\begin{pmatrix} 1+(1-p)z & (1-p)x - i(1-p)y \\ (1-p)x + i(1-p)y & 1-(1-p)z \end{pmatrix},$$

so it gets mapped to the state which has rescaled Bloch vector $(1-p)\vec{r}$. In other words, the depolarizing channel $\mathcal{D}_p$ shrinks the Bloch sphere by a factor $1-p$. Note that the fixed point $\vec{r} = 0$ of this operation corresponds to the maximally mixed state.

As a final example we take the *dephasing channel*

$$\mathcal{P}_p(M_A) = (1-p)M_A + p(\langle 0|M_A|0\rangle|0\rangle\langle 0| + \langle 1|M_A|1\rangle|1\rangle\langle 1|). \tag{5.4}$$

Given $\rho(\vec{r})$ you can check this gets mapped as

$$\frac{1}{2}\begin{pmatrix} 1+z & x-iy \\ x+iy & 1-z \end{pmatrix} \mapsto \frac{1}{2}\begin{pmatrix} 1+z & (1-p)x - i(1-p)y \\ (1-p)x + i(1-p)y & 1-z \end{pmatrix}.$$

We see that this corresponds to mapping $\vec{r} = (x, y, z)$ to $((1-p)x, (1-p)y, z)$. Visually this corresponds to shrinking the Bloch ball, but only in the x, y direction (so the result is an ellipsoid along the z-axis).

5.2 Measurements as quantum channels

Given a measurement $\mu = \{\mu_A(x)\}_{x\in\mathcal{X}}$ on a quantum system A with outcomes in a classical register X, we can model this measurement as the channel $\Phi^M_{A\to X}$

$$\Phi^{\mu}_{A\to X}(M_A) = \sum_x \mathrm{tr}[\mu_A(x)M_A]\,|x\rangle\langle x|. \tag{5.5}$$

In Exercise 4.7 you have shown that this indeed defines a quantum channel, and that any quantum-to-classical channel corresponds to a measurement. We will now use the quantum channel framework to address two aspects of quantum measurements that were so far perhaps not entirely satisfactory.

How to implement a measurement?

We introduced the set of measurements in Lecture 1, but the 'natural' set of measurements one finds in the pure state formulation of quantum mechanics are projective measurements. However, similar to the above discussion on the Stinespring dilation, we will see that any measurement can be constructed using only a projective measurement. This fact is known as *Naimark's theorem.*

Theorem 5.2 (Naimark)**.** *Suppose* $\mu_A \in \mathrm{Meas}(A, X)$*. Then there exists an auxiliary system* F *and a* projective *measurement* $\nu_{AF} \in \mathrm{Meas}(AF, X)$ *such that*

$$\Phi^{\mu}_{A\to X}(\rho_A) = \Phi^{\nu}_{AF\to X}(\rho_A \otimes |0_F\rangle\langle 0_F|)$$

for some state $|0_F\rangle\langle 0_F| \in \mathrm{S}(F)$*.*

Proof. Consider a unitary extension of the measurement channel as in Eq. (5.2)

$$\Phi^{\mu}_{A\to X}(\rho_A) = \mathrm{tr}_E[U(\rho_A \otimes |0_F\rangle\langle 0_F|)U^\dagger].$$

for unitary U and some pure state $|0_F\rangle$ on F. Then the probability of outcome x is given by

$$\begin{aligned} p_x(\rho_A) &= \langle x|\Phi^{\mu}_{A\to X}(\rho_A)|x\rangle = \langle x|\,\mathrm{tr}_E[U(\rho_A \otimes |0_F\rangle\langle 0_F|)U^\dagger]|x\rangle \\ &= \mathrm{tr}[(\langle x| \otimes \mathbb{1}_E)U(\rho_A \otimes |0_F\rangle\langle 0_F|)U^\dagger(|x\rangle \otimes \mathbb{1}_E)] \\ &= \mathrm{tr}[(U^\dagger(|x\rangle\langle x| \otimes \mathbb{1}_E)U)(\rho_A \otimes |0_F\rangle\langle 0_F|)]. \end{aligned}$$

Therefore, if we define

$$P_x = U^\dagger(|x\rangle\langle x| \otimes \mathbb{1}_E)U \in \mathrm{Lin}(AF)$$

these are projection operators and

$$\begin{aligned} \sum_x P_x &= \sum_x U^\dagger(|x\rangle\langle x| \otimes \mathbb{1}_E)U \\ &= U^\dagger\underbrace{(\sum_x |x\rangle\langle x| \otimes \mathbb{1}_E)}_{=\mathbb{1}_X\otimes\mathbb{1}_E}U = U^\dagger U = \mathbb{1}_{AF} \end{aligned}$$

and the P_x define the desired projective measurement $\nu_{AF}(x) = P_x$. □

What happens to a quantum state after measurement?

So far we modelled measurements in a 'destructive' way, in the sense that we said that after measurement we just have the classical outcome and no longer have a quantum state. This corresponds to the channel in Exercise 4.7. By our channel formulation it is now clear what happens to a bipartite state $\rho_{AB} \in \mathrm{S}(AB)$ if we only measure the A system (so we do a *partial measurement*), which will be given by

$$(\Phi^{\mu}_{A\to X} \otimes \mathcal{I}_B)(\rho_{AB}) = \sum_x |x\rangle\langle x| \otimes \mathrm{tr}_A[(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}]$$

In particular, if we *see* outcome x in our classical register, the state on B must be

$$\frac{\mathrm{tr}_A[(\mu_A(x) \otimes \mathbb{1}_B)\rho_{AB}]}{\mathrm{tr}[\mu_A(x)\rho_A]}.$$

More generally, we can look at quantum channels which output classical information X but also keep some quantum system B. Such channels must be of the form (see Exercise 5.4)

$$\Phi_{A\to BX}(M_A) = \sum_x \Theta_{A\to B,x}(M_A) \otimes |x\rangle\langle x|$$

where each $\Theta_{A\to B,x} \in \mathrm{CP}(A,B)$ but need not be trace-preserving. However, for $\Phi_{A\to BX}$ to be a quantum channel, we do require that

$$\sum_x \Theta_{A\to B,x} \in \mathrm{C}(A,B)$$

(by assumption the sum is completely positive, so the nontrivial condition is that the sum is trace-preserving). From this we see that a channel which outputs some classical information x is captured by the following definition:

Definition 5.3 (Instrument)**.** A collection of maps $\{\Theta_{A\to B,x}\} \subset \mathrm{CP}(A,B)$ is called an *instrument* if

$$\Phi_{A\to B} = \sum_x \Theta_{A\to B,x}$$

is a quantum channel (i.e. it is trace preserving).

If our initial state is $\rho_A \in \mathrm{S}(A)$ we now find classical outcome x with probability

$$p_x = \mathrm{tr}[\Theta_{A\to B,x}(\rho_A)]$$

and if we observe x, then the *post-measurement state* on the register B is given by

$$\sigma_{B,x} = \frac{\Theta_{A\to B,x}(\rho_A)}{\mathrm{tr}[\Theta_{A\to B,x}(\rho_A)]}.$$

Given a measurement $\mu_A \in \mathrm{Meas}(A,X)$ we can define an instrument by

$$\Theta_{A,x}(\rho_A) = \sqrt{\mu_A(x)}\rho_A\sqrt{\mu_A(x)}$$

which gives the post-measurement state, given outcome x,

$$\rho_{A,x} = \frac{\sqrt{\mu_A(x)}\rho_A\sqrt{\mu_A(x)}}{\mathrm{tr}[\mu_A(x)\rho_A]}.$$

In the special case of a basis measurement $\mu_x = |\psi_x\rangle\langle\psi_x|$, the post-measurement state is just given by $|\psi_x\rangle\langle\psi_x|$.

Finally, we comment on an important difference between quantum and classical information. The fact that measurements have some 'destructive property' is very central to quantum information theory. The most basic incarnation of this phenomenon is the *no-cloning* theorem. Suppose that you had a 'cloning device' which took as input a quantum state and returned you two copies of the state. This would allow non-destructive measurements: one would simply clone the state ρ, and measure only one copy. However, such cloning devices do not exist, not even if we restrict to pure states:

Theorem 5.4 (No cloning)**.** *For any system A with $|A| \geq 2$ there is no quantum channel cloning (pure) states, i.e. there does not exist a channel $\Phi_{A\to AA}$ with the property that for all pure $\rho_A \in \mathrm{S}(A)$*

$$\Phi_{A\to AA}(\rho_A) = \rho_A^{\otimes 2}.$$

The reason is that such an operation would violate linearity (so there is not even a cloning superoperator). You will prove this in Exercise 5.3.

5.2.1 Local operations and classical communication (LOCC)

In the previous section we discussed the precise mathematical formulation of measurements in terms of quantum channels. A measurement is any situation where we extract *classical information* from a quantum system. An important scenario in information theory is the following: we have two parties, Alice and Bob. They are able to locally manipulate their quantum systems at will, but they can only communicate classically. For instance, we may imagine a situation where Alice and Bob each have their own lab, and they are spatially separated. They can not send over quantum bits, but they can call each other on the telephone to tell each other what measurement outcomes they found in their experiments.

Such a set-up is formalized by the notion of *Local Operations and Classical Communication*, LOCC in short. The idea is that we allow Alice to perform any operation on her system, and then communicate classical information to Bob. Then Bob is allowed to do any operation on his system and send classical information to Alice, and Alice and Bob may do as many rounds of this as they would like. The precise definition is a bit complicated and unwieldy, as we will see below. However, we will rarely use this formal definition, and the *concept* of LOCC, as expressed in the above scenario should be clear.

For the formal definition, let us first define what a single round of LOCC is.

Definition 5.5 (One-way LOCC)**.** Suppose Alice and Bob have quantum systems A and B, and they apply some quantum channel $\Phi_{AB\to A'B'}$ after which Alice has system A' and Bob B'. We say that $\Phi_{AB\to A'B'}$ is:

(a) one-way LOCC from Alice to Bob if there exists an instrument $\{\Theta_{A\to A',x}\}_x$ and a channel $\Psi_{B\to B',x}$ for each x such that

$$\Phi_{AB\to A'B'} = \sum_x \Theta_{A\to A',x} \otimes \Psi_{B\to B',x}$$

(b) one-way LOCC from Bob to Alice if there exists an instrument $\{\Theta_{B\to B',x}\}_x$ and a channel $\Psi_{A\to A',x}$ for each x such that

$$\Phi_{AB\to A'B'} = \sum_x \Psi_{A\to A',x} \otimes \Theta_{B\to B',x}.$$

In this definition, for the one-way LOCC from Alice to Bob, x represents the classical data Alice obtains from her instrument. This classical information may be sent from Alice to Bob and Bob performs the channel $\Psi_{B\to B',x}$ depending on the value of x. A one-way LOCC channel from Bob to Alice has a similar interpretation with the roles of Alice and Bob reversed.

Definition 5.6 (LOCC)**.** Suppose Alice and Bob have quantum systems A and B, and they apply some quantum channel $\Phi_{AB\to A'B'}$ after which Alice has system A' and Bob B'. The channel $\Phi_{AB\to A'B'}$ is called LOCC if it can be written as a composition of channels which are either one-way LOCC from Alice to Bob or one-way LOCC from Bob to Alice.

Again, this is a rather complicated definition. Besides being a long definition, it is also mathematically difficult to work with. This is mainly because of the unbounded number of rounds in the definition. This means it may be challenging to check if a given channel is LOCC or not, or perform mathematical analysis in the class of LOCC channels. The following is a more general notion, which has a less clear operational interpretation but is often easier to work with.

Definition 5.7 (Separable channels)**.** A quantum channel $\Phi_{AB\to A'B'} \in \mathrm{C}(AB, A'B')$ is called *separable* (between AA' and BB') if there exist collections of completely positive maps $\Psi_{A\to A',\omega} \in \mathrm{CP}(A, A')$ and $\Theta_{B\to B',\omega} \in \mathrm{CP}(B, B')$ for $\omega \in \Omega$ such that

$$\Phi_{AB\to A'B'} = \sum_{\omega\in\Omega} \Psi_{A\to A',\omega} \otimes \Theta_{B\to B',\omega}.$$

Note that the definition of separability involves a choice of both the input and output systems of the channel into two subsystems. We have the following:

Lemma 5.8. *A composition of separable channels is separable. Every LOCC channel is separable.*

The proof is Exercise 5.9. See Exercise 5.10 and Exercise 5.11 for a characterization of separable channels. A state is separable if and only if it can be prepared by an LOCC channel, as you can show in Exercise 5.11.

5.3 Superdense coding and teleportation

We will address two basic questions in information theory:

(a) If I can send over a quantum bit, can I also send classical information, and if so, how much?

(b) If I can send over classical bits, can I also send over quantum information, and if so, how much?

In this lecture we will not give a complete answer to this question, but we will show two particular and famous protocols. These protocols assume that Alice and Bob do not only exchange (qu)bits, but that they also share a maximally entangled state!

To analyze these protocols, we introduce the *Bell basis* of $\mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^2 \otimes \mathbb{C}^2$, consisting of

$$|\Phi^{(00)}_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) \qquad |\Phi^{(01)}_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$$
$$|\Phi^{(10)}_{AB}\rangle = \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle) \qquad |\Phi^{(11)}_{AB}\rangle = \frac{1}{\sqrt{2}}(|10\rangle - |01\rangle)$$

This basis is such that $|\Phi^{(00)}_{AB}\rangle = |\Phi^+_{AB}\rangle$ is the usual maximally entangled state, and

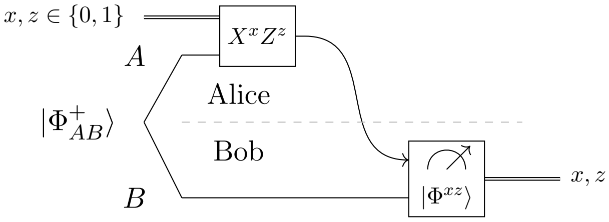
$$|\Phi^{(xz)}_{AB}\rangle = (X^x Z^z \otimes \mathbb{1}_B)|\Phi^+_{AB}\rangle. \tag{5.6}$$

Note that by the transpose trick in Lemma 2.18

$$|\Phi^{(xz)}_{AB}\rangle = (\mathbb{1}_A \otimes (X^x Z^z)^{\mathsf{T}})|\Phi^+_{AB}\rangle = (\mathbb{1}_A \otimes Z^z X^x)|\Phi^+_{AB}\rangle. \tag{5.7}$$

Superdense coding

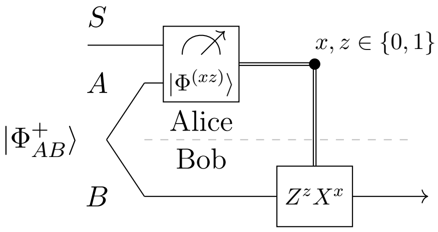
We start with *superdense coding*, which is a protocol to send over two bits of classical information, sending over only a single qubit and using a shared maximally entangled qubit between Alice and Bob. The idea is simple: if Alice wants to send over bits x and z, she applies $X^x Z^z$ to her system. If she then sends over her subsystem, Bob will possess the state $|\Phi^{(xz)}_{AB}\rangle$, and he can just measure in the Bell basis to find x and z.



Teleportation

There is a dual protocol, which also requires a shared maximally entangled state between Alice and Bob, and which allows them to transfer a single qubit by sending over two classical bits.

Let us first describe the procedure in words. Alice considers her part of the maximally entangled state and the qubit system S she wants to send over. She measures in the Bell basis, obtaining an outcome xz. She sends over this outcome to Bob (at a cost of two classical bits of communication). Bob then applies the unitary $Z^z X^x$ to his system. This is described in the following diagram:



This is an example of an LOCC protocol! Let us now write out the teleportation protocol as a quantum channel and check that it actually performs as promised and the qubit comes out on Bob's side! Let $M_S \in \mathrm{Lin}(S)$, then the protocol consists of the following steps:

(a) Prepare a maximally entangled state on AB

$$M_S \mapsto M_S \otimes |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|.$$

(b) Measure in the Bell basis

$$M_S \otimes |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}| \mapsto \sum_{x,z\in\{0,1\}} \left(\langle\Phi^{(xz)}_{SA}| \otimes \mathbb{1}_B (M_S \otimes |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|)|\Phi^{(xz)}_{SA}\rangle \otimes \mathbb{1}_B\right) \otimes |zx\rangle\langle zx|$$

(c) In the next step we send over the classical system with the information about x and z to Bob, and Bob applies $Z^z X^x$, which yields

$$M_S \mapsto \sum_{x,z\in\{0,1\}} Z^z X^x \left(\langle\Phi^{(xz)}_{SA}| \otimes \mathbb{1}_B (M_S \otimes |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|)|\Phi^{(xz)}_{SA}\rangle \otimes \mathbb{1}_B\right) X^x Z^z \tag{5.8}$$

using $(Z^z X^x)^\dagger = X^x Z^z$.

We conclude that after the teleportation protocol, we have implemented the channel defined by Eq. (5.8) which we will denote by $\Phi_{S\to B}$. Now, S and B are both qubits and we claim that this channel is really the identity channel from S to B, so $\Phi_{S\to B} = \mathcal{I}_{S\to B}$. This would confirm that we have indeed sent over the qubit using this protocol! To see this, we rewrite

$$\Phi_{S\to B}(M_S) = \sum_{x,z\in\{0,1\}} Z^z X^x \left(\langle \Phi^+_{SA}| \otimes \mathbb{1}_B (X^x Z^z M_S Z^z X^x \otimes |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|) |\Phi^+_{SA}\rangle \otimes \mathbb{1}_B \right) X^x Z^z$$

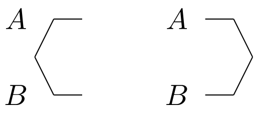
using Eq. (5.7). Now we may verify that

$$\begin{aligned}(\mathbb{1}_S \otimes \langle \Phi^+_{AB}|)(|\Phi^+_{SA}\rangle \otimes \mathbb{1}_B) &= \frac{1}{2} \sum_{i,j\in\{0,1\}} (\mathbb{1}_S \otimes \langle i_A| \otimes \langle i_B|)(|j_S\rangle \otimes |j_A\rangle \otimes \mathbb{1}_B) \\ &= \frac{1}{2} \sum_{i,j\in\{0,1\}} |j_S\rangle \otimes \langle i_A | j_A\rangle \otimes \langle i_B| \\ &= \frac{1}{2} \sum_{i\in\{0,1\}} |i_S\rangle\langle i_B|\end{aligned} \tag{5.9}$$

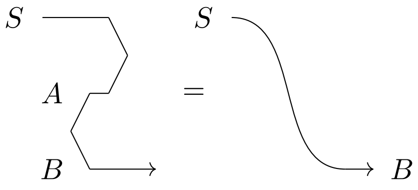
and hence

$$\begin{aligned}\Phi_{S\to B}(M_S) &= \frac{1}{4} \sum_{x,z} \sum_{i,j} Z^z X^x |i_B\rangle\langle i_S| (X^x Z^z) M_S (Z^z X^x) |j_S\rangle\langle j_B| X^x Z^z \\ &= \frac{1}{4} \sum_{x,z} (Z^z X^x)(X^x Z^z) M_S (Z^z X^x)(X^x Z^z) = M_S.\end{aligned}$$

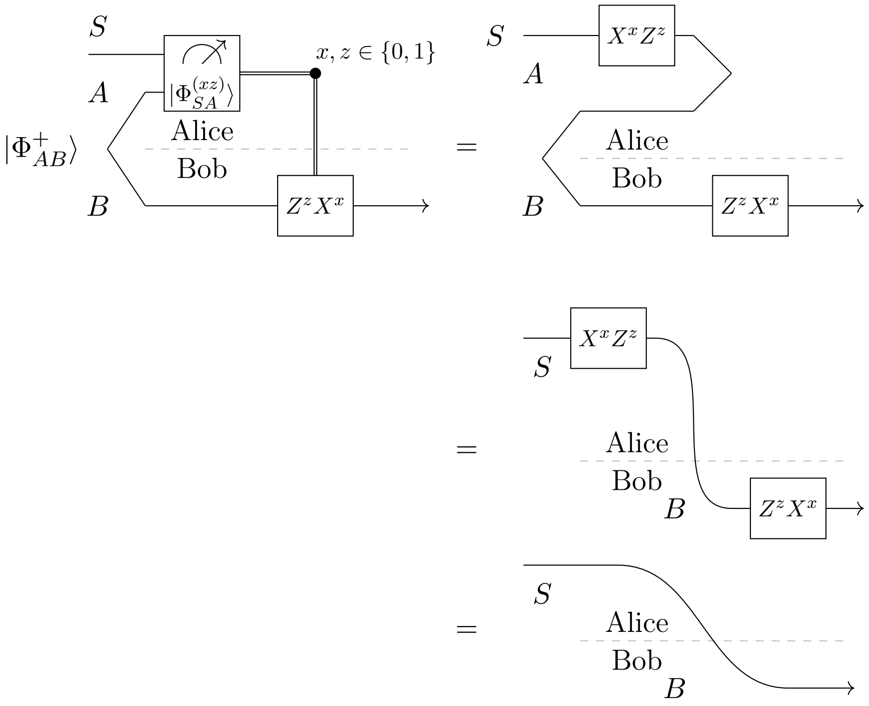
This is what we wanted to show, and the teleportation protocol indeed transmits the qubit from Alice's side to Bob! To give a little more intuition for how this works, we note that if we visually represent the operations of preparing a maximally entangled state $|\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|$ and projecting onto this state (i.e. $M_{AB} \mapsto \langle \Phi^+_{AB}|M_{AB}|\Phi^+_{AB}\rangle$) as respectively



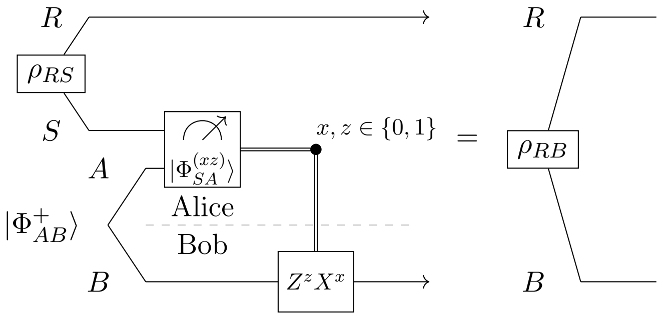
then the equality in Eq. (5.9), ignoring the normalization, can be visualized as



This gives the following visual 'proof' of $\Phi_{S\to B}(M_A) = \mathcal{I}_{S\to B}$



which is valid for each outcome x, z. Since we have shown that teleportation really implements the *channel* mapping the system S to B, if we have an additional reference system R, it must also hold that if we start with a state $\rho_{SR} \in \mathrm{S}(SR)$, the teleportation protocol preserves the correlations with the R system. For instance, if S is maximally entangled with R, then after teleportation B will share a maximally entangled state with R:



showing that $\mathcal{I}_R \otimes \Phi_{S\to B}(\rho_{RS}) = \rho_{RB}$.

Outlook

For an elaborate discussion of LOCC and separable channels we refer to [45] and [6].

5.4 Exercises

5.1 **Unitary extension:** Confirm that given a Stinespring extension V it is possible to find a unitary as in Eq. (5.1).

5.2 **Projective measurements in Bell games:** Explain why in Lecture 3 we could restrict to quantum strategies using only projective measurements without loss of generality.

5.3 **No cloning:** In this exercise you will prove Theorem 5.4.

(a) Suppose A and A' are qubit systems. Show that there does not exist a quantum channel $\Phi_{A\to AA'}$ which is such that

$$\Phi_{A\to AA'}(\rho_A) = \rho_A^{\otimes 2} \tag{5.10}$$

for all *classical* $\rho_A = p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1|$ for $p \in [0,1]$.

(b) Show that there is no channel $\Phi_{A\to AA'}$ such that Eq. (5.10) holds for all *pure states*. *Hint: look at two different bases.*

(c) Prove Theorem 5.4 (note that this concerns any system A with $|A| \geq 2$).

(d) What are the qubit states which are both pure and classical? Show that there *does* exist a quantum channel which clones all qubit states which are both pure and classical.

5.4 **Instruments:** Suppose that $\Phi_{A\to BX}$ is a quantum channel such that for any input ρ_A the resulting state on X is classical, so

$$\Phi_{A\to BX}(\rho_A) = \sum_x p(x)\rho_{B,x} \otimes |x\rangle\langle x|$$

for some probability distribution p and states $\rho_{B,x} \in \mathrm{S}(B)$ depending on ρ_A. Show that there exists an instrument $\{\Theta_{A\to B,x} \in \mathrm{CP}(A,B) : x \in \Omega_X\}$ such that

$$\Phi_{A\to BX} = \sum_x \Theta_{A\to B,x} \otimes |x\rangle\langle x|.$$

5.5 **Unitaries on the Bloch sphere:** This exercise concerns unitaries on a single qubit.

(a) Show that for $\vec{r} = (x,y,z)$ on the Bloch sphere

$$H(\vec{r}) = xX + yY + zZ$$

satisfies $H(\vec{r})^2 = \mathbb{1}$ and $\mathrm{tr}[H(\vec{r})] = 0$.

(b) Suppose that H is a hermitian qubit operator such that $H^2 = \mathbb{1}$ and $\mathrm{tr}[H] = 0$. Show that

$$e^{i\theta H} = \cos(\theta)\mathbb{1} + i\sin(\theta)H.$$

Hint: use the spectral decomposition. The conditions on H completely determine the values of the two eigenvalues.

(c) For the special case $H = X$ (so $\vec{r} = (1,0,0)$), check that

$$U(\theta) = e^{i\theta X}$$

acts as rotation by an angle 2θ around the x-axis. That is, show that if $\rho(\vec{s})$ is a state with Bloch vector $\vec{s} = (x', y', z')$ it gets transformed as

$$\rho \mapsto U(\theta)\rho U(\theta)^\dagger = \rho(\vec{t})$$

where

$$\vec{t} = (x', \cos(2\theta)y' - \sin(2\theta)z', \sin(2\theta)y' + \cos(2\theta)z').$$

5.6 **Qubit quantum channels:**

(a) Show that the qubit depolarizing channel can be written as in Eq. (5.3).

(b) Show that the qubit dephasing channel in Eq. (5.4) has Kraus operators $\{\sqrt{1-\frac{p}{2}}\mathbb{1}, \sqrt{\frac{p}{2}}Z\}$.

5.7 **Kraus decompositions for instruments:** Given a POVM $\{\mu_i\}_{i=0}^{m-1}$ on a system A, we can construct an instrument of the form

$$\Lambda_{A\to B}(M_A) = \sum_{i=0}^{m-1} |i_B\rangle\langle i_B| \operatorname{tr}[\mu_i M_A] ,$$

where $\{|i_B\rangle\}_{i=0}^{m-1}$ form a basis for B.

Define a POVM on $\mathbb{C}^2$ with elements

$$\mu_0 = \frac{2}{3}|+\rangle\langle +| , \quad \mu_1 = \frac{2}{3}|\psi_+\rangle\langle\psi_+| , \quad \mu_2 = \frac{2}{3}|\psi_-\rangle\langle\psi_-| ,$$

where

$$|\psi_\pm\rangle = \frac{1}{\sqrt{2}}\Big(|0\rangle + e^{\pm\frac{2\pi i}{3}}|1\rangle\Big) ,$$

and let $\Lambda_{A\to B}$ be the above map. Here $\dim(A) = 2$, $\dim(B) = 3$.

(a) Verify that $\{\mu_i\}_{i=0}^{2}$ defines a POVM.

(b) Write down a set $\{X_i\}_{i=0}^{2}$ of Kraus operators for this instrument, and the corresponding Stinespring dilation isometry $V \in \mathrm{Lin}(A, BF)$.

(c) Let $|\alpha_i\rangle \in A$ be such that $\mu_i = |\alpha_i\rangle\langle\alpha_i|$. Show that $W = \sum_{i=0}^{2}|i_B\rangle\langle\alpha_i|$ defines an isometry in $\mathrm{Lin}(A, B)$.

(d) Let $Q = WW^\dagger \in \mathrm{Lin}(B)$. Show that the vector

$$|z\rangle = \frac{1}{\sqrt{3}}\big(|0\rangle + e^{\frac{2\pi i}{3}}|1\rangle + e^{-\frac{2\pi i}{3}}|2\rangle\big)$$

is in the kernel of Q.

(e) Define

$$|\beta_i\rangle = |\alpha_i\rangle \otimes |0_E\rangle + \langle z|i_B\rangle|0\rangle \otimes |1_E\rangle , \quad \text{for } i = 0, 1, 2.$$

Show that these vectors form an orthonormal set in $A \otimes E$, for E an auxiliary system with $\dim E = 2$. Find a normalised vector $|\beta_3\rangle$ such that $\{|\beta_i\rangle\}_{i=0}^{3}$ is a basis for $A \otimes E$.

(f) Show that the projectors $P_i = |\beta_i\rangle\langle\beta_i|$ for $i = 0, 1, 2$ form a Naimark extension of our POVM. In other words, show that

$$\operatorname{tr}[\mu_i M_A] = \operatorname{tr}[P_i M_A \otimes |0\rangle\langle 0|] , \quad \text{for all } M_A \in \mathrm{Lin}(A).$$

5.8 **Entanglement swapping:** Suppose that Alice and Bob both share a maximally entangled qubit state with a third party Charlie. Describe a procedure by which Alice and Bob can generate a maximally entangled qubit pair between them using LOCC operations.

Remark: This trick is used in practice to generate entanglement over long distances from shorter distance entanglement!

5.9 **Separable channels:** Prove Lemma 5.8.

5.10 **Characterization of separable channels:** Let $\Phi_{AB\to A'B'} \in \mathrm{C}(AB, A'B')$. Show that the following are equivalent:

(a) The channel $\Phi_{AB\to A'B'}$ is separable.
(b) The Choi matrix $J(\Phi) \in \mathrm{PSD}(ABA'B')$ is separable under the bipartitioning AA' versus BB'.
(c) There exists a Kraus represention with tensor product Kraus operators, that is, there exist operators $X_i \in \mathrm{Lin}(A, A')$ and $Y_i \in \mathrm{Lin}(B, B')$ such that

$$\Phi_{AB\to A'B'}(M_{AB}) = \sum_i (X_i \otimes Y_i) M_{AB} (X_i^\dagger \otimes Y_i^\dagger).$$

Hint: prove (a) ⇒ (b) ⇒ (c) ⇒ (a), following the argument in the proof of Theorem 4.16.

5.11 **Separable channels and states:**

(a) Show that a separable channel maps separable states to separable states.
(b) Show that if a channel maps all separable states to separable states, it is separable. *Hint: use the characterization in terms of the Choi matrix.*
(c) Show that every separable state can be prepared (from a trivial state) by an LOCC channel.

5.12 **Remote state preparation:** This question concerns a protocol known as *remote state preparation*, which is closely related to quantum teleportation, but here only one bit of classical communication is required to remotely prepare a given qubit state. In contrast to teleportation, the sender knows a classical description of the state to prepare, and has access to a larger number of entangled qubits.

(a) Let $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \in \mathbb{C}^2$ be a pure qubit state. Show that

$$\big(|\psi\rangle\langle\psi|\big)^T = |\bar{\psi}\rangle\langle\bar{\psi}| \ , \quad \text{and} \quad \mathbb{1} - |\bar{\psi}\rangle\langle\bar{\psi}| = |\bar{\psi}^\perp\rangle\langle\bar{\psi}^\perp|$$

where the transpose T is taken with respect to the computational basis, and

$$|\bar{\psi}\rangle = \bar{\alpha}|0\rangle + \bar{\beta}|1\rangle \ , \quad |\bar{\psi}^\perp\rangle = \bar{\beta}|0\rangle - \bar{\alpha}|1\rangle \ .$$

(b) Suppose Alice and Bob share a maximally entangled state $|\Phi^+_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Alice would like to give Bob the state $|\psi\rangle$ as a gift, but to keep it a surprise she doesn't want to give Bob any information about α or β.
Alice performs a projective measurement on her half of the maximally entangled state, corresponding to projectors $\{\Pi_0, \Pi_1\}$ where $\Pi_0 = |\bar{\psi}\rangle\langle\bar{\psi}|$, and $\Pi_1 = \mathbb{1} - |\bar{\psi}\rangle\langle\bar{\psi}|$. Show that the outcome probabilities for this measurement are

$$p_x(|\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|) = \frac{1}{2} \ , \quad x = 0, 1 \ .$$

(c) Alice then sends Bob the single-bit outcome of her measurement x in a classical system C. Show that Bob now holds the state

$$\rho_{BC} = \frac{1}{2}|\psi_B\rangle\langle\psi_B| \otimes |0_C\rangle\langle 0_C| + \frac{1}{2}|\psi^\perp_B\rangle\langle\psi^\perp_B| \otimes |1_C\rangle\langle 1_C| \ ,$$

where $|\psi^\perp\rangle = \beta|0\rangle - \alpha|1\rangle$.
(d) Assume, for the moment, that $|\psi\rangle$ is of the form $|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{i\theta}|1\rangle)$. Show that

$$Z|\psi^\perp\rangle\langle\psi^\perp|Z = |\psi\rangle\langle\psi| \ ,$$

and hence describe how Bob can recover the state $|\psi\rangle$ from ρ_{BC}.

(e) Now suppose that Alice wants to send Bob *many* nice quantum gifts $|\psi_1\rangle, |\psi_2\rangle, \ldots, |\psi_n\rangle \in \mathbb{C}^2$, which do not necessarily take the above form. Assume that Alice and Bob share $n \cdot m$ maximally entangled states, where $m = 2^{n+\log n}$. They each arrange their qubits in a rectangle, so that the qubit from the (i, j)th maximally entangled pair lies in the ith row and jth column (for $i = 1, \ldots, n$, $j = 1, \ldots, m$). For each $i = 1, \ldots, n$, Alice measures the entire ith row of qubits in the $\{|\bar{\psi_i}\rangle, |\bar{\psi_i^\perp}\rangle\}$ basis. Show that, with high probability, there will be an entire *column* of qubits for which the measurements were successful (i.e. the result corresponded to the projector $\Pi_0 = |\bar{\psi_i}\rangle\langle\bar{\psi_i}|$).

(f) Alice sends Bob the index $j = 1, \ldots, m$ of such a column classically. Deduce that in this way, Alice can remotely prepare n states in Bob's system with approximately 1 bit of classical communication per state (in the limit $n \to \infty$).

Lecture 6

Measuring distances and errors

Concept	Math translation
The trace distance between states ρ and σ corresponds to how difficult it is to *distinguish* the states ρ and σ	The trace distance $T(\rho,\sigma) = \frac{1}{2}\|\rho - \sigma\|_1$. Helström's Theorem 6.8 states that the optimal measurement distinguishes ρ and σ (when given either with 50% probability) with probability $$p = \frac{1}{2} + \frac{1}{2}T(\rho,\sigma).$$
Pure states are close if their *overlap* is close to 1.	The *fidelity* $F(\rho,\sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$ is given by $\lvert\langle\phi\vert\psi\rangle\rvert$ for pure states. The *purified distance* is $$P(\rho,\sigma) = \sqrt{1 - F(\rho,\sigma)^2}.$$
The fidelity is the maximal overlap between purifications.	Uhlmann's Theorem 6.12: $$F(\rho_A,\sigma_A) = \max_{\lvert\phi_{AR}\rangle,\lvert\psi_{AR}\rangle} \lvert\langle\phi_{AR}\vert\psi_{AR}\rangle\rvert$$ where $\lvert\phi_{AR}\rangle, \lvert\psi_{AR}\rangle$ are purifications of ρ_A and σ_A.
A quantum channel Φ_A is close to the identity channel if it preserves entanglement with a reference system.	The *entanglement purified distance* with respect to ρ_A is $$P((\Phi_A \otimes \mathcal{I}_R)(\rho_{AR}), \rho_{AR})$$ where ρ_{AR} is a purification of ρ_A.

In information theory it is often useful to allow small errors, or some small probability of having an error. For instance, if we want to transmit information over a noisy channel, we will want to use an error correcting code to protect the information from the errors. An example in classical information theory would be where we have the binary symmetric channel from Example 4.3 where a bit flips with probability p. If we want to send over a bit of information, one trick we could use is to simply send the same message over many times (say n times). The receiver obtains some string of symbols, and guesses that the original message was the one that occurs most in the string. This leads to an error in transmitting the message only if more than $n/2$ of the bits flip. If $p < \frac{1}{2}$ this happens with very small probability if n is large! In fact, for fixed p, we can make the probability of error as small as we like by taking n sufficiently large. However, if $p > 0$, the probability never becomes *exactly zero* for finite n.

The point of this story is that in many situations in information theory, if we allow *no probability of error whatsoever* we have no capacity to transfer information, while we can transfer information at arbitrarily small probability of error. This is one example where we find that it is crucial to have good measures of what we mean by 'probability of error'. In this lecture we will introduce various such measures for quantum states and channels and derive their most important properties.

What we need is a notion to quantify the *distance* between two quantum states. This means we would like to have a *metric* on the set of quantum states, i.e. a function $d(\rho, \sigma)$ which measures the distance between the states ρ and σ. Formally, a metric on the set of quantum states $\mathrm{S}(\mathcal{H})$ is a function $d : \mathrm{S}(\mathcal{H}) \times \mathrm{S}(\mathcal{H}) \to \mathbb{R}_{\geq 0}$ such that

(a) $d(\rho, \sigma) = d(\sigma, \rho)$ for all $\rho, \sigma \in \mathrm{S}(\mathcal{H})$.

(b) $d(\rho, \sigma) \geq 0$ with equality $d(\rho, \sigma) = 0$ if and only if $\rho = \sigma$.

(c) The *triangle inequality* $d(\rho, \sigma) + d(\sigma, \tau) \geq d(\rho, \tau)$ holds for all $\rho, \sigma, \tau \in \mathrm{S}(\mathcal{H})$.

We will see two options for defining distances on $\mathrm{S}(\mathcal{H})$. One is based on the *trace norm*, and the other is based on the *fidelity*.

6.1 Norms of operators

Since quantum states are operators, to measure whether two quantum states are close, we need a notion of distance on spaces of operators. There are various such notions, and here we will introduce an appropriate *norm* for linear operators. For us, the most relevant norm is the trace norm.

Definition 6.1. If $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ then the *trace norm*, also known as the 1-norm, is the sum of the singular values of M.

Basic properties of the trace norm are:

Lemma 6.2. *Let $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$.*

(a) The trace norm can be computed as $\|M\|_1 = \mathrm{tr}[\sqrt{M^\dagger M}]$.

(b) $\|M\|_1 = \|M^\dagger\|_1 = \|M^\mathsf{T}\|_1 = \|\overline{M}\|_1$.

(c) If V and W are isometries $\|VMW\|_1 = \|M\|_1$.

(d) If $\mathcal{H} = \mathcal{K}$ and $M = M^\dagger$ has spectrum $\lambda_1, \ldots, \lambda_n$,

$$\|M\|_1 = \sum_{i=1}^{n} |\lambda_i|.$$

Verifying these properties is Exercise 6.2.

The last observation in Lemma 6.2 is that for Hermitian M the trace norm equals the sum of the *absolute values of the eigenvalues*. In particular, the trace norm of a quantum state $\rho \in \mathrm{S}(\mathcal{H})$ is $\|\rho\|_1 = 1$, since ρ has nonnegative eigenvalues summing to 1. This property makes the trace norm especially suited for measuring distances between quantum states.

Two other norms which are occasionally useful are the following:

(a) The Hilbert-Schmidt norm $\|\cdot\|_{\mathrm{HS}}$, or 2-norm, derives from an inner product, the *Hilbert-Schmidt inner product*

$$\langle M, N\rangle_{\mathrm{HS}} := \operatorname{tr}[M^\dagger N]$$
$$\|M\|_2 := \sqrt{\langle M, M\rangle_{\mathrm{HS}}} = \sqrt{\operatorname{tr}[M^\dagger M]}.$$

If we express M, N as matrices in some basis, it is easy to verify that

$$\langle M, N\rangle_{\mathrm{HS}} = \sum_{i,j} \overline{M_{ij}} N_{ij}.$$

This inner product is equivalent to viewing the matrices as vectors and taking the usual inner product. The *Cauchy-Schwarz inequality* states in this case that

$$|\operatorname{tr}[M^\dagger N]|^2 = |\langle M, N\rangle_{\mathrm{HS}}|^2 \leq \|M\|_2^2 \|N\|_2^2 = \operatorname{tr}[M^\dagger M] \operatorname{tr}[N^\dagger N],$$

and hence

$$|\operatorname{tr}[MN]| \leq \|M\|_2 \|N\|_2.$$

(b) The *operator norm*, or ∞-norm, is defined to be the largest singular value of the operator and can alternatively be expressed as

$$\|M\|_\infty = \max_{\|v\|=1} \|M|v\rangle\| = \max_{\|v\|=1} \sqrt{\langle v|M^\dagger M|v\rangle} = \max_{\|v\|=\|w\|=1} |\langle w|M|v\rangle|$$

It has the property that it is submultiplicative:

$$\|MN\|_\infty \leq \|M\|_\infty \|N\|_\infty. \tag{6.1}$$

These three norms are special cases of the Schatten p-norms, for $p = 1, 2, \infty$, as explained in Appendix A.3.

A useful fact about the trace norm is that it has a variational characterization.

Lemma 6.3. *The trace norm has the following characterization: for $M \in \mathrm{Lin}(\mathcal{H})$*

$$\|M\|_1 = \max_{U \in \mathrm{U}(\mathcal{H})} |\operatorname{tr}[MU]|. \tag{6.2}$$

Moreover, for all $N \in \mathrm{Lin}(\mathcal{H})$

$$|\operatorname{tr}[MN]| \leq \|MN\|_1 \leq \|M\|_1 \|N\|_\infty. \tag{6.3}$$

Proof. Let

$$M = \sum_{i=1}^r s_i |e_i\rangle\langle f_i|$$

be a singular value decomposition. If $N \in \mathrm{Lin}(\mathcal{H})$

$$|\mathrm{tr}[MN]| = \left|\mathrm{tr}\Big[\sum_i s_i |e_i\rangle\langle f_i|N\Big]\right| \leq \sum_i s_i \underbrace{|\langle f_i|N|e_i\rangle|}_{\leq \|N\|_\infty} \leq \|M\|_1 \|N\|_\infty. \tag{6.4}$$

In particular, for any $U \in \mathrm{U}(\mathcal{H})$, we have $|\mathrm{tr}[MU]| \leq \|M\|_1$. On the other hand, extend e_i and f_i to a basis, and define the unitary $V = \sum_i |f_i\rangle\langle e_i|$. Then

$$\max_{U \in \mathrm{U}(\mathcal{H})} |\mathrm{tr}[MU]| \geq \mathrm{tr}[MV] = \mathrm{tr}\Big[\sum_i s_i |e_i\rangle\langle e_i|\Big] = \|M\|_1$$

proving Eq. (6.2). We may now use Eq. (6.2) to prove Eq. (6.3): we have $|\mathrm{tr}[MN]| \leq \|MN\|_1$ and there exists U such that

$$\|MN\|_1 = |\mathrm{tr}[MNU]| \leq \|M\|_1 \|NU\|_\infty = \|M\|_1 \|N\|_\infty$$

using Eq. (6.4) for the inequality. □

6.2 Trace distance

For the trace distance we use the trace norm (normalized by a factor of $\frac{1}{2}$) as a distance measure on the set of states on a Hilbert space $\mathcal{H}$.

Definition 6.4 (Trace distance)**.** Let $\rho, \sigma \in \mathrm{S}(\mathcal{H})$. Then the *trace distance* between ρ and σ is

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

The trace distance defines a metric on the set of quantum states (this is immediate as it derives from a norm). The following lemma is useful for computing it:

Lemma 6.5. *For $\rho, \sigma \in \mathrm{S}(\mathcal{H})$ we have*

(a) If $\rho - \sigma$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, then the trace distance is the sum of the positive eigenvalues $T(\rho, \sigma) = \sum\limits_{\lambda_i > 0} \lambda_i$.

(b) The trace distance has the following variational characterization:

$$T(\rho, \sigma) = \max_{0 \leq Q \leq \mathbb{1}} \mathrm{tr}[Q(\rho - \sigma)].$$

The maximum is attained by an operator Q which is a projection.

Proof. (a) Note that $\mathrm{tr}[\rho - \sigma] = \mathrm{tr}[\rho] - \mathrm{tr}[\sigma] = 0$. So, if the Hermitian operator $M = \rho - \sigma$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ we have

$$\mathrm{tr}[\rho - \sigma] = \sum_{i=1}^n \lambda_i = 0$$
$$T(\rho, \sigma) = \frac{1}{2}\|\rho - \sigma\|_1 = \frac{1}{2}\sum_{i=1}^n |\lambda_i| = \sum_{i:\lambda_i > 0} \lambda_i.$$

In the last equality we use that the λ_i sum to zero, so

$$\sum_{i=1}^{n} |\lambda_i| = \sum_{i:\lambda_i>0} \lambda_i - \sum_{i:\lambda_i<0} \lambda_i = 2 \sum_{i:\lambda_i>0} \lambda_i.$$

(b) If

$$\rho - \sigma = \sum_{i=1}^{d} \lambda_i |e_i\rangle\langle e_i|$$

is a spectral decomposition then for any $0 \leq Q \leq \mathbb{1}$

$$\begin{aligned} \mathrm{tr}[(\rho-\sigma)Q] = \sum_{i=1}^{d} \lambda_i \underbrace{\langle e_i|Q|e_i\rangle}_{\geq 0} &= \sum_{\lambda_i>0} \lambda_i \langle e_i|Q|e_i\rangle + \sum_{\lambda_i<0} \lambda_i \langle e_i|Q|e_i\rangle \\ &\leq \sum_{\lambda_i>0} \lambda_i \underbrace{\langle e_i|Q|e_i\rangle}_{\leq 1} \leq \sum_{\lambda_i>0} \lambda_i = T(\rho,\sigma) \end{aligned}$$

using (a) in the last step. Moreover, equality can be achieved if we let Q be the projection operator $Q = \sum_{\lambda_i>0} |e_i\rangle\langle e_i|$, we have equality.

□

Example 6.6. Let

$$\rho_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}| \quad \text{and} \quad \sigma_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$$

be the two-qubit maximally entangled state and maximally correlated state respectively. For their trace distance we take the difference

$$\rho_{AB} - \sigma_{AB} = \frac{1}{2}(|00\rangle\langle 11| + |11\rangle\langle 00|) = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

The two nonzero eigenvalues are $\pm\frac{1}{2}$. The trace distance is $T(\rho_{AB}, \sigma_{AB}) = \frac{1}{2}(\frac{1}{2} + \frac{1}{2}) = \frac{1}{2}$.

It has the following useful properties:

Lemma 6.7. *For $\rho, \sigma \in \mathrm{S}(\mathcal{H})$ we have*

(a) $0 \leq T(\rho,\sigma) \leq 1$, *and* $T(\rho,\sigma) = 0$ *if and only if* $\rho = \sigma$.

(b) The trace distance is invariant under isometries: if $V \in \mathrm{Isom}(\mathcal{H},\mathcal{K})$ *then*

$$T(V\rho V^\dagger, V\sigma V^\dagger) = T(\rho,\sigma).$$

(c) The trace distance is monotonic under partial trace: if $\rho_{AB}, \sigma_{AB} \in \mathrm{S}(AB)$ *then*

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

(d) The trace distance is monotonic under quantum channels: if $\Phi_{A\to B} \in \mathrm{C}(A,B)$ *and* $\rho_A, \sigma_A \in \mathrm{S}(A)$ *then*

$$T(\Phi_{A\to B}(\rho_A), \Phi_{A\to B}(\sigma_A)) \leq T(\rho_A,\sigma_A).$$

Proof. (a) Since it is a norm we have $\|\rho - \sigma\|_1 \geq 0$ with equality if and only if $\rho = \sigma$. The upper bound follows from the triangle inequality:

$$T(\rho,\sigma) = \frac{1}{2}\|\rho-\sigma\|_1 \leq \frac{1}{2}\left(\|\rho\|_1 + \|\sigma\|_1\right) = 1.$$

(b) We have

$$T(V\rho V^\dagger, V\sigma V^\dagger) = \frac{1}{2}\|V\rho V^\dagger - V\sigma V^\dagger\|_1 = \frac{1}{2}\|V(\rho-\sigma)V^\dagger\|_1$$

and the result follows from the invariance of the 1-norm under isometries in Lemma 6.2.

(c) We use Lemma 6.5 (b):

$$\begin{aligned} T(\rho_A,\sigma_A) &= \max_{0\leq Q_A \leq \mathbb{1}_A} \operatorname{tr}[Q_A(\rho_A - \sigma_A)] \\ &= \max_{0\leq Q_A \leq \mathbb{1}_A} \operatorname{tr}[(Q_A \otimes I_B)(\rho_{AB} - \sigma_{AB})] \\ &\leq \max_{0\leq Q_{AB} \leq \mathbb{1}_{AB}} \operatorname{tr}[Q_{AB}(\rho_{AB} - \sigma_{AB})] \\ &= T(\rho_{AB},\sigma_{AB}). \end{aligned}$$

The inequality holds because if an operator Q_A satisfies $0 \leq Q_A \leq \mathbb{1}_A$, then the operator $Q_{AB} := Q_A \otimes \mathbb{1}_B$ satisfies $0 \leq Q_{AB} \leq \mathbb{1}_{AB}$.

(d) This follows from parts (b) and (c) by considering a Stinespring representation of the channel $\Phi_{A\to B}$ (Theorem 4.16).

□

[MW: I change this a bit since I felt the 'bias of $\mu(x)$ for a general measurement' discussion was a bit obfuscating the state discrimination task. Let's discuss!] The trace distance has a natural operational interpretation, based on (b) of Lemma 6.5. Suppose we are given an unknown quantum state, and we know that with probability $\frac{1}{2}$ we received a state ρ and with probability $\frac{1}{2}$ we received a state σ. We are allowed to do a measurement, and then we have to guess whether the state was ρ or σ. In this scenario the optimal probability of guessing correctly is directly related to $T(\rho,\sigma)$. Indeed, suppose μ is any two-outcome measurement, where outcome 0 means that we guess that the state is ρ, while outcome 1 means that we guess the state is σ. Then the probability of guessing correctly is given by the formula $\frac{1}{2}(\operatorname{tr}[\mu(0)\rho] + \operatorname{tr}[\mu(1)\sigma])$. Using (b) of Lemma 6.5 and the fact that $\mu(0) + \mu(1) = \mathbb{1}$, one readily obtains the following result:

Theorem 6.8 (Helstrom). *Let $\rho, \sigma \in \mathrm{S}(\mathcal{H})$. Suppose that with probability $\frac{1}{2}$ we received the state ρ and with probability $\frac{1}{2}$ we received the state σ. Then the optimal probability of identifying the correct staste by a two-outcome measurement is given by*

$$p_{opt} = \frac{1}{2} + \frac{1}{2} T(\rho, \sigma).$$

We already sketched the proof; verifying the details is Exercise 6.3.

In (d) of Lemma 6.7 we saw that the trace distance is monotone under quantum channels. The interpretation of this result is quite intuitive in view of Helstrom's theorem: when we apply an operation to a quantum system, we potentially add some *noise*, making it harder to distinguish states. If the operation is not noisy (an isometry), then this does not change the trace distance between states. On the other hand, if we consider the completely depolarizing channel or another channel that is 'maximally noisy' in that it maps all states to the same state, we lose all distinguishability.

We can also use the operational interpretation furnished by Helstrom's theorem to give a second proof of the monotonicity. Consider a quantum channel $\Phi_{A\to B}$ be a quantum channel and two states $\rho_A, \sigma_A \in \mathrm{S}(A)$. Suppose that μ_B is a two-outcome measurement on B that optimally distinguishes $\Phi_{A\to B}(\rho_A)$ and $\Phi_{A\to B}(\sigma_A)$. Let $\Phi^{\mu}_{B\to X}$ denote the corresponding measurement channel (Section 5.2). Then $\Phi^{\mu}_{B\to X} \circ \Phi_{A\to B}$ is a quantum-to-classical channel and hence corresponds to a measurement ν_A such that

$$\mathrm{tr}[\nu_A(x) M_A] = \mathrm{tr}[\mu_B(x) \Phi_{A\to B}(M_A)]$$

for any M_A and hence in particular for $M_A \in \{\rho_A, \sigma\}$. Hence the measurement ν_A distinguishes ρ_A and σ_A at least as well as the measurement μ_B distinguishes $\Phi_{A\to B}(\rho_A)$ and $\Phi_{A\to B}(\sigma_A)$. The monotonicity now follows from Helstrom's theorem.

[MW: I find the argument by passing to measurement channels and back a bit complicated. Unfortunately we don't have adjoint channels available, otherwise we could simply define $\nu_A(x) := \Phi^*(\mu_B(x))$. However, we could either restrict to the partial trace situation (it's very concrete: measurements on A are special measurements on B) or consider a Stinespring representation to define ν_A in erms of μ_B. Perhaps that would be more transparent?]

Trace distance for classical states

If p_X and q_X are probability distributions on some classical system X with associated density matrices

$$\rho_X = \sum_x p_X(x) |x\rangle\langle x| \quad \text{and} \quad \sigma_X = \sum_x q_X(x) |x\rangle\langle x|$$

then we see directly (since ρ_X and σ_X are diagonal in the same basis) that

$$T(\rho_X, \sigma_X) = \frac{1}{2} \sum_x |p_X(x) - q_X(x)|$$

which corresponds to the usual *statistical distance* between probability distributions, and we will also write $T(p_X, q_X)$ for $T(\rho_X, \sigma_X)$.

Trace distance for pure states

If $\phi, \psi \in \mathcal{H}$ and $\rho = |\phi\rangle\langle\phi|$ and $\sigma = |\psi\rangle\langle\psi|$ are pure states, we may also compute their trace distance to be

$$T(\rho, \sigma) = \sqrt{1 - |\langle\phi|\psi\rangle|^2} \tag{6.5}$$

as you may show in Exercise 6.4.

6.3 Fidelity and purified distance

From Eq. (6.5) we see that the trace distance between pure states may be computed in terms of their *overlap* $|\langle\phi|\psi\rangle|$, and that the states are close if their overlap is close to 1. Indeed, if $\phi, \psi \in \mathcal{H}$ are normalized vectors, with an angle $\theta \in [-\pi, \pi]$ between them $|\langle\phi|\psi\rangle| = \cos(\theta)$ and at small angle, the overlap is close to 1. The quantity $|\langle\phi|\psi\rangle|$ is also known as the *fidelity* and there is a natural extension to mixed states. We observe that if $\rho = |\phi\rangle\langle\phi|$ and $\sigma = |\psi\rangle\langle\psi|$ then

$$|\langle\phi|\psi\rangle| = \sqrt{\langle\psi|\phi\rangle\langle\phi|\psi\rangle} = \sqrt{\operatorname{tr}[\rho\sigma]}.$$

A natural guess could therefore be to take $\operatorname{tr}[\rho\sigma]$ as our measure. However, if $\rho = \sigma$ we do not necessarily have $\operatorname{tr}[\rho^2] = 1$ and this is not quite the right choice. It turns out that the correct definition is the following:

Definition 6.9 (Fidelity)**.** Let $\rho, \sigma \in \mathrm{S}(\mathcal{H})$. Then the *fidelity* between ρ and σ is defined as

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

Let us unpack this definition. First of all,

$$\|\sqrt{\rho}\sqrt{\sigma}\|_1 = \operatorname{tr}\left[\sqrt{(\sqrt{\rho}\sqrt{\sigma})^\dagger\sqrt{\rho}\sqrt{\sigma}}\right] = \operatorname{tr}\left[\sqrt{\sqrt{\sigma}\sqrt{\rho}\sqrt{\rho}\sqrt{\sigma}}\right] = \operatorname{tr}\left[\sqrt{\sqrt{\sigma}\rho\sqrt{\sigma}}\right]$$

Note that it is *not* the case that $\sqrt{\sqrt{\sigma}\rho\sqrt{\sigma}} = \sigma^{\frac{1}{4}}\sqrt{\rho}\sigma^{\frac{1}{4}}$ in general. While it is not immediate from the definition, the fidelity is symmetric in its arguments, since

$$\|\sqrt{\rho}\sqrt{\sigma}\|_1 = \|(\sqrt{\rho}\sqrt{\sigma})^\dagger\|_1 = \|\sqrt{\sigma}\sqrt{\rho}\|_1 = \operatorname{tr}\left[\sqrt{\sqrt{\sigma}\rho\sqrt{\sigma}}\right].$$

If $\rho = |\phi\rangle\langle\phi|$ is pure, then $\sqrt{\rho} = \rho$, and we have

$$\sqrt{\rho}\sigma\sqrt{\rho} = |\phi\rangle\langle\phi|\sigma|\phi\rangle\langle\phi|$$

which has rank one. This implies that we can take the square root outside the trace

$$\operatorname{tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right] = \sqrt{\operatorname{tr}[|\phi\rangle\langle\phi|\sigma|\phi\rangle\langle\phi|]}.$$

We find that

$$F(\rho, \sigma) = \sqrt{\langle\phi|\sigma|\phi\rangle} \tag{6.6}$$

In particular, if $\sigma = |\psi\rangle\langle\psi|$ is also pure

$$F(\rho, \sigma) = |\langle\phi|\psi\rangle|.$$

Remark 6.10. Around half of the quantum information community defines the fidelity as the *square* of our $F(\rho, \sigma)$ (so for pure states it would be the overlap squared). This is good to keep in mind when consulting the literature.

Example 6.11. Let us continue with Example 6.6 and compute the fidelity between the maximally entangled and the maximally correlated qubit states ρ_{AB} and σ_{AB}. Since ρ_{AB} is pure,

$$\begin{aligned} F(\rho_{AB}, \sigma_{AB})^2 &= \langle \Phi^+_{AB} | \sigma_{AB} | \Phi^+_{AB} \rangle \\ &= \frac{1}{4} \left(\langle 00| + \langle 11| \right) \left(|00\rangle\langle 00| + |11\rangle\langle 11| \right) \left(|00\rangle + |11\rangle \right) \\ &= \frac{1}{4}(1+1) = \frac{1}{2} \end{aligned}$$

so $F(\rho_{AB}, \sigma_{AB}) = \frac{1}{\sqrt{2}}$.

Why did we introduce the fidelity as a distance measure? The trace distance had a good operational interpretation. The fidelity has as its main advantage its compatibility with taking purifications. We have the following central result.

Theorem 6.12 (Uhlmann)**.** *Suppose $\rho_A, \sigma_A \in \mathrm{S}(A)$. Let R be a reference system such that both ρ_A and σ_A have purifications on AR. Then,*

$$F(\rho_A, \sigma_A) = \max_{|\phi_{AR}\rangle, |\psi_{AR}\rangle} |\langle \phi_{AR} | \psi_{AR} \rangle|,$$

where the maximum is over purifications $|\phi_{AR}\rangle$ and $|\psi_{AR}\rangle$ of ρ_A and σ_A, respectively. Alternatively, if $|\phi_{AR}\rangle$ and $|\psi_{AR}\rangle$ are some fixed purifications of ρ_A and σ_A, then

$$F(\rho_A, \sigma_A) = \max_{U_R \in \mathrm{U}(R)} |\langle \phi_{AR} | \mathbb{1}_A \otimes U_R | \psi_{AR} \rangle|.$$

[FW: Maybe compare to the Amsterdam version, I tried to condense the proof a bit (especially the second part with arbitrary register R). Perhaps good to check whether it's an improvement.] [MW: I made another iteration to condense it some more. What do you think? The old version is commented out.]

Proof. The second statement, maximizing over U_R, is equivalent to the first one, since by Lemma 2.12 any two purifications on R of the same state are related by a unitary on R. This also shows that the maximum in the second statement does not depend on the choice of purifications, so it suffices to prove it for a *single* pair of purifications.

We first consider the case when $\mathcal{H}_R = \mathcal{H}_A$ and use the standard purifications from Eq. (2.8):

$$\begin{aligned} |\phi_{AR}\rangle &= \sum_a (\sqrt{\rho_A} \otimes \mathbb{1}_R)|aa\rangle = \sqrt{d}\,(\sqrt{\rho_A} \otimes \mathbb{1}_R)|\Phi^+_{AR}\rangle, \\ |\psi_{AR}\rangle &= \sum_a (\sqrt{\sigma_A} \otimes \mathbb{1}_R)|aa\rangle = \sqrt{d}\,(\sqrt{\sigma_A} \otimes \mathbb{1}_R)|\Phi^+_{AR}\rangle, \end{aligned}$$

for some choice of basis $|a\rangle$ of $\mathcal{H}_A$ and $d := |A|$. Then, for any unitary $U \in \mathrm{U}(R) = \mathrm{U}(A)$,

$$|\langle \phi_{AR} | \mathbb{1}_A \otimes U_R | \psi_{AR} \rangle| = \left| d \langle \Phi^+_{AR} | \sqrt{\rho_A}\sqrt{\sigma_A} \otimes U_R | \Phi^+_{AR} \rangle \right| = \left| \operatorname{tr}\left[\sqrt{\rho_A}\sqrt{\sigma_A} U_A^{\mathsf{T}} \right] \right|$$

using Lemma 2.18, and hence, by Lemma 6.3,

$$\max_{U \in \mathrm{U}(R)} |\langle \phi_{AR} | \mathbb{1}_A \otimes U_R | \psi_{AR} \rangle| = \|\sqrt{\rho_A}\sqrt{\sigma_A}\|_1 = F(\rho_A, \rho_\sigma).$$

Thus the theorem is proved in case that $\mathcal{H}_R = \mathcal{H}_A$.

To extend it to arbitrary reference systems, it suffices to show that for any $|R| \leq |S|$ and for any two purifications $|\phi_{AR}\rangle, |\psi_{AR}\rangle$ of ρ_A, σ_A, there are purifications $|\phi_{AS}\rangle, |\psi_{AS}\rangle$ such that

$$\max_{U \in \mathrm{U}(R)} |\langle\phi_{AR}|\mathbb{1}_A \otimes U_R|\psi_{AR}\rangle| = \max_{W \in \mathrm{U}(S)} |\langle\phi_{AS}|\mathbb{1}_A \otimes W_S|\psi_{AS}\rangle|.$$

To this end, pick an arbitrary isometry $V \in \mathrm{Isom}(R,S)$ and choose $|\phi_{AS}\rangle := (I_A \otimes V_{R\to S})|\phi_{AR}\rangle$ and $|\psi_{AS}\rangle := (I_A \otimes V_{R\to S})|\psi_{AR}\rangle$. Then, using Lemma 6.3,

$$\max_{U \in \mathrm{U}(R)} |\langle\phi_{AR}|\mathbb{1}_A \otimes U_R|\psi_{AR}\rangle| = \max_{U \in \mathrm{U}(R)} |\mathrm{tr}[M_R U_R]| = \|M_R\|_1,$$

where $M_R := \mathrm{tr}_A[|\psi_{AR}\rangle\langle\phi_{AR}|]$, while

$$\begin{aligned} \max_{W \in \mathrm{U}(S)} |\langle\phi_{AS}|\mathbb{1}_A \otimes W_S|\psi_{AS}\rangle| &= \max_{W \in \mathrm{U}(S)} |\langle\phi_{AR}|\mathbb{1}_A \otimes V^\dagger_{R\to S} W_S V_{R\to S}|\psi_{AR}\rangle| \\ &= \max_{W \in \mathrm{U}(S)} \left|\mathrm{tr}\left[V^\dagger_{R\to S} W_S V_{R\to S} M_R\right]\right| \\ &= \max_{W \in \mathrm{U}(S)} \left|\mathrm{tr}\left[W_S V_{R\to S} M_R V^\dagger_{R\to S}\right]\right| = \|V_{R\to S} M_R V^\dagger_{R\to S}\|_1 = \|M_R\|_1, \end{aligned}$$

using the invariance under isometries of the trace norm in the last step. □

The fidelity has the following basic properties.

Lemma 6.13. *Suppose $\rho, \sigma \in \mathrm{S}(\mathcal{H})$.*

(a) $0 \leq F(\rho,\sigma) \leq 1$ and $F(\rho,\sigma) = 1$ if and only if $\rho = \sigma$.

(b) The fidelity is invariant under isometries: if $V \in \mathrm{Isom}(\mathcal{H},\mathcal{K})$ then

$$F(V\rho V^\dagger, V\sigma V^\dagger) = F(\rho,\sigma).$$

(c) The fidelity is monotonic under partial trace: if $\rho_{AB}, \sigma_{AB} \in \mathrm{S}(AB)$ then

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

(d) The fidelity is monotonic under quantum channels: if $\Phi_{A\to B} \in \mathrm{C}(A,B)$ and $\rho_A, \sigma_A \in \mathrm{S}(A)$, then

$$F(\Phi_{A\to B}(\rho_A), \Phi_{A\to B}(\sigma_A)) \geq F(\rho_A, \sigma_A).$$

The proof is Exercise 6.5. The inequality in the monotonicity is in the other direction than for the trace distance. This is sensible: the states may get closer to each other as we apply a quantum channel, so their fidelity *increases.*

One may also compute the fidelity for classical states. If p_X and q_X are probability distributions on some classical system X with associated density matrices

$$\rho_X = \sum_x p_X(x)|x\rangle\langle x| \quad \text{and} \quad \sigma_X = \sum_x q_X(x)|x\rangle\langle x|$$

then we see directly (since ρ_X and σ_X are diagonal in the same basis) that

$$F(\rho_X, \sigma_X) = \sum_x \sqrt{p_X(x) q_X(x)}$$

which one may also denote by $F(p_X, q_X)$. This is less commonly used as a similarity measure for in probability theory (the fidelity is mainly useful because of its relation to purifications as in Theorem 6.18, so it is more natural in the quantum setting).

The fidelity and trace distance can be related to each other. We already saw in Eq. (6.5) how they may converted for pure states. In general we have the following *Fuchs-van de Graaf inequalities*.

Lemma 6.14. *For any $\rho, \sigma \in \mathrm{S}(\mathcal{H})$ it holds that*

$$1 - F(\rho, \sigma) \leq T(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)^2}. \tag{6.7}$$

The proof is Exercise 6.10.

Remark 6.15. The Fuchs-van de Graaf inequalities show that the trace distance and fidelity are very similar distance measures: if ρ and σ are close to each other in the one measure, they are also close in the other measure. Note that the bounds in Eq. (6.7) are independent of the dimension of the Hilbert space. If these two measures are similar, why are we using both of them? The reason for this is pragmatic: the trace distance and the fidelity have useful properties in different situations:

- The trace distance has a good operational interpretation, by Theorem 6.8 related to distinguishing states. The fact that is derives from a norm can be useful when computing bounds.
- The fidelity is convenient for pure states. Moreover, Uhlmann's theorem is often a powerful tool (as we will see later).

When analyzing a information processing protocol one should choose the distance measure which is most convenient for the analysis. If you would like to use properties from both measures (say, you want to compute the probability of distinguishing two states, but you would also like to apply Uhlmann's theorem) you can simply convert between the two measures using Lemma 6.14, possibly at the cost of incurring a square root on the dependence on the error.

The purified distance

The fidelity is not a metric (this is clear since states are close when they have fidelity close to 1). We may define a metric based on the fidelity which is known as the purified distance.

Definition 6.16 (Purified distance)**.** Suppose $\rho, \sigma \in \mathrm{S}(\mathcal{H})$. Then the *purified distance* between ρ and σ is defined as

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

We see that for pure states ρ, σ we have

$$P(\rho, \sigma) = T(\rho, \sigma). \tag{6.8}$$

The following properties are a direct consequence of the corresponding properties of the fidelity in Lemma 6.13

Lemma 6.17. *Suppose $\rho, \sigma \in \mathrm{S}(\mathcal{H})$.*

(a) $0 \leq P(\rho, \sigma) \leq 1$ *and* $P(\rho, \sigma) = 0$ *if and only if* $\rho = \sigma$.

(b) The purified distance is invariant under isometries: if $V \in \mathrm{Isom}(\mathcal{H}, \mathcal{K})$ *then*

$$P(V\rho V^\dagger, V\sigma V^\dagger) = P(\rho, \sigma).$$

(c) The purified distance is monotonic under partial trace: if $\rho_{AB}, \sigma_{AB} \in \mathrm{S}(AB)$ *then*

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

(d) The purified distance is monotonic under quantum channels: if $\Phi_{A\to B} \in \mathrm{C}(A, B)$ *and* $\rho_A, \sigma_A \in \mathrm{S}(A)$ *then*

$$P(\Phi_{A\to B}(\rho_A), \Phi_{A\to B}(\sigma_A)) \leq P(\rho_A, \sigma_A).$$

As a direct consequence of Uhlmann's theorem and the monotonicity of the purified distance we have the following result, which you may prove in Exercise 6.7.

Theorem 6.18. *For* $\rho_A, \sigma_A \in \mathrm{S}(A)$ *we have*

$$P(\rho_A, \sigma_A) = \min_{\rho_{AR}, \sigma_{AR}} P(\rho_{AR}, \sigma_{AR})$$

where the minimum is over all states ρ_{AR}, σ_{AR} *such that* $\operatorname{tr}_R[\rho_{AR}] = \rho_A$ *and* $\operatorname{tr}_R[\sigma_{AR}] = \sigma_A$. *We can restrict the minimization to purifications* ρ_{AR} *and* σ_{AR} *of* ρ_A *and* σ_A *respectively.*

We will now verify that the purified distance defines a metric.

Lemma 6.19. *The purified distance defines a metric on* $\mathrm{S}(\mathcal{H})$.

Proof. From Lemma 6.13 it follows that for all $\rho, \sigma \in \mathrm{S}(\mathcal{H})$ we have $P(\rho, \sigma) = P(\sigma, \rho)$ and $P(\rho, \sigma) \geq 0$ with equality if and only if $\rho = \sigma$. It remains to prove the triangle inequality. We let $\mathcal{H} = \mathcal{H}_A$, and $\rho_A, \sigma_A, \tau_A \in \mathrm{S}(A)$. We need to show that $P(\rho_A, \sigma_A) + P(\sigma_A, \tau_A) \geq P(\rho_A, \tau_A)$. To this end, let ρ_{AR}, σ_{AR} and τ_{AR} be purifications on some additional system R such that $P(\rho_{AR}, \sigma_{AR}) = P(\rho_A, \sigma_A)$ and $P(\sigma_{AR}, \tau_{AR}) = P(\sigma_A, \tau_A)$ (note that we can chose a fixed purification σ_{AR} and then apply Theorem 6.18 to find ρ_{AR} and τ_{AR}). Then

$$\begin{aligned} P(\rho_A, \sigma_A) + P(\sigma_A, \tau_A) &= P(\rho_{AR}, \sigma_{AR}) + P(\sigma_{AR}, \tau_{AR}) \\ &= T(\rho_{AR}, \sigma_{AR}) + T(\sigma_{AR}, \tau_{AR}) \\ &\geq T(\rho_{AR}, \tau_{AR}) = P(\rho_{AR}, \tau_{AR}) \geq P(\rho_A, \tau_A) \end{aligned}$$

using Eq. (6.8), the triangle inequality for the trace distance and monotonicity of the purified distance. □

Finally, we note that Lemma 6.14 directly implies the following inequalities relating the trace distance and the purified distance:

$$T(\rho, \sigma) \leq P(\rho, \sigma) \leq \sqrt{1 - (1 - T(\rho, \sigma))^2} \leq \sqrt{2T(\rho, \sigma)}. \tag{6.9}$$

6.3.1 Gentle measurement lemma

Measurement can drastically perturb a quantum state. For example, if we measure the $|+\rangle$ in the standard basis, the post-measurement state will be $|0\rangle$ or $|1\rangle$ depending on the equally likely outcomes. However, there are also situations where the measurement does not change the state! If the state is $|0\rangle$, and we measure in the standard basis, then we get outcome 0 with probability 1 and the post-measurement state is just $|0\rangle$. In general, one could expect that if the measurement gives one specific outcome with probability very close to one, then the measurement will not disturb the state too much. As a concrete example, suppose we measure on a qubit in the standard basis on a state ρ. If the probability of finding outcome 0 is $1-\varepsilon$ for small $\varepsilon \geq 0$ then

$$1-\varepsilon = \mathrm{tr}[|0\rangle\langle 0|\rho] = \langle 0|\rho|0\rangle = F(\rho, |0\rangle\langle 0|)^2$$

so

$$P(\rho, |0\rangle\langle 0|) = \sqrt{1-F(\rho, |0\rangle\langle 0|)^2} = \sqrt{\varepsilon}$$

and by Fuchs-van de Graaf $T(\rho, |0\rangle\langle 0|) \leq \sqrt{\varepsilon}$. In other words, if the outcome 0 is very likely, the state must be close to $|0\rangle\langle 0|$.

The *gentle measurement lemma* shows that a similar fact is true for general measurements. Recall that if we perform a measurement μ on a state ρ, then after the measurement, upon finding outcome x, the post-measurement state will be

$$\rho_x = \frac{\sqrt{\mu(x)}\rho\sqrt{\mu(x)}}{\mathrm{tr}[\mu(x)\rho]}.$$

Lemma 6.20 (Gentle measurement lemma)**.** *Let ρ be a quantum state and let $\mu = \{\mu(x)|x \in \mathcal{X}\}$ be measurement such that for some $x = x_0$ and $0 \leq \varepsilon \leq 1$ the probability of outcome x_0 is at least*

$$\mathrm{tr}(\mu(x_0)\rho) \geq 1-\epsilon. \tag{6.10}$$

Then the post-measurement state given the measurement outcome x_0

$$\rho' = \frac{\sqrt{\mu(x_0)}\rho\sqrt{\mu(x_0)}}{\mathrm{tr}(\mu(x_0)\rho)}$$

satisfies

$$T(\rho, \rho') \leq P(\rho, \rho') \leq \sqrt{\varepsilon}.$$

and hence

$$T(\rho, \rho') \leq \sqrt{\epsilon}. \tag{6.11}$$

Proof. We have that

$$\begin{aligned}
F(\rho, \rho') &= \frac{1}{\sqrt{\mathrm{tr}(\mu(x_0)\rho)}}\,\mathrm{tr}\left(\sqrt{\sqrt{\rho}\sqrt{\mu(x_0)}\rho\sqrt{\mu(x_0)}\sqrt{\rho}}\right) \\
&= \frac{1}{\sqrt{\mathrm{tr}(\mu(x_0)\rho)}}\,\mathrm{tr}\left(\sqrt{\rho}\sqrt{\mu(x_0)}\sqrt{\rho}\right) \\
&= \frac{1}{\sqrt{\mathrm{tr}(\mu(x_0)\rho)}}\,\mathrm{tr}\left(\sqrt{\mu(x_0)}\rho\right)
\end{aligned}$$

$$\geq \sqrt{\operatorname{tr}(\mu(x_0)\rho)},$$

where the second equality uses that $\sqrt{\rho}\sqrt{\mu(x_0)}\sqrt{\rho}$ is positive semi-definite and

$$\left(\sqrt{\rho}\sqrt{\mu(x_0)}\sqrt{\rho}\right)^2 = \sqrt{\rho}\sqrt{\mu(x_0)}\rho\sqrt{\mu(x_0)}\sqrt{\rho}.$$

The last inequality uses $\operatorname{tr}(\sqrt{\mu(x_0)}\rho) \geq \operatorname{tr}(\mu(x_0)\rho)$ as the eigenvalues of the positive matrix $\mu(x_0)$ are bounded by 1. Therefore, $F(\rho, \rho')^2 \geq 1 - \epsilon$ and $P(\rho, \rho') \leq \sqrt{\varepsilon}$. The Fuchs-van de Graaf inequality gives $T(\rho, \rho') \leq \sqrt{\epsilon}$. □

6.4 Error measures for channels

Now that we have good measures for when states are close, the next natural question is to get a good measure for when *quantum channels* are close. Intuitively, two quantum channels are close if they are such that if you input the same state to them, they should give nearby states as output. This is indeed the way one typically defines distance measures for quantum channels. For such a definition, we need to specify two things: what measure do we use to distinguish states, and perhaps less obviously, what states are allowed to serve as input?

The entanglement fidelity

For our purposes it will suffice to be able to measure how close a channel is to the *identity channel*. The reason for this is that typically we will be interested in some process and we would like the quantum system that comes out is a good approximation to the input system. For example, this could be a communication scenario where we want the output of the communication protocol to be a reliable transmission of the input. In the next chapter we will see a concrete example!

The situation we are interested in is one where we know that a quantum system is in state ρ_A, we apply some channel Φ_A and we would like to know whether this channel acted similarly to the identity channel $\mathcal{H}_A$. A naive definition could be that we are close to the action of the identity channel if

$$F(\Phi_A(\rho_A), \rho_A) \geq 1 - \varepsilon$$

for some small $\varepsilon > 0$ (or one could think of a similar definition using the trace distance). However, this definition does not quite capture what we are looking for! For instance, the channel which discards the system A and prepares the state ρ_A clearly satisfies $\Phi_A(\rho_A) = \rho_A$. However, suppose that ρ_A is a mixture of two states ρ_1 and ρ_2 with probabilities p_1 and p_2, then just discarding and preparing the state ρ_A does not preserve this decomposition.

If $\rho_A = \sum_x p_x \rho_{A,x}$ we would like that if we think of ρ_A as an ensemble of states $\rho_{A,x}$ with probability p_x, we should have in expectation

$$\sum_x p_x P(\Phi_A(\rho_{A,x}), \rho_{A,x}) \leq \varepsilon$$

Now, it is easy to see from Exercise 6.6 that this is equivalent to

$$(\Phi_A \otimes \mathcal{I}_X)(\rho_{AX}) \approx_\varepsilon \rho_{AX}$$

for

$$\rho_{AX} = \sum_x p(x)\rho_{A,x} \otimes |x\rangle\langle x|.$$

However, ensemble interpretations are not unique, and in general we would like to allow any coupling to a reference system. A reasonable definition is the following:

Definition 6.21. Let $\Phi_A \in \mathrm{C}(A)$ and $\rho_A \in \mathrm{S}(A)$. Then the *entanglement fidelity* of Φ_A with respect to ρ_A is defined to be

$$F_E(\Phi_A, \rho_A) := \inf_{\sigma_{AR}} F((\Phi_A \otimes \mathcal{I}_R)(\sigma_{AR}), \sigma_{AR})$$

where the infimum is over all systems R and states σ_{AR} such that $\operatorname{tr}_R[\sigma_{AR}] = \rho_A$.

We see that we have to take an infimum over all possible reference systems R and any extension σ_{AR} of ρ_A. Fortunately, this infimum is attained by an *arbitrary* purification of ρ_A.

Lemma 6.22. *Suppose $\rho_{AR} = |\phi_{AR}\rangle\langle\phi_{AR}|$ is a purification of ρ_A. Then*

$$F_E(\Phi_A, \rho_A) := F((\Phi_A \otimes \mathcal{I}_R)(\rho_{AR}), \rho_{AR}) = \sqrt{\langle\phi_{AR}|\big((\Phi_A \otimes \mathcal{I}_R)(\rho_{AR})\big)|\phi_{AR}\rangle}.$$

Proof. Suppose R is an arbitrary system and σ_{AR} is such that $\operatorname{tr}_R[\sigma_{AR}] = \rho_A$. Denote $\tau_{AR} = (\Phi_A \otimes \mathcal{I}_R)(\sigma_{AR})$. Choose any purification of σ_{ARS} on some additional system S and let $\tau_{ARS} = (\Phi_A \otimes \mathcal{I}_R)(\sigma_{ARS})$. Then, by monotonicity

$$F(\tau_{ARS}, \sigma_{ARS}) \geq F(\tau_{AR}, \sigma_{AR})$$

and hence in the infimum in Definition 6.21 we may restrict to pure states $\sigma_{AR} = |\phi_{AR}\rangle\langle\phi_{AR}|$. Moreover, if we have two different purifications of ρ_A, then they are related by an isometry on the extending system, and since the fidelity is invariant under isometries we find that the value of

$$\langle\phi_{AR}|(\Phi_A \otimes \mathcal{I}_R)(\rho_{AR})|\phi_{AR}\rangle$$

is independent of the choice of purification $|\phi_{AR}\rangle$. □

Based on Lemma 6.22 we may compute $F_E(\Phi_A, \rho_A)$ for instance in terms of a Kraus representation of Φ_A.

Lemma 6.23. *Suppose that $\Phi_A(M_A) = \sum_i X_i M_A X_i^\dagger$ is a Kraus representation. Then*

$$F_E(\Phi_A, \rho_A) = \sqrt{\sum_i \left|\operatorname{tr}[X_i \rho_A]\right|^2}.$$

The proof is Exercise 6.13.

Finally, we define the analog of the entanglement fidelity using the purified distance:

Definition 6.24. Let $\Phi_A \in \mathrm{C}(A)$ and $\rho_A \in \mathrm{S}(A)$. Then the *entanglement purified distance*

$$P_E(\Phi_A, \rho_A) := \sup_{\sigma_{AR}} P((\Phi_A \otimes \mathcal{I}_R)(\sigma_{AR}), \sigma_{AR})$$

where the supremum is over all systems R and states σ_{AR} such that $\operatorname{tr}_R[\sigma_{AR}] = \rho_A$.

By Lemma 6.22, if ρ_{AR} is an arbitrary purification of ρ_A we have

$$P_E(\Phi_A, \rho_A) = P((\Phi_A \otimes \mathcal{I}_R)(\rho_{AR}), \rho_{AR}).$$

6.5 Exercises

6.1 **Distances between states:** Find the trace distance and the fidelity between the following single-qubit states:

(a) $\rho = \frac{1}{2}(|0\rangle + |1\rangle)(\langle 0| + \langle 1|)$ and $\sigma = |0\rangle\langle 0|$.
(b) $\rho = \frac{1}{3}|+\rangle\langle +| + \frac{2}{3}|-\rangle\langle -|$, $\sigma = \frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|1\rangle\langle 1|$.
(c) $\rho = \frac{1}{11}(5|0\rangle\langle 0| + 6|1\rangle\langle 1| - 4|0\rangle\langle 1| - 4|1\rangle\langle 0|)$ and $\sigma = \frac{1}{3}(|0\rangle\langle 0| + 2|1\rangle\langle 1| + |1\rangle\langle 0| + |0\rangle\langle 1|)$.
Hint: you may find the following fact useful:

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}^2 = 5\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} .$$

6.2 **Basic properties of the trace norm:** Prove Lemma 6.2.

6.3 **Helstrom's theorem:** Prove Theorem 6.8.

6.4 **Trace distance between pure states:** Let $\rho = |\phi\rangle\langle\phi|$ and $\sigma = |\psi\rangle\langle\psi|$ be pure states.

(a) By writing $|\psi\rangle = \alpha|\phi\rangle + \beta|\phi^\perp\rangle$, where $|\phi^\perp\rangle$ is some state orthogonal to $|\phi\rangle$, show that the matrix $(\rho - \sigma)$ has non-zero eigenvalues $\pm|\beta|$.
(b) Deduce that

$$T(\rho, \sigma) = \sqrt{1 - |\langle\phi|\psi\rangle|^2} .$$

6.5 **Properties of the fidelity:** Prove Lemma 6.13.

6.6 **Fidelity between classical-quantum states:** Show that for a pair of classical-quantum states

$$\rho_{XA} = \sum_x p(x)|x\rangle\langle x| \otimes \rho_{A,x} \text{ and } \sigma_{XA} = \sum_x q(x)|x\rangle\langle x| \otimes \sigma_{A,x}$$

for probability distibutions p, q on a classical register X and collections of states $\rho_{A,x}, \sigma_{A,x} \in \mathrm{S}(A)$ it holds that

$$F(\rho_{XA}, \sigma_{XA}) = \sum_x \sqrt{p(x)q(x)} F(\rho_{A,x}, \sigma_{A,x}).$$

6.7 **Uhlmann's theorem for the purified distance:** Prove Theorem 6.18.

6.8 **Extensions and fidelity:** Let $\rho_{AB} \in \mathrm{S}(AB)$ and $\sigma_A \in \mathrm{S}(A)$. Show that there exists an extension σ_{AB} (not necessarily pure) of σ_A such that $F(\rho_A, \sigma_A) = F(\rho_{AB}, \sigma_{AB})$.

6.9 **Fidelity inequalities:**

(a) Show that $|\langle\psi_1|\phi\rangle|^2 + |\langle\psi_2|\phi\rangle|^2 \leq 1 + |\langle\psi_1|\psi_2\rangle|$ for all vector vectors $|\psi_1\rangle, |\psi_2\rangle, |\phi\rangle \in \mathcal{H}$.
Hint: Upper bound the left-hand side by the largest eigenvalue of some rank-2 matrix.
(b) Show that $F(\rho_1, \sigma)^2 + F(\rho_2, \sigma)^2 \leq 1 + F(\rho_1, \rho_2)$ for all states $\rho_1, \rho_2, \sigma \in \mathrm{S}(\mathcal{H})$.
(c) Show the following 'triangle inequality': If $F(\alpha, \beta) \geq 1 - \delta$ and $F(\beta, \gamma) \geq 1 - \delta$ for any three states $\alpha, \beta, \gamma \in D(\mathcal{H})$, then $F(\alpha, \gamma) \geq 1 - 4\delta$.

6.10 **Fuchs-van de Graaf:** The goal of this exercise is to prove Lemma 6.14.

(a) Given $\rho_A, \sigma_A\,\mathrm{S}(A)$, argue that there exist purifications ρ_{AR}, σ_{AR} of ρ_A and σ_A such that

$$T(\rho_{AR}, \sigma_{AR}) = \sqrt{1 - F(\rho_A, \sigma_a)^2}$$

and use this to show that

$$T(\rho_A, \sigma_A) \leq \sqrt{1 - F(\rho_A, \sigma_A)^2}.$$

(b) Show that for any probability distributions p_X, q_X we have

$$1 - F(p_X, q_X) \leq T(p_X, q_X).$$

(c) Show that

$$1 - F(\rho_A, \sigma_A) \leq T(\rho_A, \sigma_A)$$

where you may use without proof the fact that there exists some measurement such that if p_X and q_X denote the outcome probabilities after measuring ρ_A and σ_A it holds that $F(\rho_A, \sigma_A) = F(p_X, q_X)$.

6.11 **Optimally distinguishing between quantum states:** Let $\rho \in S(A)$ be a pure state $\rho = |\psi\rangle\langle\psi|$, and let $\tau = \frac{1}{|A|}\mathbb{1}_A$ be the maximally mixed state on A.

(a) Show that ρ can be distinguished from τ using a two-outcome measurement with optimal probability

$$p_{\text{opt}} = \frac{2|A| - 1}{2|A|} \ .$$

(b) Write down the measurement that optimally distinguishes ρ from τ in this case.

(c) Using a similar measurement, show that ρ and σ can be distinguished by a two-outcome measurement with probability $1 - \frac{1}{2}\langle\psi|\sigma|\psi\rangle$. Deduce that in this case when one of our states is pure we can obtain the following improvement on the Fuchs-van de Graaf lower bound:

$$1 - F(\rho, \sigma)^2 \leq T(\rho, \sigma) \ .$$

6.12 **Entanglement fidelity:** Consider a qubit system A and let $\mathcal{D}_p \in \mathrm{C}(A)$ and $\mathcal{P}_p$ be the depolarizing and dephasing channels with parameter $p \in [0, 1]$.

(a) Compute the entanglement fidelity $F(\mathcal{D}_p, \tau_A)$ for the maximally mixed state $\tau_A = \frac{\mathbb{1}_A}{2}$.

(b) Compute the entanglement fidelity $F(\mathcal{P}_p, \tau_A)$ for the maximally mixed state $\tau_A = \frac{\mathbb{1}_A}{2}$.

6.13 **Kraus representation and entanglement fidelity:** Prove Lemma 6.23.

6.14 **Continuity of Stinespring extensions:** Given two quantum channels $\Phi_1, \Phi_2 : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$, we can define their entanglement fidelity as

$$\begin{aligned} F_E(\Phi_1, \Phi_2) &:= F\big((\Phi_1 \otimes \mathcal{I}_A)(|\Omega\rangle\langle\Omega|_{AA'}), (\Phi_2 \otimes \mathcal{I}_A)(|\Omega\rangle\langle\Omega|_{AA'})\big) \\ &= F\big(\frac{1}{|A|}J(\Phi_1), \frac{1}{|A|}J(\Phi_2)\big) \ , \end{aligned}$$

where F is the usual fidelity between states, and $|\Omega\rangle_{AA'} = \frac{1}{\sqrt{|A|}}\sum_a |aa\rangle$ is the maximally entangled state between two copies of $\mathcal{H}_A$.

(a) For a quantum channel $\Phi : \mathrm{Lin}(A) \to \mathrm{Lin}(B)$, suppose the Choi operator $J(\Phi)$ has a (non-normalised!) purification $|\phi_{BAR}\rangle$ such that

$$J(\Phi) = \mathrm{tr}_R[|\phi_{BAR}\rangle\langle\phi_{BAR}|] \ .$$

Prove that this defines a Stinespring isometry V of Φ via

$$V : \mathcal{H}_A \to \mathcal{H}_B \otimes \mathcal{H}_R \ , \quad |a\rangle \mapsto (\mathbb{1}_B \otimes \langle a|_A \otimes \mathbb{1}_R)|\phi_{BAR}\rangle \ ,$$

where $\{|a\rangle\}_a$ is the basis of $\mathcal{H}_A$ used in the Choi operator.

Prove also that conversely any Stinespring isometry of Φ defines a purification of the Choi matrix.

(b) Let $|\phi^1_{BAR}\rangle$ and $|\phi^2_{BAR}\rangle$ be purifications of the Choi operators $J(\Phi_1)$ and $J(\Phi_2)$ respectively, corresponding to Stinespring isometries V_1 and V_2 as defined above. Show that

$$\langle\Omega|(\mathbb{1} \otimes V_1^\dagger)(\mathbb{1} \otimes V_2)|\Omega\rangle = \frac{1}{|A|}\langle\phi^1_{BAR}|\phi^2_{BAR}\rangle \ .$$

(c) Use Uhlmann's theorem to deduce that

$$F_E(\Phi_1, \Phi_2) = \mathrm{tr}[V_1^\dagger V_2] \ ,$$

for some Stinespring isometries $V_1, V_2 \in \mathrm{Isom}(A, BR)$ for Φ_1 and Φ_2.

(d) Deduce that if the channels Φ_1, Φ_2 are "close" in the sense that $F_E(\Phi_1, \Phi_2) \geq |A|(1-\epsilon)$, then they admit Stinespring isometries V_1, V_2 which are also "close" in the Schatten 2-norm:

$$\|V_1 - V_2\|_2^2 \leq 2\epsilon|A| \ .$$

6.15 **The diamond norm:** Suppose you are given a black-box quantum device by an experimentalist who refuses to tell you which channel it implements. Luckily the experimentalist only knows how to make two different channels $\Phi_{A\to B}$ and $\Psi_{A\to B}$, so you can assume that you have been given one of these uniformly at random.

You can attempt to distinguish between the channels in the following way: prepare a state ρ_{AR} in your system $\mathcal{H}_A \otimes \mathcal{H}_R$, apply the mysterious quantum device to the A system, and then use a two-outcome POVM on the resulting state in $S(BR)$ to try to work out which channel has been applied.

(a) Show that your optimal probability of success is given by

$$p_{\mathrm{opt}} = \frac{1}{2} + \frac{1}{4} \max_{\rho_{AR} \in S(AR)} \|\big((\Phi_{A\to B} - \Psi_{A\to B}) \otimes \mathcal{I}_R\big)(\rho_{AR})\|_1 \ .$$

(b) In fact this quantity is closely related to the *diamond norm* of a superoperator $\Lambda_{A\to B}$, which is defined as

$$\|\Lambda_{A\to B}\|_\diamond := \sup_n \max_{\rho_{AR} \in S(AR)} \|(\Lambda_{A\to B} \otimes \mathcal{I}_R)(\rho_{AR})\|_1 \ ,$$

where $|R| = n$.

Prove the following properties of the diamond norm:

i. The map $\|\cdot\|_\diamond \to \mathbb{R}$ actually defines a norm on the space of channels $\mathrm{Lin}(A) \to \mathrm{Lin}(B)$.

ii. For fixed n, maximum over ρ_{AR} is attained by a pure state.

iii. The supremum over n is attained by $n = |A|$. *Hint: reduce to pure states and use the Schmidt decomposition.*

Deduce that, assuming the auxiliary system R has $|R| \geq |A|$, the optimal success probability from the previous part can be written

$$p_{\text{opt}} = \frac{1}{2} + \frac{1}{4}\|\Phi_{A\to B} - \Psi_{A\to B}\|_\diamond \,.$$

(c) Prove that

$$\|\Lambda_{A\to B}\|_\diamond \leq \|J(\Phi)\|_1 \leq |A| \, \|\Lambda_{A\to B}\|_\diamond \,.$$

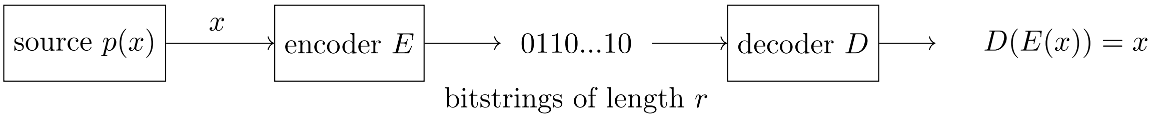
Hint: For the lower bound, first prove that you can write the maximising state $|\psi_{AR}\rangle$ *as* $(\mathbb{1}_A \otimes X_R)|\Omega_{AR}\rangle$*, where* $|\Omega_{AR}\rangle$ *is the maximally entangled state and* $\mathrm{tr}[X_R^\dagger X_R] = 1$. *Recall that* $\|MN\|_1 \leq \|M\|_1\|N\|_\infty$ *and argue that* $\|X_R\|_\infty \leq 1$.

Lecture 7

Compression

In these lectures we have so far not been very explicit about the meaning of the term *information*. As already alluded to in Lecture 1, one way to quantitatively define what information is, is by seeing how much a source can be *compressed*.

Let us first consider the case of classical compression. Suppose that we have a *source* generating symbols x according to a probability distribution p_X. How expensive is it to store x in memory? In other words, we would like to encode x into r bits in a way that we can recover the original symbol x:



The minimal r at which we can do this is a measure for the amount of information in the source p_X. We will start by defining this set-up in a more formal way in the classical case, and next we will formulate its quantum generalization where we try to compress a quantum source. In this chapter we will investigate compression in the case where we have a single symbol x (or a single quantum state in the quantum case), which is so called *one-shot compression.* In the next chapter we will see what happens if we have many samples from an IID source and study the asymptotic behaviour.

7.1 Classical compression

Suppose we have a classical source on a register X, with alphabet $\mathcal{X}$, and with probability distribution p_X. We model this as a random variable $\mathbf{X}$, which takes value x with probability $p_X(x)$. As alluded to in the above figure, the idea of compression is that there exists a classical *encoding* channel E to a *smaller* classical system C, representing the *compressed* information, and a classical *decoding* channel D back from C to X. We say that E and D form a zero-error r*-compression code* for p_X if the alphabet $\mathcal{C}$ of C is such that $\log(|\mathcal{C}|) \leq r$. You can think of the system C as r classical bits (i.e. bitstrings of length r).

For a good compression scheme, this should be such that while we reduce the size needed to store information, we can still recover the original content. There are two options for what we may demand of the recovery. A first option is to demand that we can always recover correctly, this is the zero-error scenario. In other words, we need that

$$D(E(\mathbf{X})) = \mathbf{X}$$

or in other words

$$D(E(x)) = x$$

for *all x such that* $p(x) \neq 0$. It is intuitively clear that in this scenario we can compress to a system C which has as size the number of x such that $p(x) \neq 0$. Note that E and D are classical channels, and in principle are defined as maps on probability distributions, but, in a small abuse of notation, we write $E(x)$ for the random variable we get when we apply E to the distribution which takes value x with probability 1.

Definition 7.1. If $p_X \in \mathrm{P}(X)$ then the *support* of $p_X \in \mathrm{P}(X)$ is given by

$$\mathrm{supp}(p_X) = \{x \in \mathcal{X} \text{ such that } p_X(x) > 0\}$$

and we let

$$H_0(p_X) = \log_2(|\mathrm{supp}(p_X)|)$$

be the *Rényi-0 entropy* of p_X.

From now on, we take the convention that logarithms are to base 2 (since we want to count in terms of bits) and write $\log = \log_2$. Clearly, $H_0(p_X)$ gives a first crude estimate of how much information a source p_X contains: it is simply the number of bits needed to describe all outcomes with nonzero chance of occurring. Formally we have the following result. As the notation suggests, $H_0(p_X)$ is a special case of a family of entropic quantities, which we will define later.

Lemma 7.2. *Suppose $p_X \in \mathrm{P}(X)$. Then there exists a zero-error r-compression code for p_X if and only if $r \geq H_0(p_X)$.*

Proof. If $r \geq H_0(p_X)$ then for $\mathcal{C} = \{0,1\}^r$ we have $|\mathrm{supp}(p_X)| \leq |\mathcal{C}|$ and we define the encoding channel by mapping each x such that $p(x) \neq 0$ to a different element in $\mathcal{C}$ and we map elements x with $p(x) = 0$ to some arbitrary element in $\mathcal{C}$. It is clear that this can be decoded correctly. On the other hand, if $r < H_0(p_X)$, let E, D be the encoder and decoder for a zero-error $r$compression code. Then $|\mathrm{supp}(p_X)| > |\mathcal{C}|$ and with nonzero probability E will map $x \neq y$ (with $x, y \in \mathrm{supp}(p_X)$) to the same element, and this implies that with nonzero probability either $D(E(x)) \neq x$ or $D(E(y)) \neq y$. □

There are situations where allowing a small probability of error makes a large difference in how well one can compress. For instance, consider a source which produces with probability $1-\varepsilon$ a fixed symbol 0, and with probability ε it produces a symbol from some large set Ω. Then, if ε is very small (for instance 10^{-20}), we would like to see that we can effectively completely compress the source to the single symbol 0 (so zero bits), rather than to $\log(|\Omega|+1)$ bits if we tolerate a very small error. In other words, we see that zero-error compression is very sensitive to small deformations of the source distribution. What is the optimal compression if you allow a small probability of error? The idea is simple: you correctly encode and decode the most likely symbols, and you allow error for a set of symbols that have total probability at most ε. Here is a concrete example.

Example 7.3. Consider a random variable $\mathbf{X}$, taking values $x = 1, \ldots, N+1$ where

$$p_X(x) = 2^{-x} \text{ for } x = 1, \ldots, N \quad \text{and} \quad p_X(N+1) = 2^{-N}.$$

If we do not allow any error, we need $\lceil \log(N) \rceil$ bits. However, if we allow a small probability of error, we may compress in the following way: for $x \leq K$ we express x using $\lceil \log(K) \rceil$ bits and decode accordingly. If $x > K$ we assign an error message (or any random bitstring). Then we see that the probability of error equals

$$\Pr(D(E(\mathbf{X})) \neq \mathbf{X}) = \sum_{x=K+1}^{N+1} p_X(x) = 2^{-K}.$$

In other words, if we allow $\varepsilon = 2^{-K}$ error, we need $\log(K) = \log(\log(\varepsilon^{-1}))$ bits. We conclude that we can compress $\mathbf{X}$ to

$$r = \left\lceil \log \log \left(\frac{1}{\varepsilon} \right) \right\rceil$$

bits with probability of error at most ε. This can be a huge saving! While N can be arbitrarily large, if we allow the truly microscopic probability of error 2^{-30} we only need $\lceil \log(30) \rceil = 5$ compressed bits!

This example captures the idea of *lossy compression*, where we allow some fixed probability of error in the decoding process. This captures formally in the following definition for compression into r bits with probability of error ε.

Definition 7.4. We define a *classical* (ε, r)*-compression code* of a source $\mathbf{X}$ with distribution $p_X \in \mathrm{P}(X)$ to r bits with error $\varepsilon > 0$ as a system C on an alphabet $\mathcal{C}$ of size $\log(|\mathcal{C}|) \leq r$ and a pair of classical channels

$$E : \mathrm{P}(X) \to \mathrm{P}(C), \qquad D : \mathrm{P}(C) \to \mathrm{P}(X)$$

which are such that

$$\Pr(D(E(\mathbf{X})) = \mathbf{X}) = \sum_x p_X(x) \Pr(D(E(x)) = x) \geq 1 - \varepsilon. \tag{7.1}$$

Then, we consider the best compression we can do if we allow error at most ε:

$$C^\varepsilon(p_X) = \min\{r : \text{there exists a } (\varepsilon, r)\text{-compression code for } p_X\}.$$

It is easy to see that there exist (ε, r)-compression codes for *all* $r \geq C^\varepsilon(p_X)$. For $\varepsilon = 0$ we found that $C^0(p_X) = H_0(p_X)$, which was the logarithm of the size of the support of p_X. The natural approach to lossy compression, as in Example 7.3, is to allow error for the outcomes with smallest probability. This corresponds to finding a distribution q_X which is ε-close to p_X but has as small support as possible.

Definition 7.5. The classical smooth Rényi-0 entropy of a probability distribution $p_X \in \mathrm{P}(X)$ is given by

$$H_0^\varepsilon(p_X) = \min_{q_X \in \mathrm{P}(X), T(p_X, q_X) \leq \varepsilon} H_0(q_X) \tag{7.2}$$

Such a construction (taking an information-theoretic quantity which is not necessarily continuous and optimizing it on a neighbourhood around it) is known as *smoothing*, and it yields a quantity which is by construction continuous. It gives an optimal lossy compression protocol. If $p_X, q_X \in \mathrm{P}(X)$, then by Lemma 6.7

$$T(p_X, q_X) = \frac{1}{2}\sum_x |p_X(x) - q_X(x)| = \sum_{p_X(x) > q_X(x)} p_X(x) - q_X(x) \tag{7.3}$$

Lemma 7.6. *Let $p_X \in \mathrm{P}(X)$, then*

$$H_0^\varepsilon(p_X) = \min\{\log(|\Omega|) : \Omega \textit{ such that } \sum_{x\in\Omega} p_X(x) \geq 1 - \varepsilon\} \tag{7.4}$$

Proof. If q_X is the minimizer in Eq. (7.2) and $\mathbf{X}$ is distributed according to p_X

$$\Pr(\mathbf{X} \notin \mathrm{supp}(q_X)) \leq \sum_{p_X(x) > q_X(x)} p_X(x) - q_X(x) = T(p_X, q_X) \leq \varepsilon.$$

On the other hand, if Ω is a minimizer in Eq. (7.4), let $q_X(x) = 0$ for $x \notin \Omega$ and for $x \in \Omega$

$$q_X(x) = \frac{p_X(x)}{\sum_{x\in\Omega} p_X(x)}.$$

Then $\mathrm{supp}(q_X) \subseteq \Omega$ and by Eq. (7.3)

$$T(p_X, q_X) = \sum_{x\notin\Omega} p_X(x) \leq \varepsilon.$$

□

Theorem 7.7 (Classical one-shot compression)**.** *Suppose $p_X \in \mathrm{P}(X)$ and $\varepsilon \geq 0$. Then there exists an (ε, r)-compression code for p_X if and only if $r \geq H_0^\varepsilon(p_X)$, so*

$$C^\varepsilon(p_X) = H_0^\varepsilon(p_X).$$

Proof. Suppose $H_0^\varepsilon(p_X) \leq r$, and let Ω be the minimizer in Eq. (7.2). Then an exact r-compression code compressing correctly on Ω has probability of error at most ε. Conversely, suppose that we have an (ε, r)-compression code for p_X with encoder E and decoder D. We may write the encoder and decoder as a convex combination of deterministic encoders and decoders. The probability of successful recovery is now the average over the success probabilities of each pair of deterministic encoders and decoders. Therefore, at least one pair of the deterministic encoder $\tilde{E}$ and decoder $\tilde{D}$ has recovery probability at least $1 - \varepsilon$. Let Ω be the set of size 2^r such that we have correct encoding and decoding. Then the probability that x is in Ω is at least $1 - \varepsilon$, so by Lemma 7.6 $H_0^\varepsilon(p_X) \leq r$. □

7.2 Quantum compression

Our discussion of classical compression was hopefully rather intuitive: we can perform lossy compression by discarding outcomes which together have probability at most ε if we allow probability of error ε. For the generalization to the quantum setting we will replace p_X by some quantum state $\rho_A \in \mathrm{S}(A)$. It is then natural to allow quantum channels as encoder $\mathcal{E} \in \mathrm{C}(A, C)$ and decoder $\mathcal{D} \in \mathrm{C}(C, A)$. Finally, for classical compression the condition

$$\Pr(D(E(\mathbf{X})) = \mathbf{X}) \geq 1 - \varepsilon$$

was a natural way to express that the protocol has small probability of error (or zero error if $\varepsilon = 0$).

The appropriate notion for quantum compression is *not* that we simply require

$$\mathcal{D}(\mathcal{E}(\rho_A)) \approx_\varepsilon \rho_A.$$

Indeed, this would be similar to only demanding in the classical case that p_X and $q_X = D(E(p_X))$ are close as distributions. However, in that case we could just 'compress' by discarding our initial x (so $r = 0$) and as decoder one would sample from the distribution p_X. In this case we have $q_X = p_X$ as distributions, but in general we do not have that $D(E(x)) = x$ with high probability. One way to interpret Eq. (7.1) is that it is equivalent to the condition that for *any* joint random variables $\mathbf{XY}$ on XY with distribution p_{XY} with marginal distribution of $\mathbf{X}$ equal to p_X, we find that if q_{XY} is the distribution we get by applying $D \circ E$ to $\mathbf{X}$, we have

$$T(p_{XY}, q_{XY}) \leq \varepsilon.$$

You will prove this in Exercise 7.8.

As was also discussed in Lecture 6, the correct way to measure whether a channel acts similar to the identity channel on some prescribed state is by allowing an arbitrary reference system on which we act by the identity channel, leading to the notion of the entanglement fidelity or the entanglement purified distance. In other words, the natural requirement for quantum compression is to demand that

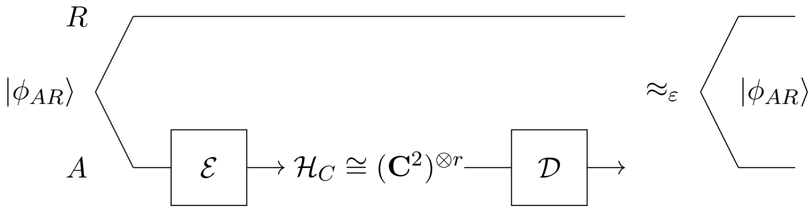
$$((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\sigma_{AR}) \approx \sigma_{AR}$$

for any state σ_{AR} extending ρ_A (so $\sigma_A = \rho_A$).

What is convenient about using the entanglement purified distance, which we saw as Definition 6.24, is that we only need to verify closeness on an arbitrary purification of ρ_A to verify that we have this for arbitrary reference systems:

$$P_E(\mathcal{D} \circ \mathcal{E}, \rho_A) = \sup_{\sigma_{AR}, \sigma_A = \rho_A} P(((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\sigma_{AR}), \sigma_{AR}) = P(((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\tau_{AR}), \tau_{AR})$$

for an arbitrary choice of purification $\tau_{AR} = |\phi_{AR}\rangle\langle\phi_{AR}|$ of ρ_A:



We summarize the above discussion in the following definition for quantum compression.

Definition 7.8. We define a *quantum* (ε, r)*-compression code* of a state $\rho_A \in \mathrm{S}(A)$ to r qubits with error $\varepsilon > 0$ as a system C with dimension $|C|$ such that $\log(|C|) \leq r$ and pair of quantum channels

$$\mathcal{E} \in \mathrm{C}(A, C), \qquad \mathcal{D} \in \mathrm{C}(C, A)$$

which are such that

$$P_E(\mathcal{D} \circ \mathcal{E}, \rho_A) \leq \varepsilon.$$

Similar to the classical case the optimal compression with error at most ε is given by

$$C^\varepsilon(\rho_A) = \min\{r : \text{there exists a } (\varepsilon, r)\text{-compression code for } \rho_A\}.$$

Definition 7.9. The quantum Rényi-0 entropy of $\rho_A \in \mathrm{S}(A)$ is given by

$$H_0(\rho_A) = \log(\mathrm{rank}(\rho_A))$$

For $0 < \varepsilon < 1$ the smooth Rényi-0 entropy is given by

$$H_0^\varepsilon(\rho_A) = \min_{\sigma_A \in \mathrm{S}(A), P(\rho_A, \sigma_A) \leq \varepsilon} H_0(\sigma_A).$$

Remark 7.10. Note that Definition 7.8 and Definition 7.9 do not reduce to the corresponding definitions for probability distributions. The reason is purely that we chose different distance measures. Otherwise, we see that if we let ρ_X be the classical density matrix corresponding to a probability distribution p_X, then $|\mathrm{supp}(p_X)| = \mathrm{rank}(\rho_X)$ and hence $H_0(p_X) = H_0(\rho_X)$. The motivation for our different choice of distance measures is that for probability distributions the trace distance is much more natural, while for the quantum case the purified distance is also natural and easier to work with. However, we could have used the trace distance for our quantum definitions as well! For small error ε we can use Eq. (6.9) to convert between the two distance measures at small cost. Another variation on Definition 7.9 is that one can generalize to *subnormalized states* (so $\mathrm{tr}[\rho_A] \leq 1$ instead of $\mathrm{tr}[\rho_A] = 1$). In certain cases this simplifies arguments (but we will not use this convention). However, it is good to be aware that there are different definitions in the literature of $H_0^\varepsilon(\rho_A)$, and that they can all be related and do not lead to fundamentally different notions.

To get a better idea of how this smoothing notion is related to the approach we took for classical compression, note that there we were essentially looking for a subset $\Omega \subset \mathcal{X}$ such that Ω was as small as possible, while $\Pr(X \in \Omega) \geq 1 - \varepsilon$. We replace this by looking for a projection operator Π_A which has as small as possible rank, but $\mathrm{tr}[\rho_A \Pi_A]$ is close to 1.

Lemma 7.11. *Let $\rho_A \in \mathrm{S}(A)$, let p_X be the probability distribution given by the spectrum of ρ_A.*

(a) We have $H_0^\varepsilon(\rho_A) \leq r$ if and only if there exists a projection operator $\Pi_A \in \mathrm{Lin}(A)$ of rank at most 2^r such that $\mathrm{tr}[\Pi_A \rho_A] \geq 1 - \varepsilon^2$.

(b) The quantum Rényi-0 entropy for ρ_A with smoothing ε equals the classical Rényi-0 entropy for its spectrum with smoothing $\delta = \varepsilon^2$:

$$H_0^\varepsilon(\rho_A) = H_0^\delta(p_X).$$

Proof. First suppose there exists a projection operator Π_A such that $\operatorname{tr}[\Pi_A \rho_A] \geq 1 - \varepsilon^2$. We may then take

$$\sigma_A = \frac{\Pi_A \rho_A \Pi_A}{\operatorname{tr}[\Pi_A \rho_A]}$$

which has rank at most the rank of Π_A. By the gentle measurement lemma, Lemma 6.20, $P(\rho_A, \sigma_A) \leq \sqrt{1 - \operatorname{tr}[\Pi_A \rho_A]} \leq \varepsilon$, and hence $H_0^\varepsilon(\rho_A) \leq r$. Conversely, let $r = H_0^\varepsilon(\rho_A)$. Let σ_A have rank 2^r and $F(\rho_A, \sigma_A)^2 \geq 1 - \varepsilon^2$. Choose purifications ϕ_{AR} and ψ_{AR} of ρ_A and σ_A as in Uhlmann's theorem and let Π_A be the projection onto the image of σ_A. By Cauchy-Schwartz

$$\begin{aligned} 1 - \varepsilon^2 \leq F(\rho_A, \sigma_A)^2 = |\langle \phi_{AR} | \psi_{AR} \rangle|^2 = |\langle \phi_{AR} | (\Pi_A \otimes \mathbb{1}_R) | \psi_{AR} \rangle|^2 \\ \leq \langle \phi_{AR} | (\Pi_A \otimes \mathbb{1}_R) | \phi_{AR} \rangle = \operatorname{tr}[\Pi_A \rho_A]. \end{aligned}$$

This proves (a).

For the second claim of the lemma, let

$$\rho_A = \sum_x^n p_x |\psi_x\rangle\langle\psi_x|$$

be a spectral decomposition. We claim that a projection Π_A with minimal rank such that $\operatorname{tr}[\Pi_A \rho_A] \geq 1 - \varepsilon^2$ can be taken to be of the form

$$\Pi_A = \sum_{x \in \Omega} |\psi_x\rangle\langle\psi_x|$$

for some set Ω, in which case $\operatorname{tr}[\Pi_A \rho_A] = \sum_{x \in \Omega} p_x$. You may prove this in Exercise 7.6. This implies that

$$\begin{aligned} H_0^\varepsilon(\rho_A) &= \min\{\log(\operatorname{rank}(\Pi_A) \ : \operatorname{tr}[\Pi_A \rho_A] \geq 1 - \varepsilon^2)\} \\ &= \min\{\log(|\Omega|) : \sum_{x \in \Omega} p_x \geq 1 - \varepsilon^2\} \end{aligned}$$

which equals $H_0^{\varepsilon^2}(p_X)$ by Lemma 7.6. □

We will now argue that we can perform zero-error r-compression (so $\varepsilon = 0$) for $r = H_0(\rho_A)$. The intuition is that this is similar to the classical case (where we restricted to the support of the distribution) and we 'restrict' to the subspace that is the image of ρ_A. Overall, the intuition behind one-shot quantum compression is straightforward. In the zero-error version, we can at most restrict to the subspace that is the image of ρ_A; if we allow some error then we look for a subspace such that projecting on this subspace does not change ρ too much. However, there will be some technicalities in making this precise and proving that such a procedure is optimal.

Lemma 7.12. *There exists a $(0, H_0(\rho_A))$-compression code for $\rho_A \in \mathrm{S}(A)$.*

Proof. Let $n = \operatorname{rank}(\rho_A) = 2^r$ and let $\mathcal{H}_C = \mathbb{C}^n$. Let

$$\rho_A = \sum_{i=0}^{n-1} p_i |\psi_i\rangle\langle\psi_i|$$

be a spectral decomposition of ρ_A. Let

$$V = \sum_{i=0}^{n-1} |\psi_i\rangle\langle i|$$

then $V \in \mathrm{Isom}(C, A)$ is an isometry. The idea is that we would like to simply restrict $\mathcal{H}_A$ to the image of ρ_A, and then $V^\dagger$ encodes this into $\mathbb{C}^n$, and V decodes back to $\mathcal{H}_A$. This would correspond to $\mathcal{E}(M_A) = V^\dagger M_A V$ and $\mathcal{D}(M_C) = V M_E V^\dagger$. This is almost correct, the only issue is that this way $\mathcal{E}$ is not trace-preserving. One way to fix this is the following: we choose an arbitrary $\tau_C \in \mathrm{S}(C)$, and we let

$$\begin{aligned}\mathcal{E}(M_A) &= V^\dagger M_A V + \mathrm{tr}[(\mathbb{1}_A - VV^\dagger)M_A]\tau_C \\ \mathcal{D}(M_C) &= V M_E V^\dagger.\end{aligned}$$

The second term in the definition of $\mathcal{E}$ can be interpreted as taking any states in the complement of the image of V and mapping them to a fixed state τ_C. It is clear that $\mathcal{D}$ defines a channel (as it is an isometric channel). It is also easy to see (try to find a Kraus representation!) that $\mathcal{E}$ is completely positive. Finally, we check that $\mathcal{E}$ is trace-preserving as well:

$$\begin{aligned}\mathrm{tr}[\mathcal{E}(M_A)] &= \mathrm{tr}[V^\dagger M_A V + \mathrm{tr}[(\mathbb{1}_A - VV^\dagger)M_A]\tau_C] \\ &= \mathrm{tr}[V^\dagger M_A V] + \mathrm{tr}[(\mathbb{1}_A - VV^\dagger)M_A] \\ &= \mathrm{tr}[VV^\dagger M_A + (\mathbb{1}_A - VV^\dagger)M_A] = \mathrm{tr}[M_A].\end{aligned}$$

Now, it is easy to verify that for a purification ρ_{AR} of ρ_A we have $((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\rho_{AR}) = \rho_{AR}$. □

From this lemma we see that we can compress exactly for $r = H_0(\rho_A)$. The idea behind the proof that we can not do any better is that if we take a purification ρ_{AR} of ρ_A it has *Schmidt rank* (which is defined as the number of terms in the Schmidt decomposition, or equivalently, $\mathrm{rank}(\rho_A)$), so $H_0(\rho_A)$ is a measure of how much entanglement there is between A and R. Applying the encoding and decoding to A should not increase the entanglement between A and R, but after encoding the entanglement is upper bounded by the dimension of C, suggesting that $|C| \geq \mathrm{rank}(\rho_A)$ and hence $r \geq H_0(\rho_A)$.

In order to make this intuition precise, also in the case of lossy compression, we generalize the idea of the Schmidt rank to mixed states.

Definition 7.13. If $\rho_{AB} \in \mathrm{S}(AB)$, then the entanglement rank of ρ_{AB} (between the subsystems A and B) is the minimum r such that ρ_{AB} can be written as

$$\rho_{AB} = \sum_i p_i |\phi_{AB,i}\rangle\langle\phi_{AB,i}|$$

where each $|\phi_{AB,i}\rangle$ has Schmidt rank at most r, so

$$\mathrm{rank}(\mathrm{tr}_B[|\phi_{AB,i}\rangle\langle\phi_{AB,i}|]) \leq r.$$

The key fact is that entanglement rank does not increase under separable channels.

Lemma 7.14. *Suppose $\Phi_{AB\to A'B'}$ is a separable channel, and ρ_{AB} has entanglement rank r. Then $\Phi_{AB\to A'B'}(\rho_{AB}) \in \mathrm{S}(A'B')$ has entanglement rank (between A' and B') at most r.*

The proof is Exercise 7.4. We now state and prove the quantum analog of Theorem 7.7

Theorem 7.15 (Quantum one-shot compression)**.** *We have*

$$H_0^{\varepsilon}(\rho_A) \leq C^{\varepsilon}(\rho_A) \leq H_0^{\frac{\varepsilon}{2}}(\rho_A).$$

Proof. We start by proving that $C^{\varepsilon}(\rho_A) \leq H_0^{\frac{\varepsilon}{2}}(\rho_A)$. By definition, we have $\sigma_A \in \mathrm{S}(A)$ such that $P(\rho_A, \sigma_A) \leq \frac{\varepsilon}{2}$ and $H_0^{\frac{\varepsilon}{2}}(\rho_A) = H_0(\sigma_A)$. Let $\mathcal{E} \in \mathrm{C}(A,C)$, $\mathcal{D} \in \mathrm{C}(C,A)$ be an exact $H_0(\sigma_A)$-compression code. Let R be a reference system and let ρ_{AR}, σ_{AR} be purifications of ρ_A and σ_A such that $P(\rho_A, \sigma_A) = P(\rho_{AR}, \rho_{AR})$, which exist by Uhlmann's theorem. Then by the triangle inequality for the purified distance

$$\begin{aligned} P_E(\mathcal{D} \circ \mathcal{E}, \rho_A) &= P(((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\rho_{AR}), \rho_{AR}) \\ &\leq P(((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\rho_{AR}), \sigma_{AR}) + P(\sigma_{AR}, \rho_{AR}) \end{aligned}$$

Now since we have an exact compression code for σ_A, we have

$$P(((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\rho_{AR}), \sigma_{AR}) = P(((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\rho_{AR}), ((\mathcal{D} \circ \mathcal{E}) \otimes \mathcal{I}_R)(\sigma_{AR})) \leq P(\rho_{AR}, \sigma_{AR})$$

using the monotonicity of the purified distance in the last inequality. In conclusion

$$P_E(\mathcal{D} \circ \mathcal{E}, \rho_A) \leq 2P(\rho_{AR}, \sigma_{AR}) \leq \varepsilon.$$

Conversely, suppose that we have an (ε, r)-compression code for ρ_A for some r with encoding and decoding quantum channels $\mathcal{E} \in \mathrm{C}(A,C)$ and $\mathcal{D} \in \mathrm{C}(C,A)$. Let $\rho_{AR} = |\phi_{AR}\rangle\langle\phi_{AR}|$ be a purification of ρ_A. Now let $\omega_{CR} = (\mathcal{E} \otimes \mathcal{I}_R)(\rho_{AR})$. It has entanglement rank between C and R at most $n = 2^r$. Therefore, by Lemma 7.14 the state $\sigma_{AR} = (\mathcal{D} \otimes \mathcal{I}_R)(\omega_{CR})$ has entanglement rank at most n. Let

$$\sigma_{AR} = \sum_i p_i |\psi_{AR,i}\rangle\langle\psi_{AR,i}|$$

where each $|\psi_{AR,i}\rangle$ is such that $\sigma_{A,i} = \mathrm{tr}_R[|\psi_{AR,i}\rangle\langle\psi_{AR,i}|]$ has rank at most n. Then

$$\begin{aligned} F(\rho_{AR}, \sigma_{AR}) &= \sqrt{\langle\phi_{AR}|\sigma_{AR}|\phi_{AR}\rangle} \\ &= \sqrt{\sum_i p_i \langle\phi_{AR}|\psi_{AR,i}\rangle\langle\psi_{AR,i}|\phi_{AR}\rangle} \\ &= \sqrt{\sum_i p_i |\langle\phi_{AR}|\psi_{AR,i}\rangle|^2} \\ &\leq \max_i |\langle\phi_{AR}|\psi_{AR,i}\rangle|. \end{aligned}$$

Let $|\psi_{AR,i}\rangle$ be the state which has maximal overlap with $|\phi_{AR}\rangle$ and let $\tau_{AR} = |\psi_{AR,i}\rangle\langle\psi_{AR,i}|$. Then $\mathrm{rank}(\tau_A) \leq n = 2^r$ and

$$P(\rho_A, \tau_A) \leq P(\rho_{AR}, \tau_{AR}) \leq P(\rho_{AR}, \sigma_{AR}) \leq \varepsilon.$$

We conclude that $H_0^{\varepsilon}(\rho_A) \leq r$ for any r such that we have an (ε, r)-compression code for ρ_A. The smallest such r equals $C^{\varepsilon}(\rho_A)$ by definition and hence $H_0^{\varepsilon}(\rho_A) \leq C^{\varepsilon}(\rho_A)$. □

7.3 Exercises

7.1 **Smooth entropy:** Let $\rho = 0.9|0\rangle\langle 0| + 0.09|1\rangle\langle 1| + 0.01|2\rangle\langle 2|$ be a state on $\mathbb{C}^3$.

(a) What is $H_0(\rho)$? What is $H_0(\rho^{\otimes n})$ for $n \geq 1$?
(b) What is $H^\varepsilon(\rho)$ for $\varepsilon = 0.02$?
(c) What is $H^\varepsilon(\rho^{\otimes 2})$ for $\varepsilon = 0.02$?

Hint: note that by Lemma 7.11 you only need to use diagonal states in the standard basis for the smoothing.

7.2 **Distance measures for product states:** Suppose $\rho_A, \sigma_A \in \mathrm{S}(A)$ and $\rho_B, \sigma_B \in \mathrm{S}(B)$.

(a) Show that $P(\rho_A \otimes \rho_B, \sigma_A \otimes \sigma_B) \leq P(\rho_A, \sigma_A) + P(\rho_B, \sigma_B)$.
(b) Show that $T(\rho_A \otimes \rho_B, \sigma_A \otimes \sigma_B) \leq T(\rho_A, \sigma_A) + T(\rho_B, \sigma_B)$.
(c) Show that $F(\rho_A \otimes \rho_B, \sigma_A \otimes \sigma_B) = F(\rho_A, \sigma_A)F(\rho_B, \sigma_B)$.

7.3 **Compressing multiple systems:** Let $\rho_{AB} \in \mathrm{S}(AB)$ with reduced states ρ_A and ρ_B. Suppose $\mathcal{E}_A, \mathcal{D}_A$ and $\mathcal{E}_B, \mathcal{D}_B$ are (ε, r_A)- and (ε, r_B)-compression codes for ρ_A and ρ_B respectively. Let $\mathcal{E} = \mathcal{E}_A \otimes \mathcal{E}_B$ and $\mathcal{D} = \mathcal{D}_A \otimes \mathcal{D}_B$.

(a) Show that if ρ_{ABS} is a purification of ρ_{AB}

$$\sigma_{ABS} = ((\mathcal{D}_A \circ \mathcal{E}_A) \otimes \mathcal{I}_{BS})(\rho_{ABS})$$

is such that $\sigma_B = \rho_B$.
(b) Show that

$$P((\mathcal{I}_A \otimes (\mathcal{D}_B \circ \mathcal{E}_B) \otimes \mathcal{I}_S)(\sigma_{ABS}), \sigma_{ABS}) \leq \varepsilon.$$

(c) Show that

$$P_E(\mathcal{D} \circ \mathcal{E}, \rho_{AB}) \leq P((\mathcal{I}_A \otimes (\mathcal{D}_B \circ \mathcal{E}_B) \otimes \mathcal{I}_S)(\sigma_{ABS}), \sigma_{ABS}) + P(\sigma_{ABS}, \rho_{ABS}).$$

(d) Conclude that

$$P_E(\mathcal{D} \circ \mathcal{E}, \rho_{AB}) \leq 2\varepsilon.$$

(e) Show that for any $\rho_{AB} \in \mathrm{S}(AB)$ with reduced states ρ_A and ρ_B

$$C^{2\varepsilon}(\rho_{AB}) \leq C^\varepsilon(\rho_A) + C^\varepsilon(\rho_B).$$

(f) Deduce that

$$H_0^{4\varepsilon}(\rho_{AB}) \leq H_0^\varepsilon(\rho_A) + H_0^\varepsilon(\rho_B). \tag{7.5}$$

7.4 **Entanglement rank:** Prove Lemma 7.14. *Hint: use the Kraus representation from Exercise 5.10.*

7.5 **Zero error quantum compression:** Verify that the channels $\mathcal{E}$ and $\mathcal{D}$ in Lemma 7.12 indeed are a zero error compression code for ρ_A.

7.6 **Smoothing quantum states and projections:** Let $\rho_A \in \mathrm{S}(A)$ have spectral decomposition

$$\rho_A = \sum_{i=1}^{n} p_i |\psi_i\rangle\langle\psi_i|$$

where $p_1 \geq p_2 \geq \dots$. The goal of this exercise is to show that if Π_A is a projection of rank r,

$$\operatorname{tr}[\Pi_A \rho_A] \leq \sum_{i=1}^{r} p_i. \tag{7.6}$$

This will confirm a claim in Lemma 7.11.

(a) Argue that if $V \subset \mathcal{H}_A$ is the subspace spanned by all eigenvectors $|\psi_i\rangle$ for $i \geq r$, and W is the image of Π_A, there is a nonzero vector $|v\rangle$ in the intersection $V \cap W$.

(b) Show that $\langle v|\rho_A|v\rangle \leq p_r$.

(c) Prove Eq. (7.6) by induction.

(d) Conclude that a projection with minimal rank such that $\operatorname{tr}[\Pi_A \rho_A] \geq 1 - \varepsilon^2$ can be taken of the form

$$\Pi_A = \sum_{i=1}^{r} |\psi_i\rangle\langle\psi_i|$$

for some r, confirming a claim made in the proof of Lemma 7.11.

7.7 **Coupling and trace distance:**

(a) Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are random variables with distributions p, q on a set Ω. Show that

$$T(p,q) = \max_{O \subset \Omega} \left(\Pr(\mathbf{X} \in O) - \Pr(\mathbf{Y} \in O) \right). \tag{7.7}$$

(b) Now suppose that $\mathbf{X}$ and $\mathbf{Y}$ are again random variables, but now having a joint distribution on $\Omega \times \Omega$, with marginal distributions p and q respectively. Such a joint probability distribution is called a *coupling* for p and q Show that

$$T(p,q) \leq \Pr(\mathbf{X} \neq \mathbf{Y}).$$

7.8 **Condition for classical compression:** Show that E and D are an (ε, r)-compression code for p_X, so

$$\Pr(D(E(\mathbf{X})) = \mathbf{X}) \geq 1 - \varepsilon$$

if and only if for any joint distribution p_{XY} for random variables $\mathbf{XY}$ which is such that $\mathbf{X}$ has marginal distribution p_X we have that applying $D \circ E$ to $\mathbf{X}$ gives a joint distribution q_{XY} with

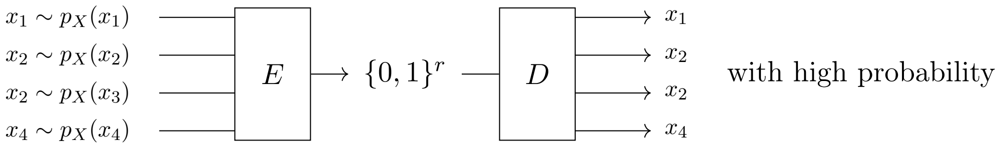
$$T(p_{XY}, q_{XY}) \leq \varepsilon.$$

Hint: use Exercise 7.7. Try the special case where $\mathbf{Y}$ *takes values in the same set as* $\mathbf{X}$ *and let* $p_{XY}(x, x') = p_X(x)\delta_{x,x'}$.

Lecture 8

Asymptotic compression and entropy

In the previous lecture we studied compression for a source which was either modelled by a probability distribution p_X or a quantum state ρ_A. A natural setting is where we have a source which produces many independent samples from the same probability distribution. We can then apply a compression code to n samples of the source (this is called *block coding* since we take blocks). So, the situation is as follows:



To formulate this more mathematically, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sequence of IID random variables with distribution p_X. We abbreviate

$$\mathbf{X}^n = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \quad \text{and} \quad x^n = (x_1, \ldots, x_n) \in \Omega^n.$$

We write p_{X^n} for the distribution of $\mathbf{X}^n$, which is given by

$$p_{X^n}(x^n) = p_X(x_1) \ldots p_X(x_n).$$

It turns out that block encoding can lead to large savings in the compression.

Example 8.1. Consider a binary random variable $\mathbf{X}$ which takes value 0 with probability 0.1 and value 1 with probability 0.9. Let us say we are willing to allow an error probability $\varepsilon = 0.05$. Then, if we get a single sample from this source we can not compress, and therefore, if we have many samples $\mathbf{X}^n$, we can not do any compression if we only allow codes that compress each $\mathbf{X}_i$ separately. However, if we look at (for instance) the distribution of $\mathbf{X}^3$ we see that we get

x^3	$p(x^3)$
000	0.001
$001, 010, 100$	0.009
$011, 101, 110$	0.081
111	0.729

Now we see that we can discard the outcomes $000, 001, 010$ and 100 to achieve error probability below 0.05 and hence we only need $\log(4) = 2$ bits instead of 3 bits!

To see how well block codes perform it is natural to see how many bits they require per source symbol, so if we compress a system X^n to $r(n)$ bits we study the *rate of compression* $r(n)/n$. How well we can compress depends on the error tolerance we allow. Of course, we would like the error to be as small as possible, and more strongly one would like to be able to pick ε *arbitrarily* small if n is large enough. We formalize this idea in the following definition.

Definition 8.2. The *optimal rate of compression* for a random variable $\mathbf{X}$ with probability distribution p_X is

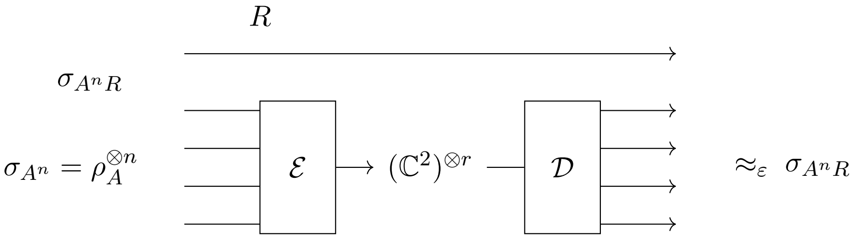
$$r(p_X) := \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}C^{\varepsilon}(p_{X^n}). \tag{8.1}$$

What does this mean more concretely? It means that $r(p_X)$ is the optimal value such that for any $\varepsilon, \delta > 0$ we can find some n_0 such that there exist block codes with error probability at most ε for $n > n_0$ which need at most $r(p_X) + \delta$ bits per symbol, i.e. $\frac{1}{n}C^{\varepsilon}(p_{X_n}) \leq r + \delta$. Note that the order of limits in Eq. (8.1) is important! If we take

$$\lim_{n\to\infty}\lim_{\varepsilon\to 0}\frac{1}{n}C^{\varepsilon}(p_{X^n})$$

we see that $\lim_{\varepsilon\to 0}\frac{1}{n}C^{\varepsilon}(p_{X^n}) = H_0(p_X)$ so we do not gain anything. In other words, in Eq. (8.1) the error probability goes to zero, but only as n goes to infinity.

The same discussion applies in the quantum setting. In that case we model a source producing many independent copies $\rho_A^{\otimes n}$ which we would like to compress simultaneously. That is, we have the following set-up, where σ_{A^nR} is a purification of $\rho_A^{\otimes n}$.


R
σ_{A^nR}
$\sigma_{A^n} = \rho_A^{\otimes n}$
$\mathcal{E}$
$(\mathbb{C}^2)^{\otimes r}$
$\mathcal{D}$
$\approx_\varepsilon\ \sigma_{A^nR}$

The optimal asymptotic rate of compression is then defined as follows.

Definition 8.3. The *optimal rate of compression* for a quantum state $\rho_A \in \mathrm{S}(A)$ is

$$r(\rho_A) := \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}C^{\varepsilon}(\rho_A^{\otimes n}).$$

From Theorem 7.7 and Theorem 7.15 we immediately find that

$$r(p_X) := \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}H_0^{\varepsilon}(p_{X^n})$$

and

$$r(\rho_A) := \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}H_0^{\varepsilon}(\rho_A^{\otimes n}).$$

We will see that the optimal asymptotic rate of compression $r(p_X)$ is given by the *Shannon entropy*, and in the quantum case $r(\rho_A)$ is given by the closely related *von Neumann entropy*. We will start by introducing these notions. Since we related compression for ρ_A to classical compression to its spectrum, we will see that the characterization of asymptotic quantum compression is a direct consequence of the characterization of asymptotic classical compression. So, it suffices to do a classical analysis; we already did all the hard work in the quantum case in the previous lecture!

The optimal rate of compression of a source (classical or quantum) is a good measure for the *amount of information* in the source. One way to understand this is the following scenario: Alice has a source described by a probability distribution p_X or quantum state ρ_A. She wants to send a realization of a stream of n outcomes from this source to Bob. The optimal rate of compression is the number of (qu)bits she has to send per copy in the limit of large n, which quantifies the amount of information in the source.

8.1 Classical and quantum entropy

We will see that the optimal rate of compression is given by the *entropy* of the source. Let us take a random variable $\mathbf{X}$ on a classical system X with probability distribution p_X. Recall that logarithms are to base 2. We will moreover take the convention that $0\log(0) = 0$ (note that $\lim_{x\downarrow 0} x\log(x) = 0$).

Definition 8.4. Let p_X be a probability distribution. Then the *Shannon entropy* of p_X is given by

$$H(p_X) = -\sum_x p_X(x)\log(p_X(x)).$$

We will also write $H(X)$ for $H(p_X)$.

What is the intuition behind this quantity? One can loosely think of it as a measure of how 'surprised' you will be when you receive a sample from the distribution p_X. For instance, if $p_X(x) = 1$ for some outcome x, then we surely obtain x and upon receiving the sample we learn nothing new (it is totally unsurprising). Indeed, we see that $H(p_X) = 0$ in this case. Conversely, if p_X is a uniform distribution, then any particular outcome is quite unlikely, so obtaining any particular outcome is rather surprising. In this case, if we have a uniform distribution on d outcomes

$$H(X) = -\sum_{x=0}^{d-1} \frac{1}{d}\log\left(\frac{1}{d}\right) = \log(d).$$

More generally, $-\log(p_X(x))$ is a measure of how 'surprising' an outcome is, or how much *uncertainty* p_X has.

Recall the notion of an *expectation value* (see Appendix B for details): if f is a function (taking values in $\mathbb{R}$), then

$$\mathbb{E}f(\mathbf{X}) = \sum_x p_X(x)f(x).$$

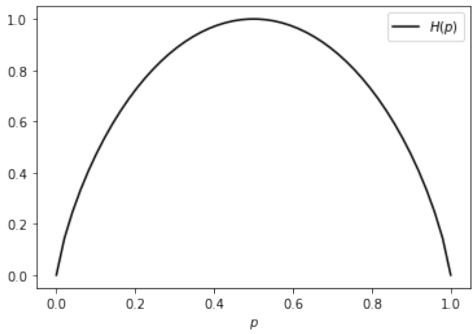
If $\mathbf{X}$ is a random variable with distribution p_X, and we let $p(\mathbf{X})$ denote the random variable taking value $p_X(x)$ with probability $p_X(x)$, then

$$H(X) = -\mathbb{E}\log(p(\mathbf{X})) = \mathbb{E}\log\left(\frac{1}{p(\mathbf{X})}\right).$$

The special case where we have the binary distribution $\{p, 1-p\}$ is also known as the *binary entropy function*

$$h(p) = -p\log(p) - (1-p)\log(1-p). \tag{8.2}$$

Here is a plot of what it looks like:



We see that it is zero when $p = 0$ or $p = 1$ and is maximal (and equal to 1) at $p = \frac{1}{2}$.

This is a special case of the following general bounds, which can be proven using Jensen's inequality. Recall that Jensen's inequality, Lemma B.2, states that if f is a convex function on a convex set I and $\mathbf{X}$ is a random variable taking values in I,

$$\mathbb{E}(f(\mathbf{X})) \geq f(\mathbb{E}\mathbf{X}).$$

If f is strictly convex, we have equality if and only if $\mathbf{X}$ is deterministic (takes a single value with probability 1). If f is concave (meaning $-f$ is convex), we have $\mathbb{E}(f(\mathbf{X})) \leq f(\mathbb{E}\mathbf{X})$.

Lemma 8.5. *Suppose $p_X \in \mathrm{P}(X)$.*

(a) $H(X) \geq 0$ and $H(X) = 0$ if and only if $\mathbf{X}$ is deterministic.

(b) $H(X) \leq H_0(p_X) \leq \log(|X|)$ and $H(X) = \log(|X|)$ if only if p_X is the uniform distribution on all outcomes.

Proof. For $x \in [0,1]$ we have $-x\log(x) \geq 0$ with equality if and only if $x \in \{0,1\}$. This implies that $H(X) \geq 0$, with equality if and only if $p_X(x)$ only takes values 0 or 1, which is only the case if $\mathbf{X}$ is deterministic. Next, we apply Jensen's inequality to the strictly concave function $x \mapsto \log(x)$

$$H(X) = \mathbb{E}\log\left(\frac{1}{p(\mathbf{X})}\right) \leq \log\left(\mathbb{E}\frac{1}{p(\mathbf{X})}\right)$$

Now

$$\mathbb{E}\frac{1}{p(\mathbf{X})} = \sum_x p_X(x)\frac{1}{p_X(x)} = \sum_{x \in \mathrm{supp}(p_X)} 1 = |\mathrm{supp}(p_X)|$$

so $H(X) \leq \log(|\mathrm{supp}(p_X)|) = H_0(p_X)$. We have equality if and only if $\frac{1}{p(\mathbf{X})}$ is deterministic, meaning that $p_X(x)$ is equal for all outcomes x. □

Another fact which can be proven using Jensen's inequality is the concavity of the Shannon entropy. That is, if $p_X^{(0)}$ and $p_X^{(1)}$ are probability distributions and $q \in [0,1]$, we may define the mixture $p_X^{(q)}$ of $p_X^{(0)}$ and $p_X^{(1)}$ as taking value x with probability

$$p_X^{(q)}(x) = qp_X^{(0)}(x) + (1-q)p_X^{(1)}(x).$$

Strict concavity of the Shannon entropy means that

$$H\left(p_X^{(q)}\right) \geq qH\left(p_X^{(0)}\right) + (1-q)H\left(p_X^{(1)}\right). \tag{8.3}$$

You may verify this in Exercise 8.3. This corresponds to the following intuition: mixing two probability distributions increases uncertainty. Next lecture we will see more properties of the entropy function!

We continue to the quantum version, which is known as the *von Neumann entropy* and which is simply the Shannon entropy of the spectrum: if a state $\rho_A \in \mathrm{S}(A)$ has spectral decomposition

$$\rho_A = \sum_x p_x |\psi_x\rangle\langle\psi_x|$$

then we let $H(\rho_A) = -\sum_x p_x \log(p_x)$.

Definition 8.6. The *von Neumann entropy* of a state $\rho_A \in \mathrm{S}(A)$ is given by

$$H(\rho_A) = -\operatorname{tr}[\rho_A \log \rho_A].$$

We will also write $H(A)_\rho = H(\rho_A)$ or just $H(A)$ if there can be no confusion about the state.

Note that $\log(\rho_A)$ need not be well-defined if ρ_A does not have full rank. However, with the convention $0\log(0) = 0$ the operator $\rho_A \log(\rho_A)$ is well-defined. It follows directly from Lemma 8.5, as you may verify in Exercise 8.3, that we have

Lemma 8.7. *Suppose $\rho_A \in \mathrm{S}(A)$. Then*

(a) $H(\rho_A) \geq 0$ and $H(A) = 0$ if and only if ρ_A is pure.

(b) $H(\rho_A) \leq H_0(\rho_A) \leq \log(d_A)$, and $H(\rho_A) = \log(d_A)$ if and only if ρ_A is the maximally mixed state.

We also observe that the entropy is invariant under isometries, i.e. if $\sigma = V\rho V^\dagger$ for an isometry V, then $H(\rho) = H(\sigma)$, since the von Neumann entropy only depends on the nonzero part of the spectrum. A final comment is that the function $x \mapsto x\log(x)$ is continuous. This implies that both the Shannon entropy and the von Neumann entropy are continuous functions (on the sets of probability distributions $\Pr(X)$ and states $\mathrm{S}(A)$ respectively). We will see sharp continuity estimates later.

8.2 Typical sets and subspaces

Let us return to the situation where we have random variable $\mathbf{X}$ with distribution p_X, and $\mathbf{X}^n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is a sequence of IID random variables with distributions p_X. We will introduce the notion of a *typical set* which are subsets of outcomes which are a relatively small subset of Ω^n, but at the same time has a high likelihood: as the name suggests they are sets of *typical outcomes*. For instance, if the random variable is a biased coin flip with probability of heads $\frac{1}{3}$, and you perform 1000 independent coin flips, then it is reasonable to expect that the outcome will have between, say, 200 - 400 times heads (i.e. that happens with high probability). On the other hand, this set is much smaller than the full set of outcomes. This of course relates back to compression in the following way: if we compress to this subspace then the probability of error is small. In other words, typical subsets give a way to estimate $H_0^\varepsilon(p_X)$. The way we will define the typical set is as follows. If we have an outcome x^n then

$$\frac{1}{n}\log\left(\frac{1}{p_{X^n}(x^n)}\right) = \frac{1}{n}\log\left(\frac{1}{p_X(x_1)}\cdots\frac{1}{p_X(x_n)}\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{p_X(x_i)}\right)$$

and for large n one expects that with high probability we will find outcomes such that this is close to the expectation value $-\mathbb{E}\log(p(\mathbf{X})) = H(X)$.

Definition 8.8. If $\mathbf{X}^n$ is an IID sequence of random variables with distribution p_X and $\varepsilon > 0$, then we define the *typical set* $T_{n,\varepsilon}(p_X)$ as

$$T_{n,\varepsilon}(p_X) = \{x^n = (x_1,\ldots,x_n) \in \Omega_X^n : \left|\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{p(x_i)}\right) - H(X)\right| \leq \varepsilon\}.$$

The following lemma shows that typical sequences are very likely to occur and bounds the size of the typical set. It shows that if $H(X) < \log(|X|)$, the typical set is exponentially smaller than the full outcome space. It is known as the asymptotic equipartition property, since it can be interpreted as saying that in the asymptotic setting the distribution of $p(x^n)$ concentrates on a set of outcomes with approximately equal probability.

Lemma 8.9 (Asymptotic equipartition property)**.** *The typical set $T_{n,\varepsilon}(p_X)$ has the following properties:*

(a) for $x^n \in T_{n,\varepsilon}(p_X)$, the probability of x_n is bounded by

$$2^{-n(H(X)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\varepsilon)}.$$

(b) The typical set has size bounded by

$$|T_{n,\varepsilon}(p_X)| \leq 2^{n(H(X)+\varepsilon)}.$$

(c) We have

$$\lim_{n\to\infty}\Pr(\mathbf{X}^n \in T_{n,\varepsilon}(p_X)) = 1.$$

Proof. (a) This is a direct consequence of the definition.

(b) Note that by (a)

$$1 \geq \Pr(\mathbf{X}^n \in T_{n,\varepsilon}(p_X)) = \sum_{x^n\in T_{n,\varepsilon}} p(x^n) \geq |T_{n,\varepsilon}(p_X)|2^{-n(H(X)+\varepsilon)}$$

and therefore

$$|T_{n,\varepsilon}(p_X)| \leq 2^{n(H(X)+\varepsilon)}.$$

(c) Consider the random variables $\mathbf{Y}_i = -\log(p(\mathbf{X}_i))$. By construction, $\mathbb{E}\mathbf{Y}_i = H(X)$. Then the law of large numbers, Theorem B.4, implies that

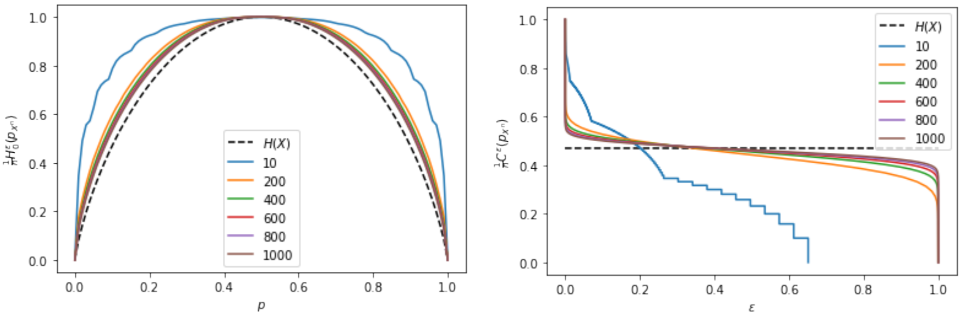
$$\lim_{n\to\infty}\Pr\left(|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i - H(X)| \geq \varepsilon\right) \to 0.$$

This implies the result (since $|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i - H(X)| \leq \varepsilon$ if and only if $\mathbf{X}^n \in T_{n,\varepsilon}(p_X)$ by definition). In case you have not seen the law of large numbers before, you are recommended to prove it for yourself in Exercise B.2!

□

8.3 Asymptotic compression

We will use typical sets to show that the optimal rate of compression is given by the entropy. To illustrate this phenomenon, we consider a binary random variable with distribution $\{p, 1-p\}$. Below we plot $\frac{1}{n}H_0^\varepsilon(X^n)$ for fixed $\varepsilon = 0.01$. We see that for large n we get closer and closer to $H(X)$ (although the convergence is pretty slow). We also plotted $H_0^\varepsilon(X^n)$ as a function of ε for fixed $p = 0.1$.



Let us now prove the phenomenon we observe in these figures!

Theorem 8.10 (Shannon's theorem)**.** *Let p_{X^n} be the distribution of $\mathbf{X}^n$, then the optimal asymptotic rate of compression is given by*

$$r(p_X) = H(X).$$

Proof. We have

$$r(p_X) = \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}H_0^\varepsilon(p_{X^n})$$

and the proof comes down to an estimate of $H_0^\varepsilon(p_X)$. Let $\delta > 0$ be arbitrary. We will show that there exists $\varepsilon(n) > 0$, which are such that $\varepsilon(n)$ goes to zero as n goes to infinity, for which

$$\frac{1}{n}H_0^{\varepsilon(n)}(p_{X^n}) \leq H(p_X) + \delta. \tag{8.4}$$

On the other hand, we will show that for any sequence $\varepsilon(n)$ going to zero as n goes to infinity

$$\frac{1}{n}H_0^{\varepsilon(n)}(p_{X^n}) \geq H(p_X) - \delta + p(n) \tag{8.5}$$

where $p(n)$ goes to zero as n goes to infinity. Together, these two bounds prove that

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}H_0^\varepsilon(p_{X^n}) = H(p_X).$$

For the first bound we fix $\delta > 0$ and define the sequence $\varepsilon(n)$ by

$$\varepsilon(n) := \Pr(\mathbf{X}^n \notin T_{n,\delta}(p_X)) \to 0$$

as n goes to infinity (using Lemma 8.9). This means that

$$\begin{aligned}\frac{1}{n} H_0^{\varepsilon(n)}(p_{X^n}) &\leq \frac{1}{n}\log(|T_{n,\delta}|)\\ &\leq H(p_X)+\delta\end{aligned}$$

using Lemma 7.6 in the first inequality and Lemma 8.9 in the second inequality. This proves Eq. (8.4).

Conversely, choose some sequence $\varepsilon(n)$ such that $\varepsilon(n)$ goes to zero as n goes to infinity and choose a set Ω_n as in Lemma 7.6 such that $H_0^{\varepsilon(n)}(p_{X^n}) = \log(|\Omega_n|)$. Again, we choose arbitrary $\delta > 0$. The idea is that Ω_n will need to contain a large subset of the typical subset $T_{n,\delta}(p_X)$. By the union bound (using that Ω_n is contained in the union of $\Omega_n \cap T_{n,\delta(p_X)}$ and the complement of $T_{n,\delta}(p_X)$)

$$1-\varepsilon(n) \leq \Pr(\mathbf{X}^n \in \Omega_n) \leq \Pr(\mathbf{X}^n \in \Omega_n \cap T_{n,\delta}(p_X)) + \Pr(\mathbf{X}^n \notin T_{n,\delta}(p_X)).$$

The first term may be estimated as

$$\sum_{x^n \in \Omega_n \cap T_{n,\delta}(p_X)} p_{X^n}(x^n) \leq |\Omega_n \cap T_{n,\delta}(p_X)| 2^{-n(H(X)-\delta)} \leq |\Omega_n| 2^{-n(H(X)-\delta)}$$

by Lemma 8.9. This implies

$$|\Omega_n| \geq (1-\varepsilon(n) - \Pr(\mathbf{X}^n \notin T_{n,\delta}(p_X)))2^{n(H(X)-\delta)}$$

Therefore,

$$\frac{1}{n} H_0^{\varepsilon}(p_X) \geq H(X) - \delta + \underbrace{\frac{1}{n}\log(1-\varepsilon(n) - \Pr(\mathbf{X}^n \notin T_{n,\delta}(p_X)))}_{:=p(n)}$$

The term $p(n)$ goes to zero as n goes to infinity by Lemma 8.9, proving Eq. (8.5). □

The typical set is the main proof tool for Theorem 8.10. However, it is generally not the *optimal* set to compress to, but the difference is sufficiently small to be irrelevant in the asymptotic setting. Its relevance is purely that it is convenient for the analysis, but we could have used different sets as well!

In the quantum case we get *Schumacher's theorem*, whose statement and proof are closely analogous to Theorem 8.10. It is a direct consequence of Theorem 8.10 and Lemma 7.11.

Theorem 8.11 (Schumacher's theorem). *The optimal asymptotic rate of compressing a quantum state $\rho_A \in \mathrm{S}(A)$ is given by*

$$r(\rho_A) = H(A)_\rho$$

Theorem 8.10 and Theorem 8.11 show that the entropy is a measure for the amount of information in a source, as it equals the optimal rate of compression of the source.

8.3.1 The typical subspace

While we do not need it at this point, we introduce the quantum version of the typical subset. This will now be a *subspace* rather than a subset. If $\rho_A \in \mathrm{S}(A)$ has spectral decomposition

$$\rho_A = \sum_x p(x) |\psi_x\rangle\langle\psi_x|$$

then the state $\rho_A^{\otimes n}$ has spectral decomposition

$$\rho_A^{\otimes n} = \sum_{x_1,\dots,x_n} p(x_1)\dots p(x_n) |\psi_{x_1}\rangle\langle\psi_{x_1}| \otimes \dots \otimes |\psi_{x_n}\rangle\langle\psi_{x_n}|.$$

The idea is to define the typical subspace $S_{n,\varepsilon}(\rho_A)$ for n copies of ρ_A and $\varepsilon > 0$, by restricting to the subspace spanned by all $|\psi_{x_1}\rangle \dots |\psi_{x_n}\rangle$ such that $x^n \in T_{n,\varepsilon}(p)$:

$$S_{n,\varepsilon}(\rho_A) = \mathrm{span}\{|\psi_{x_1}\rangle \dots |\psi_{x_n}\rangle \in \mathcal{H}_A^{\otimes n} : x^n = (x_1, \dots, x_n) \in T_{n,\varepsilon}(p)\}.$$

The *typical projector* is the projection operator $\Pi_{n,\varepsilon}$ onto $S_{n,\varepsilon}(\rho_A)$. The results of Lemma 8.9 translate to the typical subspace:

Lemma 8.12. *Let $\rho_A \in \mathrm{S}(A)$ and $\varepsilon > 0$.*

(a) The nonzero eigenvalues of $\Pi_{n,\varepsilon}\rho_A^{\otimes n} = \Pi_{n,\varepsilon}\rho_A^{\otimes n}\Pi_{n,\varepsilon}$ all lie in the interval

$$[2^{-n(H(A)+\varepsilon)}, 2^{-n(H(A)-\varepsilon)}].$$

(b) The dimension of the typical subspace is bounded as

$$\dim(S_{n,\varepsilon}(\rho_A)) \leq 2^{n(H(A)+\varepsilon)}.$$

(c) As n goes to infinity, if we measure whether we are in the typical subspace or not (corresponding to the measurement $\{\Pi_{n,\varepsilon}, \mathbb{1}_{A^n} - \Pi_{n,\varepsilon}\}$), the probability of being in the typical subspace goes to 1:

$$\lim_{n\to\infty} \mathrm{tr}[\Pi_{n,\varepsilon}\rho_A^{\otimes n}] = 1.$$

Outlook

The classical theory of information was developed by Shannon, in his 1948 paper *A mathematical theory of communication* [40], an all-time scientific masterpiece. It is recommended reading! The idea to develop compression for quantum systems came much later, and was first developed in [39].

Classical compression in practice

We have seen how to perform theoretically optimal compression for sources modelled by random variables, both in the one-shot and asymptotic settings. While doing so we completely ignored practical aspects. When you compress a file on your computer, say to a `zip`-file, it certainly does not take the approach described in these lectures! There are multiple reasons for this. First and foremost, the approaches we describe are computationally inefficient. In the one-shot compression scenario, our optimal compression protocols require an enumeration of the most likely outcomes. If the data concerned is for example a video file, we know that typical video files will have some structure (they are not just white noise) but somehow 'enumerating' all likely videos is not feasible. In the asymptotic scenario there is a bit more structure, and one can compress (for example) to the typical set (which would be asymptotically optimal). Here one could try to make an efficient enumerator for the typical set. One aspect is that for good performance one has to use very large blocks of data (i.e. compress n symbols at a time) if one wants small probability of error. This is not very flexible, and again causes large computational overhead. Another disadvantage is that the encoding and decoding depend heavily on the source distribution. In conclusion, the main point of Theorem 8.10 is that it shows the theoretically optimal rate of compression.

There are a number of elegant practical approaches which achieve good compression while overcoming the above objections. Typical practical compression algorithms are *variable-length* compression schemes. Here, one does not allow a probability of error, but rather one does not fix the number of bits r that the algorithm compresses to. A good compression algorithm is then such that the *expected* number of compressed bits is small, but with small probability the encoded message actually becomes *longer* than the original message. Two major classes of compression algorithms are *symbol codes* and *streaming codes*. A symbol code takes the set of symbols $\mathcal{X}$ and encodes each x to a bit string of variable length such that likely symbols are encoded as shorter bitstrings than less likely outcomes. One does this in such a way that when one concatenates the encodings for a multi-symbol source $x^n = (x_1, \dots, x_n)$ the result is uniquely decodeable. One can show that in this way one can achieve compression at a rate between $H(X)$ and $H(X) + 1$. Symbol codes are easy to implement and efficient. A second broad approach are streaming codes, which take in a stream of symbols (rather than blocks as in our theoretical approach to compression). In this scenario one can also model sources which are not IID, but where x_n depends in some way on the previous symbols x^{n-1}. Examples of such compression algorithms are arithmetic codes and the Lempel-Ziv compression algorithm. An explanation of these algorithms can be found in [29].

One-shot information theory and asymptotics

The results of this and the previous lecture are a paradigmatic example of a pattern in information theory. Often we have the following situation in information theory (either classical or quantum). We start with the formulation of some specific task or protocol in which you would like to find the optimal value (for us this was compression, to a minimal number of (qu)bits). Then, there is the *one-shot* version of the task, where one has to perform the task a single time. One can either study the case where the protocol has to be exact (zero-error compression) or a lossy version where a small error is allowed. Then, one can also study the *asymptotics* of the same task, where one receives many independent copies of the same source, and you look for the optimal *rate* at which you can perform the task. In our discussion of compression something very remarkable happened: we start with a complicated quantity, $H_0^\varepsilon(\rho_A)$ which was formulated as an optimization problem. By taking the asymptotic limit we found the much nicer quantity $H(A)$ which no longer requires solving an optimization problem but just has an explicit formula!

In many situations, something similar happens, where the asymptotic rate is an entropic quantity (we will see more examples of various entropic quantities in a later lecture), whereas the one-shot quantities depend in more subtle ways on the specific task and are formulated as an optimization problem and are harder to compute. However, there are also situations where we understand the one-shot quantity relatively well, but the asymptotic quantity does not significantly simplify and is hard to compute. A good introduction to one-shot quantum information theory is [42].

8.4 Exercises

8.1 **Entropy and typical sets:**

(a) Let $\mathbf{X}$ be a random variable on $\{0,1\}$ with distribution function

$$p_X(0) = 0.8 \ , \quad p_X(1) = 0.2 \ .$$

Compute the entropy $H(p_X)$.

(b) What are the elements of the typical set $T_{5,0.1}(p_X)$?

(c) Compute $P(\mathbf{X}^5 \in T_{5,0.1}(p_X))$.

8.2 **A source with infinite outcomes:** Consider the random variable in Example 7.3, which takes value x with probability 2^{-x} for $x = 1, \ldots, N$ and takes value $N+1$ with probability 2^{-N}.

(a) Write down a formula for the entropy $H(X)$ of this random variable, and compute its value for $N = 1, 2, 3$.

(b) Use the fact that $\sum_{x=1}^{\infty} x 2^{-x} = 2$ to argue that as $N \to \infty$ we have $H(X) = 2$.

8.3 **Bounds on the entropy:**

(a) Prove Lemma 8.7.

(b) Show concavity of the Shannon entropy, as in Eq. (8.3). *Hint: show that the function defined by $g(x) = -x\log(x)$ is concave on $\mathbb{R}_{\geq 0}$.*

8.4 **Entropy under functions:** Suppose $\mathbf{X}$ is a random variable on $\{0,\ldots,d-1\}$ with a probability distribution $p_X(x)$ such that $p_X(x) \neq 0$ for all $x \in \{0,\ldots,d-1\}$.

(a) Let $d' \in \mathbb{N}$, and define a function $f : \{0,\ldots,d-1\} \to \{0,\ldots,d'-1\}$. Show that

$$H(\mathbf{X}) \geq H(f(\mathbf{X}))$$

with equality if and only if f is injective.

(b) Let $\rho \in S(A)$ be a state and $U \in U(A)$ be a unitary. Show that the von Neumann entropy satisfies

$$H(\rho) = H(U\rho U^\dagger) \ .$$

Does this hold for general quantum channels?

8.5 **Typical subspaces:** If $\rho_A \in \mathrm{S}(A)$ has spectral decomposition

$$\rho_A = \sum_x p(x) |\psi_x\rangle\langle\psi_x|$$

and $\varepsilon > 0$, recall that we defined the typical subspace $S_{n,\varepsilon}(\rho_A)$ as

$$S_{n,\varepsilon}(\rho_A) = \mathrm{span}\{|\psi_{x_1}\rangle \ldots |\psi_{x_n}\rangle \in \mathcal{H}_A^{\otimes n} : x^n = (x_1,\ldots,x_n) \in T_{n,\varepsilon}(p)\}.$$

The *typical projector* is the projection operator $\Pi_{n,\varepsilon}$ onto $S_{n,\varepsilon}(\rho_A)$. The goal of this exercise is to translate the results of Lemma 8.9 to the typical subspace by proving Lemma 8.12.

(a) Show that the nonzero eigenvalues of $\Pi_{n,\varepsilon}\rho_A^{\otimes n} = \Pi_{n,\varepsilon}\rho_A^{\otimes n}\Pi_{n,\varepsilon}$ all lie in the interval $[2^{-n(H(A)+\varepsilon)}, 2^{-n(H(A)-\varepsilon)}]$.

(b) Prove that the dimension of the typical subspace is bounded as

$$\dim(S_{n,\varepsilon}(\rho_A)) \leq 2^{n(H(A)+\varepsilon)}.$$

(c) Show that

$$\lim_{n\to\infty} \mathrm{tr}[\Pi_{n,\varepsilon}\rho_A^{\otimes n}] = 1.$$

8.6 **Lower bound size typical set:** Show that for any $\delta > 0$ there exists n_0 such that for $n \geq n_0$

$$|T_{n,\varepsilon}(p_X)| \geq (1-\delta)2^{n(H(p_X)-\varepsilon)}.$$

8.7 **Leading corrections to asymptotic compression:** We have seen that as n goes to infinity we can let $\varepsilon(n)$ go to zero such that

$$C^{\varepsilon}(p_{X^n}) = nH(X) + f(n, \varepsilon(n)), \qquad \lim_{n\to\infty} \frac{1}{n} f(n, \varepsilon(n)) = 0.$$

We would like to understand the behavior of $f(n)$ in more detail. We will see that the central limit theorem allows us to determine the leading term to be $f(n) \sim \sqrt{n}$.

(a) Given IID $\mathbf{X}_1, \ldots, \mathbf{X}_n$, consider the random variables $\mathbf{Y}_i = -\log(p(\mathbf{X}_i))$. Apply the central limit theorem to show that for $z \in \mathbb{R}$

$$\Pr\left(\sum_{i=1}^{n} \mathbf{Y}_i - H(X) \leq z\sqrt{n}\right) = F(z) + g(n)$$

where $F(z)$ is the cumulative distribution function of a normal distribution with mean zero and variance

$$\sigma^2 = \sum_x p_X(x)(H(X) + \log(p_X))^2$$

and where $|g(n)| = \mathcal{O}(n^{-\frac{1}{2}})$.

(b) Let

$$\Omega_{n,\delta} = \{x^n : p_{X^n}(x^n) \geq 2^{-nH(X)-\sigma\delta\sqrt{n}}\}.$$

Show that

$$\Pr(\mathbf{X}^n \in \Omega_{n,\delta}) = F(\delta) + g(n).$$

(c) Show that $\log(|\Omega_{n,\delta}|) \leq nH(X) + \sigma\delta\sqrt{n}$.

(d) Show that for any fixed $\varepsilon > 0$, we have

$$H_0^{\varepsilon}(p_{X^n}) \leq nH(X) + \sqrt{n}\sigma\gamma(\varepsilon, n).$$

where

$$F^{-1}(\varepsilon + |g(n)|) \leq \gamma(\varepsilon, n) \leq F^{-1}(\varepsilon - |g(n)|)$$

Hint: note that $F(x) = 1 - F(-x)$.

(e) Argue that as n goes to infinity

$$\sqrt{n}\gamma(\varepsilon, n) = \sqrt{n}F^{-1}(\varepsilon) + \mathcal{O}(\log(n)).$$

Hint: Taylor expansion.

(f) Conclude that

$$C^{\varepsilon}(p_{X^n}) = nH(X) + \sqrt{n}\sigma F^{-1}(\varepsilon) + \mathcal{O}(\log(n)).$$

In particular, the leading order corrections are of the order $\sqrt{n}$.

8.8 **Entanglement distillation:** Suppose Alice and Bob, distantly separated, share n copies of a pure two qubit state $|\phi_{AB}\rangle \in \mathbb{C}^2 \otimes \mathbb{C}^2$, i.e. they have the state $\rho_{AB} = |\phi_{AB}\rangle\langle\phi_{AB}|^{\otimes n}$. They know the Schmidt decomposition of $|\phi_{AB}\rangle$, which is

$$|\phi_{AB}\rangle = \sqrt{p}|0_A\rangle|0_B\rangle + \sqrt{1-p}|1_A\rangle|1_B\rangle \ ,$$

for some $p \in [0, 1]$ where $p \neq 1/2$. Using only local operations, Alice and Bob want to manufacture an r-dimensional maximally entangled state $|\Phi^+_{r,AB}\rangle := \sum_{i=1}^{r}|i_A\rangle|i_B\rangle$ between them, for M as large as possible, up to some small error ε. In other words, Alice will apply a channel Φ_A on her n qubits and Bob will apply a channel Φ_B on his n qubits so that their shared state will become

$$\sigma_{AB} = (\Phi_A \otimes \Phi_B)(\rho_{AB}) \ \approx_{\varepsilon} |\Phi^+_{r,AB}\rangle\langle\Phi^+_{r,AB}|$$

We are interested in the *asymptotic* scenario. If Alice and Bob have a large number of copies n of the state $|\phi_{AB}\rangle$, and we allow a small error, we would like the rate $\log(r)/n$ (i.e. the number of maximally entangled qubit pairs per copy of $|\phi_{AB}\rangle$) to be as large as possible. In this exercise you will show a protocol that achieves a good result (and in fact it turns out to be essentially optimal, but we will not discuss this now).

(a) Show that the marginal von Neumann entropy of $\rho_{AB} = |\phi_{AB}\rangle\langle\phi_{AB}|$ is

$$H(\mathrm{tr}_A[\rho_{AB}]) = H(\mathrm{tr}_B[\rho_{AB}]) = h(p) \ ,$$

where $h(p) = -p\log(p) - (1-p)\log(1-p)$ is the classical binary entropy. What is the von Neumann entropy of the joint state ρ_{AB}?

(b) Show that n copies of the pure state $|\phi\rangle$ can be written in the form

$$|\phi_{AB}\rangle^{\otimes n} = \sum_{\mathbf{x}\in\{0,1\}^n} p^{\frac{\#(\mathbf{x})}{2}}(1-p)^{\frac{n-\#(\mathbf{x})}{2}}|\mathbf{x}_A\rangle|\mathbf{x}_B\rangle \ ,$$

where $\#(\mathbf{x})$ is the number of zeroes in the binary string $\mathbf{x} = (x_1, \ldots, x_n) \in \{0, 1\}^n$.

(c) Alice performs a projective measurement $\{\mu_{n,k}\}_{k=0}^{n}$, such that the operator $\mu_{n,k}$ is the projection onto the subspace

$$Z_{n,k} = \mathrm{span}\{|\mathbf{x}\rangle \mid \mathbf{x} \in \{0,1\}^n \ , \ \#(\mathbf{x}) = k\} \ .$$

If Alice obtains the measurement result k, what is the post-measurement state?

(d) Suppose Bob performs the same measurement on his system. Are the measurements of Alice and Bob independent?

(e) Show that the dimension of $Z_{n,k}$ satisfies

$$\frac{\log \dim Z_{n,k}}{n} = h\left(\frac{k}{n}\right) + O(\log(n)/n) \ .$$

Hint: You can use Stirling's approximation $\log N! = N \log N - N + O(\log N)$.

(f) Let Alice's measurement outcome be described by the random variable $\mathbf{k}$. Using Lemma 8.9 or otherwise, show that for any $\delta > 0$,

$$\Pr\left(h(p) - \delta \leq h\left(\frac{\mathbf{k}}{n}\right) \leq h(p) + \delta\right) \to 1 \quad \text{as } n \to \infty \ .$$

(g) Show that for any $\varepsilon, \delta > 0$ there exists a protocol such that with probability at least $1 - \varepsilon$ the protocol produces a maximally entangled state of dimension $r(n)$ where the rate is at least $\log(r(n))/n \geq h(p) - \delta$.

Lecture 9

Quantum entropy for multiple parties

Concept	Math translation
Discarding subsystem decreases the amount of randomness	Monotonicity: $H(XY) \geq H(X)$. False for quantum states!
The total information is at most the sum of its parts.	Subadditivity: $H(AB) \leq H(A) + H(B)$.
Quantum states satisfy a weak version of monotonicity: 'adding' monotonicity relations for AB and BC	$H(A) + H(C) \leq H(AB) + H(BC)$.
Many equivalent formulations of this entropy inequality	Strong subadditivity: $H(ABC) + H(B) \leq H(AB) + H(BC)$.

We have now seen that by Shannon's Theorem 8.10 and Schumacher's Theorem 8.11 the entropy of a distribution or quantum state is a good measure of the information content of a source producing independent samples of this distribution or quantum state. We argued that the entropy was a measure of *information*, but at the same time it is also a measure of *randomness*: the entropy is large for a uniform distribution, or maximally mixed state. This can be a little counterintuitive on first encounter (as you may think of highly random processes as not being very informative). The right way to think of *informative* in this context is 'how much you can expect to learn' once you receive a sample of the state or distribution.

The Shannon and von Neumann entropy are the building blocks of information theory. Information theory beyond compression typically involves multiple systems. Paradigmatic tasks in (quantum) information theory are:

- *Communication:* Alice wants to send Bob (classical or quantum) information over a noisy channel.
- *Entanglement distillation:* Alice and Bob share some entangled states and want to extract maximally entangled states.
- *Cryptography:* Alice and Bob share quantum states and would like to exchange secret messages.

One finds that how well one can perform these tasks is often captured by entropic quantities, now involving multiple systems. In this lecture we will discuss further properties of the entropy, focusing on relations between entropies if there are multiple parties. This gives rise to a 'calculus' of information theory, allowing one to manipulate information-theoretic quantities. The main achievement of this lecture will be to prove *strong subadditivity*, a nontrivial relation between entropies on three parties.

9.1 Entropy inequalities

We will investigate situations where there is more than one system. For instance, if we have two systems A and B with some state $\rho_{AB} \in \mathrm{S}(AB)$, then we have entropies $H(A)$, $H(B)$ and $H(AB)$. These are not independent! Here are two basic examples:

Lemma 9.1. *If ρ_{AB} is pure,*

$$H(AB) = 0, \quad H(A) = H(B).$$

If $\rho_{AB} = \rho_A \otimes \rho_B$ is a product state

$$H(AB) = H(A) + H(B).$$

Proof. We already saw in Lemma 8.7 that $H(AB) = 0$ if ρ_{AB} is pure. Moreover, the nonzero eigenvalues of ρ_A and ρ_B are equal (as we saw in the Schmidt decomposition) and hence $H(A) = H(B)$. If $\rho_{AB} = \rho_A \otimes \rho_B$, then one can show (this is your Exercise 9.1) that

$$\rho_{AB} \log(\rho_{AB}) = (\rho_A \log(\rho_A)) \otimes \rho_B + \rho_A \otimes (\rho_B \log(\rho_B))$$

and by taking the trace of this expression we see that $H(AB) = H(A) + H(B)$. □

There are also a number of relations between entropies of subsystems in terms of *inequalities*. In the classical case we have the following two basic properties.

Lemma 9.2. *Let $p_{XY} \in \mathrm{P}(XY)$, then the Shannon entropy satisfies the following two inequalities:*

(a) The Shannon entropy is monotonic*:*

$$H(XY) \geq H(X).$$

(b) The Shannon entropy is subadditive*:*

$$H(XY) \leq H(X) + H(Y).$$

We have equality if and only if X and Y are independent.

The proof is Exercise 9.4. These properties have intuitive interpretations: X and Y together contain more information than just X (monotonicity) and the joint information in X and Y is at most the sum of the information in X and Y.

Subadditivity is also valid for the von Neumann entropy: if $\rho_{AB} \in \mathrm{S}(AB)$ then

$$H(AB) \leq H(A) + H(B). \tag{9.1}$$

Here, we have equality if and only if ρ_{AB} is a product state. You can prove (9.1) in Exercise 9.7 based on the operational interpretation in terms of compression, and you can prove the condition for equality later in Exercise 10.15. The intuition is that we can either separately compress A and B at rate $H(A) + H(B)$ or we can compress the joint system which may lead to a more efficient compression at rate $H(AB)$. Monotonicity, while an intuitive property, is *not* true for the von Neumann entropy! It is easy to find a counterexample: any pure entangled stated ρ_{AB} satisfies $H(AB) = 0$ and $H(A) > 0$. However, a generalization of subadditivity, *strong subbaditivity* (SSA), is a valid inequality for the von Neumann entropy.

Theorem 9.3 (Strong subadditivity)**.** *If $\rho_{ABC} \in \mathrm{S}(ABC)$,*

$$H(ABC) + H(B) \leq H(AB) + H(BC). \tag{9.2}$$

Note that if B is an empty system, then this expression reduces to (9.1). Another way to think about SSA is by writing S for the union of A and B and T for B and C, so $B = S \cap T$ and $ABC = S \cup T$. Then SSA states that

$$H(S \cup T) + H(S \cap T) \leq H(S) + H(T).$$

There are in fact many proofs of Theorem 9.3, each offering their own insights. We will give an especially simple proof below. Proofs for entropy inequalities for the von Neumann entropy are more challenging than in the classical Shannon case. The reason is that there are no obvious relations between the spectra of ρ_{ABC}, ρ_{AB} and ρ_A in general. An equivalent statement to strong subadditivity is *weak monotonicity*.

Theorem 9.4 (Weak monotonicity)**.** *If $\rho_{ABC} \in \mathrm{S}(ABC)$, then*

$$H(A) + H(C) \leq H(AB) + H(BC). \tag{9.3}$$

This is called weak monotonicity because it is a weaker statement than the monotonicity statements $H(A) \leq H(AB)$ and $H(C) \leq H(BC)$ (which individually need not be true in the quantum case). To derive strong subadditivity from weak monotonicity we may let ρ_{ABCD} be a purification of ρ_{ABC}. Then assuming weak monotonicity as in Eq. (9.3)

$$H(B) + H(D) \leq H(BC) + H(CD).$$

Since ρ_{ABCD} is pure, $H(D) = H(ABC)$ and $H(CD) = H(AB)$, giving Eq. (9.2). A very similar argument allows one to derive weak monotonicity from strong subadditivity, which is Exercise 9.3.

Why are these inequalities so fundamental? Next lecture we will see various reformulations and consequences of strong subadditivity with important operational meanings. Strong subadditivity is the main 'constraint' for information processing protocols and is useful for showing that certain protocols are optimal (we will see a concrete example in the form of Holevo's bound next lecture).

Weak monotonicity has a direct interpretation as a statement about *monogamy of entanglement*. Monogamy of entanglement is the fact that it is not possible for Bob to be maximally entangled with both Alice and Charlie. How is this captured by weak monotonicity? Violations of monotonicity are related to entanglement: $H(AB) < H(A)$ is certainly the case if the state is pure on AB and entangled. Weak monotonicity implies that we cannot have *both* $H(AB) < H(A)$ and $H(BC) < H(C)$.

9.1.1 Entropies of classical-quantum states

If just *one* of the systems is classical we do still have monotonicity. If we have an ensemble of states $\{p_X(x), \rho_{A,x}\}$ where we have state $\rho_{A,x} \in \mathrm{S}(A)$ with probability $p_X(x)$, then we may model this by a classical-quantum system XB and a classical-quantum state

$$\rho_{XA} = \sum_x p_X(x)|x\rangle\langle x| \otimes \rho_{A,x} \tag{9.4}$$

Lemma 9.5. *Let ρ_{XA} be a classical-quantum state as in Eq. (9.4), then*

$$H(XA) = \sum_x p_X(x)H(\rho_{A,x}) + H(p_X).$$

Moreover, $H(XA) \geq H(X)$ and $H(XA) \geq H(A)$.

The proof will be Exercise 9.5, where you may also show that the von Neumann entropy is strictly concave on the set of density matrices of some Hilbert space $\mathcal{H}$, i.e. if ρ_1, ρ_2 are states on $\mathcal{H}$ and $p \in (0, 1)$, then

$$H(p\rho_1 + (1-p)\rho_2) \geq pH(\rho_1) + (1-p)H(\rho_2) \tag{9.5}$$

with equality if and only if $\rho_1 = \rho_2$. More generally, we have for any ensemble $\{p_X(x), \rho_{A,x}\}$

$$\sum_x p_X(x)H(\rho_{A,x}) \leq H(\sum_x p_X(x)\rho_{A,x}) \leq \sum_x p_X(x)H(\rho_{A,x}) + H(p_X).$$

9.2 Proof of weak monotonicity

Strong subadditivity, as formulated in Theorem 9.3, is the fundamental theorem of quantum information. We will now prove this central result by proving weak monotonicity, which we saw to be equivalent.

We need a basic fact on logarithms of positive operators. For completeness, we provide a proof.

Lemma 9.6. *If $P, Q \in \mathrm{PD}(\mathcal{H})$ and $P \leq Q$, then $\log(P) \leq \log(Q)$.*

Proof. It follows from Corollary A.3 that for any $M, N, X \in \mathrm{Lin}(\mathcal{H})$

$$M \leq N \Rightarrow XMX^\dagger \leq XNX^\dagger \tag{9.6}$$

We will use this to prove that if $P, Q \in \mathrm{PD}(\mathcal{H})$, then

$$P \leq Q \Rightarrow P^{-1} \geq Q^{-1}.$$

To see this, suppose $P, Q \in \mathrm{PD}(\mathcal{H})$ are such that $P \leq Q$. By Eq. (9.6) $\mathbb{1} \leq P^{-\frac{1}{2}}QP^{-\frac{1}{2}}$. The operator $P^{-\frac{1}{2}}QP^{-\frac{1}{2}}$ is positive, and we see that all its eigenvalues are at least 1. This implies that its inverse $P^{\frac{1}{2}}Q^{-1}P^{\frac{1}{2}}$ is positive and has eigenvalues *at most* 1. Therefore, $\mathbb{1} \geq P^{\frac{1}{2}}Q^{-1}P^{\frac{1}{2}}$. Another application of Eq. (9.6) gives $P^{-1} \geq Q^{-1}$.

The next step is that we are going to use that for any $x > 0$

$$\log(x) = \int_0^\infty \left(\frac{1}{1+t} - \frac{1}{x+t}\right) \mathrm{d}t$$

and therefore, for any positive operator $P \in \mathrm{PD}(\mathcal{H})$

$$\log(P) = \int_0^\infty \left(\frac{1}{1+t}\mathbb{1} - (P + t\mathbb{1})^{-1} \right) \mathrm{d}t. \tag{9.7}$$

Note that $P + t\mathbb{1} \in \mathrm{PD}(\mathcal{H})$. Now, if P and Q are both positive definite and $P \leq Q$, then for all t we have $P + t\mathbb{1} \leq Q + t\mathbb{1}$ and hence $(P + t\mathbb{1})^{-1} \geq (Q + t\mathbb{1})^{-1}$, and from Eq. (9.7) we conclude that $\log(P) \leq \log(Q)$. □

Remark 9.7. If we have a function $f : I \subset \mathbb{R} \to \mathbb{R}$ on some interval I, then we say that f is *operator monotone* if for all Hermitian matrices M, N with spectrum contained in I we have that $M \leq N$ implies $f(M) \leq f(N)$. Lemma 9.6 can then be interpreted as proving that $\log : \mathbb{R}_{>0} \to \mathbb{R}$ is operator monotone.

The key to the proof of weak monotonicity is the following.

Lemma 9.8. *Suppose that $\rho_{ABC} \in \mathrm{S}(ABC)$ is such that ρ_A, ρ_C and ρ_{BC} are invertible (i.e. have full rank), then*

$$\rho_{AB} \otimes \rho_C^{-1} \leq \rho_A \otimes \rho_{BC}^{-1}$$

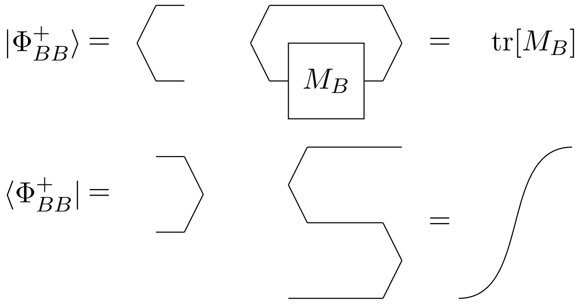
Proof. Let

$$|\Phi^+_{BB}\rangle = \sum_{b=0}^{|B|-1} |bb\rangle$$

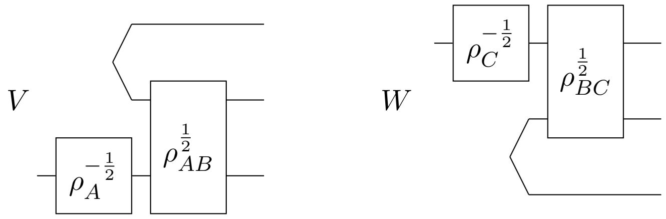
be an *unnormalized* maximally entangled state on two copies of B. We now define two linear maps $V \in \mathrm{Lin}(A, ABB)$ and $W \in \mathrm{Lin}(C, BBC)$ by

$$V = (\rho_{AB}^{\frac{1}{2}} \otimes \mathbb{1}_B)(\rho_A^{-\frac{1}{2}} \otimes |\Phi^+_{BB}\rangle)$$
$$W = (\mathbb{1}_B \otimes \rho_{BC}^{\frac{1}{2}})(|\Phi^+_{BB}\rangle \otimes \rho_C^{-\frac{1}{2}}).$$

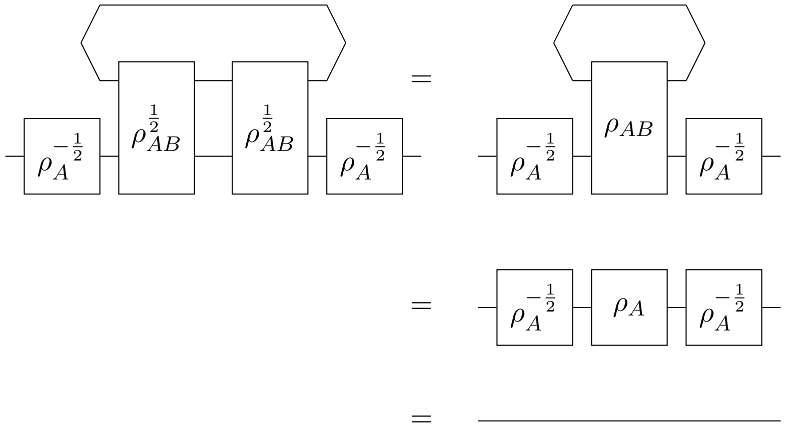
We will use a graphical proof, as in Lecture 5. Recall the following notation and facts



So, V and W are given by



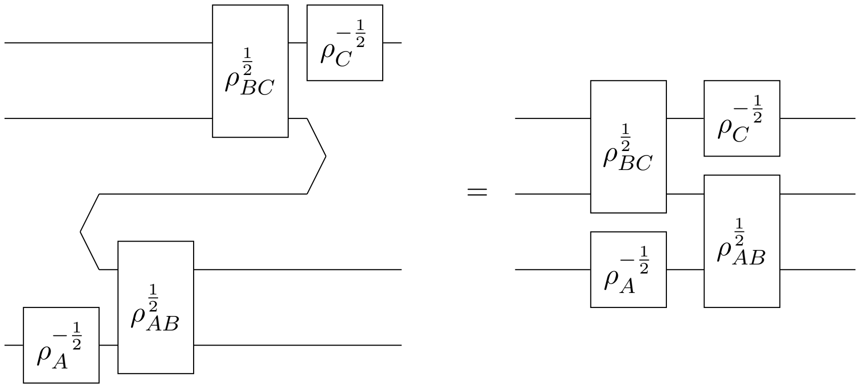
We read these diagrams from left to right (which is the converse order in which we write compositions). We may use this to show that V and W are actually isometries. For V the proof is as follows: $V^\dagger V$ is given by



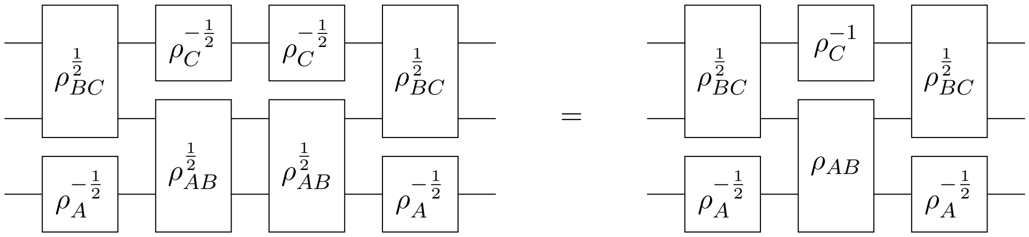
and for W the argument is analogous. Now we let K be the following operator:

$$K = (\mathbb{1}_{AB} \otimes W^\dagger)(V \otimes \mathbb{1}_{BC}).$$

We see that K is given by


$\rho_{BC}^{\frac{1}{2}}$
$\rho_C^{-\frac{1}{2}}$
$\rho_{AB}^{\frac{1}{2}}$
$\rho_A^{-\frac{1}{2}}$
$=$
$\rho_{BC}^{\frac{1}{2}}$
$\rho_C^{-\frac{1}{2}}$
$\rho_{AB}^{\frac{1}{2}}$
$\rho_A^{-\frac{1}{2}}$

and hence $K^\dagger K$ equals



The operator $K^\dagger K$ is positive by Lemma A.2. Moreover, since it is a composition of isometries and their adjoints (which all have operator norm at most 1), from the submultiplicativity of the operator norm (Eq. (6.1)) it follows that $\|K^\dagger K\|_\infty \leq 1$. Therefore, $K^\dagger K$ has eigenvalues in the interval $[0, 1]$ and

$$K^\dagger K = (\rho_A^{-1} \otimes \rho_{BC})^{\frac{1}{2}}(\rho_{AB} \otimes \rho_C^{-1})(\rho_A^{-1} \otimes \rho_{BC})^{\frac{1}{2}} \leq \mathbb{1}_{ABC}$$

This implies, by Eq. (9.6)

$$\rho_{AB} \otimes \rho_C^{-1} \leq \rho_A \otimes \rho_{BC}^{-1}.$$

□

We are now ready to prove weak monotonicity!

Proof of Theorem 9.4. Let us first assume that ρ_{ABC} is such that ρ_A, ρ_C, ρ_{AB} and ρ_{BC} all have full rank (equivalently, they are strictly positive definite, or invertible). Then, by Lemma 9.6 and Lemma 9.8 we have that

$$\log(\rho_{AB} \otimes \rho_C^{-1}) \leq \log(\rho_A \otimes \rho_{BC}^{-1})$$

Using Exercise 9.1 this gives

$$\log(\rho_A) \otimes \mathbb{1}_{BC} + \mathbb{1}_{AB} \otimes \log(\rho_C) - \log(\rho_{AB}) \otimes \mathbb{1}_C - \mathbb{1}_A \otimes \log(\rho_{BC}) \geq 0.$$

We may now take the trace with ρ_{ABC}, and by Lemma A.2 we have

$$\operatorname{tr}[\rho_{ABC} \left(\log(\rho_A) \otimes \mathbb{1}_{BC} + \mathbb{1}_{AB} \otimes \log(\rho_C) - \log(\rho_{AB}) \otimes \mathbb{1}_C - \mathbb{1}_A \otimes \log(\rho_{BC})\right)] \geq 0$$

In other words,

$$H(A) + H(C) - H(AB) - H(BC) \leq 0 \tag{9.8}$$

which is weak monotonicity, so this proves the result under the assumption that the reduced density matrices have full rank. Now, if $\rho_{ABC} \in \mathrm{S}(ABC)$ is arbitrary, let $\varepsilon > 0$ and

$$\rho^{\varepsilon}_{ABC} = (1 - \varepsilon)\rho_{ABC} + \varepsilon\tau_{ABC}$$

where τ_{ABC} is the maximally mixed state. This state satisfies the full rank assumption, and hence

$$H(A)_{\rho^\varepsilon} + H(C)_{\rho^\varepsilon} - H(AB)_{\rho^\varepsilon} - H(BC)_{\rho^\varepsilon} \leq 0$$

If we let $\varepsilon \to 0$ we recover ρ_{ABC}, and since the entropy is continuous, we must have

$$H(A)_\rho + H(C)_\rho - H(AB)_\rho - H(BC)_\rho = \lim_{\varepsilon\to 0} H(A)_{\rho^\varepsilon} + H(C)_{\rho^\varepsilon} - H(AB)_{\rho^\varepsilon} - H(BC)_{\rho^\varepsilon} \leq 0$$

which proves weak monotonicity for arbitrary ρ_{ABC}. □

Outlook

The proof in this lecture for weak monotonicity (and thereby strong subadditivity) follows [26]. However, there exists a broad variety of proofs, each providing their own insights. The first proof was given by Lieb and Ruskai in [25], based on concavity results for certain functions of operators. Conditions for *equality* in strong subadditivity are given by [20].

Entropy inequalities beyond strong subadditivity

We have so far restricted to inequalities involving at most three parties, but nothing prevents us from studying inequalities involving more parties. We can then simply apply the inequalities we have already seen above to obtain new inequalities. For the Shannon entropy we saw that the entropy was positive, and that it satisfied monotonicity and subadditivity. It turns out that there are *additional* inequalities (which are linear in the subsystem entropies) which are known as *non-Shannon type entropy inequalities* for four or more parties [50].

In the quantum case, it is an open question whether there are any entropy inequalities beyond the ones that are a direct implication of strong subadditivity and positivity. One way to formulate this question is by studying *entropy cones*. To this end, for n parties one considers a real vector space of dimension $2^n - 1$, with coefficients labeled by non-empty subsets $I \subseteq \{1, \ldots, n\}$. Every quantum states on n parties $\rho_{A_1 \ldots A_n}$ (for some arbitrary finite dimensional quantum systems) defines a vector

$$(H(A_I))_{I \subset 1, \ldots, n} \in \mathbb{R}^{2^n - 1}$$

of the values of all entropies of different choices of subsystems, and where we write A_I for the union of all A_i for which $i \in I$. For example, for $n = 2$ we would get vectors like

$$(H(A_1), H(A_2), H(A_1 A_2)) \in \mathbb{R}^3.$$

One can show that if one takes the closure of this set one obtains a *cone* Σ, meaning that if $v, w \in \Sigma$, then $v + w \in \Sigma$ and $\lambda v \in \Sigma$ for $\lambda \geq 0$. We call this cone the *quantum entropy cone* [36]. Since the entropy is positive, this cone lies in the positive orthant. However, not any vector of positive values can be realized as a vector of entropies. For example, strong subadditivity defines a hyperplane that constrains the cone to lie on one side. In general the cone is defined by a set of linear equations, and these equations are precisely the 'valid' entropy inequalities. Determining all valid entropy inequalities is equivalent to determining the shape of the quantum entropy cone. If we restrict to probability distributions instead of quantum states we obtain the classical entropy cone [50]; in this case it is known that there are infinitely many relevant inequalities (i.e. the entropy cone is not a polytope).

9.3 Exercises

9.1 **Operator logarithm:**

(a) Compute $\log(M)$ for

$$M = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \text{ for } \alpha > 0, \qquad \text{and} \qquad M = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

(b) Show that if $P > 0$, $Q > 0$ are positive definite operators

$$\log(P \otimes Q) = \log(P) \otimes \mathbb{1} + \mathbb{1} \otimes \log(Q).$$

(c) Show that if $\rho_A \in \mathrm{S}(A)$, $\rho_B \in \mathrm{S}(B)$

$$\rho_A \otimes \rho_B \log(\rho_A \otimes \rho_B) = \rho_A \log(\rho_A) \otimes \rho_B + \rho_A \otimes \rho_B \log(\rho_B).$$

9.2 **Entropy of states:** Consider the 2-qubit state

$$\rho_{AB} = \frac{1}{8}\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \in S(AB) \ ,$$

where $\mathcal{H}_A \cong \mathcal{H}_B \cong \mathbb{C}^2$.

(a) Compute the entropies $H(AB)_\rho$, $H(A)_\rho$, and $H(B)_\rho$.
(b) Compute $H(A|B)_\rho$, $H(B|A)_\rho$, and $I(A:B)_\rho$.
(c) Now consider the 3-party GHZ state $\tau_{ABC} = |\mathrm{GHZ}\rangle\langle\mathrm{GHZ}|_{ABC}$, where

$$|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}\big(|000\rangle + |111\rangle\big) \ .$$

Compute $H(ABC)_\tau$, $H(AB)_\tau$, and $H(A)_\tau$.

9.3 **Weak monotonicity and strong subadditivity:** Show (similar to the argument following Theorem 9.4) that strong subadditivity of the von Neumann entropy (Eq. (9.2)) implies weak monotonicity as in Eq. (9.3).

9.4 **Monotonicity and subadditivity of the Shannon entropy:** The aim of this exercise is to prove Lemma 9.2. Consider the joint probability distribution $p_{XY} = \Pr(XY)$.

(a) Verify that

$$\sum_y p_Y(y) H(X|Y=y) = H(XY) - H(Y). \tag{9.9}$$

Here $H(X|Y=y)$ is the entropy of the conditional distribution $p_{X|y}$ as defined in Eq. (2.5).
(b) Use Eq. (9.9) to show that $H(XY) \geq H(X)$. When is this an equality?
(c) Prove directly that $H(X) + H(Y) \geq H(XY)$, with equality if and only if X and Y are independent. *Hint: Take the difference of the left- and right-hand sides, rewrite and apply Jensen's inequality to the function $x \mapsto -\log(x)$.*
(d) Use Theorem 8.10 to obtain an alternative proof that $H(X) + H(Y) \geq H(XY)$.

9.5 **Entropies of classical-quantum states:**

(a) Let ρ_{XA} be a classical-quantum state, so

$$\rho_{XA} = \sum_x p_X(x) |x\rangle\langle x| \otimes \rho_{A,x}$$

for some states $\rho_{A,x}$ and a probability distribution p_X. Show that

$$H(XA) = \sum_x p_X(x) H(\rho_{A,x}) + H(p_X).$$

(b) Conclude that $H(XA) \geq H(X)$.
(c) Prove that the von Neumann entropy is strictly concave, as in Eq. (9.5). *Hint: use subadditivity of the von Neumann entropy.*

(d) Next, argue that you can choose a system R and an orthogonal collection of states $|\phi_{AR,x}\rangle$ such that $|\phi_{AR,x}\rangle$ is a purification of $\rho_{A,x}$. Now apply strong subadditivity to the state

$$\rho_{XAR} = \sum_x p_X(x)|x\rangle\langle x| \otimes |\phi_{AR,x}\rangle\langle\phi_{AR,x}|$$

to show that $H(XA) \geq H(A)$.

(e) Conclude that

$$\sum_x p_X(x)H(\rho_{A,x}) \leq H(\sum_x p_X(x)\rho_{A,x}) \leq \sum_x p_X(x)H(\rho_{A,x}) + H(p_X).$$

9.6 **The Araki-Lieb inequality:** Show that for $\rho_{AB} \in \mathrm{S}(AB)$ the difference between the von Neumann entropies on A and B is bounded by

$$|H(A) - H(B)| \leq H(AB).$$

This is known as the *Araki-Lieb inequality. Hint: apply subadditivity to a purification of* ρ_{AB}.

9.7 **Subadditivity of the von Neumann entropy:** Use Exercise 7.3 to prove the subadditivity of the von Neuman entropy:

$$H(AB)_\rho \leq H(A)_\rho + H(B)_\rho$$

for $\rho_{AB} \in \mathrm{S}(AB)$. Explain the operational meaning of subadditivity in terms of compression.

Lecture 10

Bounds on information processing

Concept	Math translation
Entropy differences measure the amount of information contained in B about a system A	The *conditional entropy* $$H(A\|B) = H(AB) - H(B)$$ is the amount of uncertainty left in A when learning B. The *mutual information* $$I(A:B) = H(A) + H(B) - H(AB)$$ is how much you learn about A when you learn B. These are classical interpretations.
Processing data never increases the amount of information.	Data processing inequalities for conditional entropy and mutual information: if $\sigma_{AC} = (\mathcal{I}_A \otimes \Phi_{B\to C})(\rho_{AB})$, then $$I(A:C)_\sigma \leq I(A:B)_\rho$$ $$H(A\|C)_\sigma \geq H(A\|B)_\rho.$$
It is not possible to send more than one classical bit using one qubit (without using additional entanglement)	*Holevo bound:* the information that can be extracted from an ensemble $\{p_X(x), \rho_{A,x}\}$ about X by measurement is bounded by $$I(X:Y) \leq \chi(\{p_X(x), \rho_{A,x}\}).$$
The ability to distinguish quantum states is related to the ability to send information by encoding into quantum states.	Reduction from a *quantum state learning* problem to a *communication* problem. Next apply Holevo bound to get a lower bound on the number of copies of a state required to determine an accurate approximation of the state.

In this lecture we will introduce two natural information-theoretic measures which are derived from the entropy, measuring *correlations* between two systems. We will see that strong subadditivity has a natural interpretation as a *data processing inequality* for such systems.

We then study the Holevo bound (a consequence of data processing), which sets limits on the amount of *classical* information that can be sent using quantum states. In particular this will show us that in the absence of entanglement, one can only send a single (classical) bit using one qubit of communication.

As a further application of the Holevo bound we then give an information theoretic bound on learning quantum states. Given n copies of an unknown state ρ_A we would like to determine ρ_A by performing measurements on $\rho_A^{\otimes n}$. How many copies n do we need to get an accurate answer?

10.1 Entropic correlation measures

Suppose that we would like to send information over a classical (noisy) channel. The input system is X, and we call the output of the channel Y. In order to understand to what extent we can send information over this channel, we must quantify *how much we learn about X when we obtain Y*. We will introduce two (closely related) quantities: the conditional entropy, and the mutual information.

10.1.1 Conditional entropy

Recall that for a joint probability distribution $p_{XY} \in \mathrm{P}(XY)$ we have *conditional probabilities* $p_{X|y}$, which is the probability of x given y and which is such that $p_{XY}(x, y) = p_{X|y}(x)p_Y(y)$. We may consider the entropy of X given y:

$$H(X|Y = y) := -\sum_x p_{X|y}(x) \log(p_{X|y}(x)).$$

Now, the *conditional Shannon entropy* of X given Y is the expected value of $H(X|Y = y)$

$$H(X|Y) = \sum_y p_Y(y) H(X|Y = y).$$

The intuitive interpretation of this expression is that it is the expected amount of information in X once you learn the outcome of Y. An easy computation in Exercise 9.4 shows that

$$H(X|Y) = H(XY) - H(Y). \tag{10.1}$$

This also makes sense: the information we have about X when we know Y is the total information on XY minus the amount of information in Y.

In quantum theory there is in general no way to define conditional probabilities or conditional states. However, nothing stands in our way of defining the conditional entropy for $\rho_{AB} \in \mathrm{S}(AB)$ nevertheless based on Eq. (10.1).

Definition 10.1. If $\rho_{AB} \in \mathrm{S}(AB)$ the *conditional entropy* of A conditioned on B is defined as

$$H(A|B)_\rho = H(AB)_\rho - H(B)_\rho$$

where we omit the dependence on ρ_{AB} and write $H(A|B)$ if the state is clear from the context.

It has the perhaps surprising property that it is possible that $H(A|B) < 0$ (since the von Neumann entropy is not monotonic). Later we will see a nice operational interpretation of negative conditional entropies.

Example 10.2. Let us compute the conditional entropy for three important examples of states on two qubits A and B.

- If $\rho_{AB} = \frac{1}{4}\mathbb{1}_{AB}$ is the maximally mixed state we have
$$H(AB) = 4 \times \frac{1}{4}\log(4) = 2 \qquad H(B) = 2 \times \frac{1}{2}\log(2) = 1$$
so $H(A|B) = 1 = H(A)$. We see that $H(A|B) = H(A)$ so when we get B we learn nothing about A and the amount of information in A stays the same.
- If $\rho_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$ is the maximally correlated state, we see that ρ_{AB} has nonzero eigenvalues $\frac{1}{2}, \frac{1}{2}$, while the reduced density matrices are maximally mixed, so
$$H(AB) = 2 \times \frac{1}{2}\log(2) = 1 \qquad H(B) = 2 \times \frac{1}{2}\log(2) = 1$$
and hence $H(A|B) = 0$. This makes sense with our (classical) interpretation: if we learn the outcome of B we know the value of A exactly and hence there is no information (randomness) left in A.
- If $\rho_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|$ is a maximally entangled state, $H(AB) = 0$ since the state is pure and as the reduced states are maximally mixed $H(B) = 1$ and hence $H(A|B) = -1$.

Lemma 10.3. *Let $\rho_{AB} \in \mathrm{S}(AB)$, then $H(A|B) = H(A|B)_\rho$ has the following properties.*

(a) If ρ_{AB} is pure, then $H(A|B) = -H(A) = -H(B)$. If ρ_{ABC} is pure, $H(A|B) = -H(A|C)$.

(b) We have the lower bound
$$H(A|B) \geq -H(A) \geq -\log(|A|).$$
If the system X is classical, we have $H(A|X) \geq 0$ and $H(X|A) \geq 0$.

(c) We have the upper bound
$$H(A|B) \leq H(A) \leq \log(|A|).$$
The first inequality is an equality if and only if $\rho_{AB} = \rho_A \otimes \rho_B$ is a product state.

(d) The conditional entropy is invariant under isometries on the subsystems. That is, if $V \in \mathrm{Isom}(A, A')$ and $W \in \mathrm{Isom}(B, B')$, and $\sigma_{A'B'} = (V \otimes W)\rho_{AB}(V^\dagger \otimes W^\dagger)$ then
$$H(A|B)_\rho = H(A'|B')_\sigma.$$

Proof. If ρ_{ABC} is pure, then
$$H(A|B) = H(AB) - H(B) = H(C) - H(AC) = -H(A|C).$$
If ρ_{AB} is pure, $H(AB) = 0$ and $H(A) = H(B)$, proving (a). Next, (b) is clear from the definition. If X is classical, Lemma 9.5 implies that $H(A|X)$ and $H(X|A)$ are nonnegative. Subadditivity of the von Neumann entropy (Eq. (9.1)) is equivalent to (c). Finally, (e) is a direct consequence of the invariance of the von Neumann entropy under isometries. □

We may rephrase strong subadditivity $H(ABC) + H(B) \leq H(AB) + H(BC)$ as

$$H(ABC) - H(BC) \leq H(AB) - H(B)$$

and hence

$$H(A|BC) \leq H(A|B). \tag{10.2}$$

More generally:

Theorem 10.4 (Data processing conditional entropy). *If $\Phi_{B\to C} \in \mathrm{C}(B, C)$ and we have $\rho_{AB} \in \mathrm{S}(AB)$, $\sigma_{AC} = (\mathcal{I}_A \otimes \Phi_{B\to C})(\rho_{AB})$, then*

$$H(A|C)_\sigma \geq H(A|B)_\rho$$

Proof. Let $V \in \mathrm{Isom}(B, CE)$ be a Stinespring extension, and let $\omega_{ACE} = (\mathbb{1}_A \otimes V)\rho_{AB}(\mathbb{1}_A \otimes V^\dagger)$. Then by invariance of entropy under isometries,

$$H(A|B)_\rho = H(A|CE)_\omega$$

and the result follows from Eq. (10.2) since $\omega_{AC} = \sigma_{AC}$. □

The intuition behind this result is that if C is a 'processed' version of B, then we will learn less about A once we receive the C system. While this may be intuitive, the proof relies on the nontrivial strong subadditivty property, and it is also not very clear what the intuitive meaning should be in case the conditional entropy is negative!

We may also reformulate weak monotonicity (and hence strong subadditivity) as

$$H(AB) - H(A) + H(BC) - H(C) \geq 0$$

and hence as

$$H(B|A) + H(B|C) \geq 0. \tag{10.3}$$

This corresponds to the monogamy of entanglement interpretation: we can not have both $H(B|A)$ and $H(B|C)$ negative. While $H(A|B) < 0$ is not a *necessary* condition for entanglement, it is a sufficient condition: as you may check in Exercise 10.4 every state $\rho_{AB} \in \mathrm{S}(AB)$ with $H(A|B) < 0$ is entangled.

10.1.2 The mutual information

Another natural entropic quantity is the *mutual information*.

Definition 10.5. Given $p_X \in \mathrm{P}(XY)$ we define the mutual information as

$$I(X : Y) = H(X) + H(Y) - H(XY)$$

and similarly for a quantum state $\rho_{AB} \in \mathrm{S}(AB)$

$$I(A : B)_\rho = H(A)_\rho + H(B)_\rho - H(AB)_\rho.$$

We write $I(A : B)$ if the state is clear from the context.

What is the meaning of the mutual information? The idea is that it is a measure for the correlation between A and B. You can think of it as 'the amount of information you can learn about A from B'. The mutual information is related to the conditional entropy as follows (as seen directly from the definition):

$$I(A:B) = H(A) - H(A|B) = H(B) - H(B|A). \tag{10.4}$$

Example 10.6. Let us compute the mutual information entropy for the same qubit states as in Example 10.2.

- If $\rho_{AB} = \frac{1}{4}\mathbb{1}_{AB}$ is the maximally mixed state we have

  $$H(AB) = 2 \qquad H(A) = H(B) = 1$$

  so $I(A:B) = 1 + 1 - 2 = 0$. Indeed, A and B are independent, so we learn nothing about A from B.
- If $\rho_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$ is the maximally correlated state

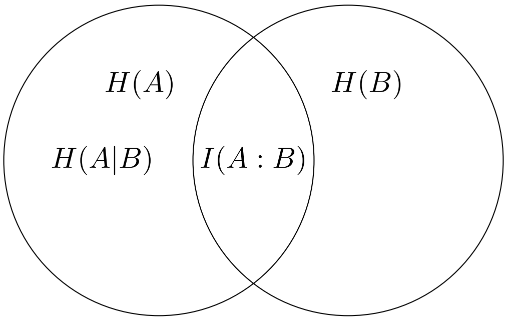
  $$H(AB) = 1 \qquad H(A) = H(B) = 1$$

  and hence $I(A:B) = 1 + 1 - 1 = 1$. The maximally correlated state indeed represents one bit of correlation.
- If $\rho_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|$ is a maximally entangled state

  $$H(AB) = 0 \qquad H(A) = H(B) = 1$$

  and therefore $I(A:B) = 1 + 1 - 0 = 2$, so this is a 'stronger' correlation than for the maximally correlated state.

The relation between the entropies $H(A)$, $H(B)$, $H(AB)$ and $H(A|B)$ and $I(A:B)$ may be visualized in the diagram



The mutual information has the following basic properties.

Lemma 10.7. *Let $\rho_{AB} \in \mathrm{S}(AB)$, then the mutual information $I(A:B) = I(A:B)_\rho$ has the following properties:*

(a) If ρ_{AB} is pure, then $I(A:B) = 2H(A) = 2H(B)$.

(b) $I(A:B) \geq 0$ with equality if and only if $\rho_{AB} = \rho_A \otimes \rho_B$.

(c) We have the upper bound

$$I(A:B) \leq 2\min(H(A), H(B)) \leq 2\min(\log(|A|), \log(|B|)).$$

(d) If the system X is classical, then

$$I(X:B) \leq \min(H(X), H(B)) \leq \min(\log(|X|), \log(|B|)).$$

(e) The mutual information is invariant under isometries on the subsystems. That is, if $V \in \mathrm{Isom}(A, A')$ and $W \in \mathrm{Isom}(B, B')$, and $\sigma_{A'B'} = (V \otimes W)\rho_{AB}(V^\dagger \otimes W^\dagger)$ then

$$I(A:B)_\rho = I(A':B')_\sigma.$$

Proof. This follows directly from the properties of the conditional entropy we proved in Lemma 10.3. □

Finally, we have a data processing inequality for the mutual information:

Theorem 10.8 (Data processing mutual information)**.** *If $\Phi_{B\to C} \in \mathrm{C}(B, C)$, then for $\rho_{AB} \in \mathrm{S}(AB)$, $\sigma_{AC} = (\mathcal{I}_A \otimes \Psi_{B\to C})(\rho_{AB})$ we have*

$$I(A:B)_\rho \geq I(A:C)_\sigma$$

Proof. This is a direct consequence of the data processing inequality for the conditional entropy in Theorem 10.4. □

As before, it has the intuitive interpretation that by only acting on one of the subsystems we can never get more information about the other subsystem. This statement is easily seen to be *equivalent* to strong subadditivity, and assigns a nice operational meaning to strong subadditivity.

10.2 Continuity estimates

The fact that the function $x \mapsto x\log(x)$ is continuous on the interval $[0,1]$ implies directly that the Shannon and von Neumann entropy are continuous functions. Often it is useful to have concrete bounds on how much the entropy changes under small deformations of the state. We have the following quantitative continuity estimate, which you may prove in Exercise 10.13.

Theorem 10.9. *If $\rho_{AB}, \sigma_{AB} \in \mathrm{S}(AB)$ satisfy $T(\rho_{AB}, \sigma_{AB}) \leq \varepsilon$, then*

$$|H(A|B)_\rho - H(A|B)_\sigma| \leq 2\varepsilon\log(|A|) + (1+\varepsilon)h\left(\frac{\varepsilon}{1+\varepsilon}\right)$$

where h is the binary entropy function from Eq. (8.2).

In the special case where there is no B system this reduces to a continuity estimate for the regular von Neumann entropy $H(A)$. Theorem 10.9 also gives a continuity estimate for the mutual information: if $\rho_{AB}, \sigma_{AB} \in \mathrm{S}(AB)$ satisfy $T(\rho_{AB}, \sigma_{AB}) \leq \varepsilon$, then

$$|I(A:B)_\rho - I(A:B)_\sigma| \leq 4\varepsilon \min(\log(|A|), \log(|B|)) + \frac{2}{1+\varepsilon} h\left(\frac{\varepsilon}{1+\varepsilon}\right). \tag{10.5}$$

10.3 The Holevo bound

We will now investigate the question of how much classical information we can encode in a quantum state. We already know, from the superdense coding protcol, that if we have entanglement available we can send over two classical bits using one qubit. Now we will look at the situation where we try to encode some classical register X into an *ensemble* of quantum states where we have state $\rho_{A,x} \in \mathrm{S}(A)$ with probability p_x. This gives rise to an associated classical-quantum state

$$\rho_{XA} = \sum_x p_x |x\rangle\langle x| \otimes \rho_{A,x}.$$

Definition 10.10. We define the *Holevo χ-quantity* of an ensemble $\{p_x, \rho_{A,x}\}$ as

$$\chi(\{p_x, \rho_{A,x}\}) = I(X:A)_\rho$$

Writing out the definition, using Lemma 9.5 we see that

$$\chi(\{p_x, \rho_{A,x}\}) = H(\sum_x p_x \rho_{A,x}) - \sum_x p_x H(\rho_{A,x}).$$

Moreover, by Lemma 10.7 we have

$$0 \leq \chi(\{p_x, \rho_{A,x}\}) \leq \min(H(p), H(\rho_A)). \tag{10.6}$$

The upper bound, which is based on the result of Exercise 9.5, relies on strong subadditivity again!

Now we think of the following set-up: Alice has a classical source X and chooses to encode this using an ensemble of quantum states (i.e. if she has classical x she encodes this into $\rho_{A,x}$). She then sends over the state to Bob, who will do a measurement and store the outcomes in a classical register Y. The question is how much Bob can learn about X. An upper bound is given by the Holevo bound.

Theorem 10.11 (Holevo bound)**.** *The mutual information between X and Y is upper bounded by*

$$I(X:Y) \leq \chi(\{p_x, \rho_{A,x}\}).$$

Proof. The final state is obtained by taking the classical-quantum state ρ_{XA} and applying a measurement channel $\Phi_{A\to Y}$ to the A-system, so the classical state between X and Y is given by $\sigma_{XY} = (\mathcal{I}_X \otimes \Phi_{A\to Y})(\rho_{XA})$. We have

$$I(X:Y)_\sigma \leq I(X:A)_\rho = \chi(\{p_x, \rho_{A,x}\}).$$

The result follows directly from the data processing inequality in Theorem 10.8. □

This shows that if we try to encode into a n-qubit system A, the Holevo quantity is upper bounded by $H(A) \leq \log(|A|) = n$, and we can not achieve more than n bits of mutual information between Alice and Bob by sending over one qubit. In other words, this proves that the use of shared entanglement in superdense coding is necessary!

Why does $I(X : Y) \leq n$ mean that we can not communicate more than n bits? Note first that if there exists a (classical) channel from Y to X' which is such that it exactly recovers the original message and we start with a unform distribution on X, this means that we get the distribution $p_{XX'}(x, x')$ which is zero if $x \neq x'$ and is uniform over pairs (x, x). This is a maximally correlated state and in analogy with Example 10.6 it has mutual information equal to $\log(|X|)$. This means that according to the Holevo bound the number of bits we can send *with zero error* is $\log(|X|) \leq n$.

If we do allow some error, then we can lower bound the probability of error by *Fano's inequality.*

Lemma 10.12 (Fano's inequality). *Let X, X' be classical systems with a joint distribution $p_{XX'}(x, x')$. Let p_e be the probability of error, where $x \neq x'$. Then*

$$h(p_e) + p_e \log(|X| - 1) \geq H(X|X').$$

In other words, if there still uncertainty left in X after learning Y, meaning that

$$H(X|X') \geq H(X|Y) > 0$$

then you can only recover X from Y with a nonzero error probability.

In the special case where we take a uniform distribution on X we find

$$I(X : X') = H(X) - H(X|X') \geq (1 - p_e) \log(|X|) - h(p_e)$$

using that $H(X) = \log(|X|)$. If we send over n qubits, $I(X : X') \leq n$ and we see that if $\log(|X|) > n$ then the probability of correct decoding is bounded as

$$1 - p_e \leq \frac{I(X : X') + 1}{\log(|X|)}.$$

In fact, Shannon's *noisy coding theorem* states that if we have a classical channel Q from X to Y, and we want to use many copies of this channel to reliably send over information, then we can do so at an optimal rate (which is the number of channel uses per bit of transferred information) of

$$C(Q) = \max_{p_X \in \mathrm{P}(X)} I(X : Y) \tag{10.7}$$

where we compute $I(X : Y)$ with respect to the joint distribution obtained from

$$p_{XY}(x, y) = q(x|y) p_X(x)$$

where $q(x|y)$ is the transition function of the channel Q (see Definition 4.1). The quantity $C(Q)$ is known as the *capacity* of the channel Q. From Eq. (10.7) and the Holevo bound we see that the classical capacity of the channel which results from encoding into n quantum bits, sending them over and performing measurements is at most $C(Q) \leq n$. A natural next question is how much quantum information we can send over if we have an arbitrary quantum channel $\Phi_{A \to B}$. The Holevo bound is the starting point for such investigations, but the answer is not as simple as in the classical case in Eq. (10.7).

10.4 Lower bounds for learning pure quantum states

Consider the following scenario: Alice receives n copies of an unknown quantum state ρ_A. She will do a measurement, process the measurement outcomes and come up with a classical description of an estimate $\hat{\rho}_A$. She desires to approximate ρ_A with accuracy ε, so

$$T(\rho_A, \hat{\rho}_A) \leq \varepsilon$$

with sufficiently high probability, regardless of which state ρ_A she received. We allow her to do measurements on all n copies at the same time, so she may measure the state $\rho_A^{\otimes n}$. She would like to succeed with probability at least $1-\delta$ for some small δ. This task is also known as *quantum state tomography.* We will discuss the special case where we assume that the unknown state is pure. A fundamental question is: how many copies of the state does Alice need to perform this task to the desired accuracy? We will ignore how *complicated* these measurements are (i.e. what kind of computations a (quantum) computer would have to perform in order to learn the state) and focus purely on the required number of copies. The required number of copies for such a learning task is called the *sample complexity.*

The value of the optimal sample complexity will depend on the accuracy ε, the dimension of the Hilbert space $|A|$ and the probability of success $1-\delta$. We will use Holevo's bound to give a lower bound (so a minimal number of copies required). Through different arguments one can show that this lower bound is close to optimal (so there is a nearly matching upper bound).

The high-level idea of the bound is simple: we will relate the ability to learn states with n copies to the ability to communicate classical information by sending n copies of the system A. We reduce the *learning scenario* to a *communication scenario*: given a learning procedure, we construct a communication protocol, which we then know to be bounded by the Holevo bound. Suppose that one can distinguish states on A to precision ε using n copies. We will then find a large set of states $\rho_{A,x} \in \mathrm{S}(A)$ such that for all $x \neq x'$ the distance between $\rho_{A,x}$ and $\rho_{A,x'}$ is at least 2ε. We then set up a communication protocol where Alice sends a classical message x from the classical system X using n quantum states by sending the state $\rho_{A,x}^{\otimes n}$. Bob will decode by trying to learn the state, and take as decoded message the x' such that $\rho_{A,x'}$ is closest to his estimate. By assumption, the decoding will succeed with probability at least δ (since there are no states $\rho_{A,y}$ with $y \neq x$ within distance ε of $\rho_{A,x}$).

What we will do is argue that there exists such a collection of pure states such that $\log(|X|)$ is of the order $|A|$ which has Holevo χ-quantity of the order $n\varepsilon^2$. Holevo's bound then implies that we must have that $n\varepsilon^2$ larger than $|A|$. If A consists of m qubits so $|A| = 2^m$, this bound tells us that the number of copies of the state required is exponential in the number of qubits.

Since we prove bounds where we are concerned with the *scaling* of the relevant quantities and not so much precise values (which are too hard to determine) we use big-O notation to suppress constant factors in the bounds. The notation we use is the following: if f, g are real-valued functions on $\mathbb{N}$ or $\mathbb{R}$, then

$$f(x) = \mathcal{O}(g(x))$$

means that there exists some constant $C > 0$ and a value x_0 such that for all $x > x_0$

$$|f(x)| \leq Cg(x).$$

For lower bounds we write

$$f(x) = \Omega(g(x))$$

if there exists a constant $C > 0$ and a value x_0 such that for all $x > x_0$

$$|f(x)| \geq Cg(x).$$

To make this approach work, we need to find a large set of states which are all at least distance ε away from each other and give an ensemble with small Holevo χ-quantity. This part is a bit technical; you can ignore the proof at this point. The main idea we want to illustrate is to use this set of states and apply the Holevo bound, relating the state learning problem to a communication scenario. We will use the fact that there exists a large set of states which do not have large overlap.

Lemma 10.13. *Given a quantum system A there exists a set of states $|\psi_x\rangle \in \mathcal{H}_A$ for $x = 0, \dots, |X| - 1$ such that*

$$|\langle \psi_x | \psi_{x'} \rangle| \leq \frac{1}{2} \quad \textit{for all } x \neq x'$$

and

$$\log(|X|) = \Omega(|A|).$$

Perhaps we will prove this fact later in the course. For now, we use it to prove the following.

Lemma 10.14. *There exists a collection of pure states $\rho_{A,x} \in \mathrm{S}(A)$ such that*

$$T(\rho_{A,x}, \rho_{A,x'}) \geq \varepsilon \quad \textit{for all } x \neq x'$$

where $\log(|X|) = \Omega(|A|)$. This set is such that when taking the uniform ensemble, the Holevo χ-quantity is upper bounded as

$$\chi(\{|X|^{-1}, \rho_{A,x}\}) = \mathcal{O}\left(\varepsilon^2 \log \frac{|A|}{\varepsilon}\right).$$

Proof. Choose a basis $|a\rangle$, $a = 0, \dots, |A| - 1$ for $\mathcal{H}_A$. Apply Lemma 10.13 to the subspace $\mathcal{H}_{A'}$ spanned by $|a\rangle$ for $a > 0$, giving a collection of $|X| = 2^{C(|A|-1)}$ states $|\psi_x\rangle$. Let $\rho_{A,x}$ be given by the pure state

$$|\phi_x\rangle = \sqrt{1 - 2\varepsilon^2}|0\rangle + \sqrt{2}\varepsilon|\psi_x\rangle.$$

Then for $x \neq x'$

$$|\langle \phi_x | \phi_{x'} \rangle| = |1 - 2\varepsilon^2 + 2\varepsilon^2 \langle \phi_x | \phi_{x'} \rangle| \leq 1 - \varepsilon^2$$

since $|\langle \phi_x | \phi_{x'} \rangle| \leq \frac{1}{2}$. We then bound the trace distance as

$$\begin{aligned} T(\rho_{A,x}, \rho_{A,x'}) &= \sqrt{1 - |\langle \phi_x | \phi_{x'} \rangle|^2} \\ &\geq \sqrt{1 - (1 - \varepsilon^2)^2} = \sqrt{2\varepsilon^2 - \varepsilon^4} \geq \varepsilon. \end{aligned}$$

Next, we need to bound the Holevo χ-quantity of the ensemble which consists of a uniform mixture of the states $\rho_{A,x}$. Since all the states of the ensemble are pure, $H(\rho_{A,x}) = 0$ and

$$\chi(\{|X|^{-1}, \rho_{A,x}\}) = H(\rho_A) \quad \rho_A = \frac{1}{|X|} \sum_x \rho_{A,x}.$$

The entropy increases under unital channels (channels Φ_A which are such that $\Phi_A(\mathbb{1}_A) = \mathbb{1}_A$), as we will see below in Theorem 10.19. We apply this to the channel Φ_A defined by

$$\Phi_A(M_A) = \langle 0|M_A|0\rangle|0\rangle\langle 0| + \sum_{a=1}^{|A|-1} \langle a|M_A|a\rangle \tau_{A'}$$

where $\tau_{A'}$ is the maximally mixed state on A'

$$\tau_{A'} = \frac{1}{|A|-1} \sum_{a=1}^{|A|-1} |a\rangle\langle a|.$$

This channel is easily seen to be unital, and when applied to ρ_A we get

$$\sigma_A = \Phi_A(\rho_A) = (1-2\varepsilon^2) + 2\varepsilon^2 \tau_{A'}.$$

It has entropy

$$H(\sigma_A) = -(1-2\varepsilon^2)\log(1-2\varepsilon^2) - 2\varepsilon^2 \log\left(\frac{2\varepsilon^2}{|A|}\right) = h(2\varepsilon^2) + 2\varepsilon^2 \log(|A|).$$

In conclusion,

$$\chi(\{|X|^{-1}, \rho_{A,x}\}) \leq H(\sigma_A) = h(2\varepsilon^2) + 2\varepsilon^2 \log(|A|) = \mathcal{O}\left(\varepsilon^2 \log \frac{|A|}{\varepsilon}\right).$$

□

Theorem 10.15 (Lower bound sample complexity of learning pure states)**.** *Suppose that μ is a measurement on n copies of a system A, which is such that it outputs an estimate $\hat{\rho}_A$ on input $\rho_A^{\otimes n}$ which is such that for all states ρ_A*

$$\Pr(T(\rho_A, \hat{\rho}_A) > \varepsilon) \leq \delta$$

for some $\varepsilon, \delta > 0$. Then

$$n = \Omega\left(\frac{|A|}{\varepsilon^2 \log(|A|/\varepsilon)}\right).$$

Proof. Use Lemma 10.14 (with precision 2ε) to choose a collection of states $\rho_{A,x}$ such that

$$T(\rho_{A,x}, \rho_{A,x'}) \geq 2\varepsilon \quad \text{for all } x \neq x'$$

where $\log(|X|) = \Omega(|A|)$ and where

$$\chi(\{|X|^{-1}, \rho_{A,x}\}) = \mathcal{O}\left(\varepsilon^2 \log \frac{|A|}{\varepsilon}\right).$$

We can now send a classical message in X using n qubits in the following way: Alice encodes a message x by encoding the message as the quantum state $\rho_{A,x}$ and sends $\rho_{A,x}^{\otimes n}$. Bob may now decode by using the measurement μ. The measurement μ gives an estimate $\hat{\rho}_A$. Bob decodes the message as the x' for which $\rho_{A,x'}$ is closest to his estimate. Suppose Bob obtains an ε-accurate estimate of $\rho_{A,x}$; all $\rho_{A,y}$ for $y \neq x$ are at least distance 2ε away from $\rho_{A,x}$ which implies that

Bob finds $x = x'$ and the message gets transmitted correctly. By assumption, the probability that Bob gets an ε-accurate estimate (and hence a correctly decoded message) is at least $1 - \delta$.

Next, we bound the Holevo χ-quantity. By subadditivity, with respect to the state $\rho_{A^n} = |X|^{-1} \sum_x \rho_{A,x}^{\otimes n}$ we have

$$\chi(\{|X|^{-1}, \rho_{A,x}^{\otimes n}\}) = H(A^n) \leq nH(A) = n\chi(\{|X|^{-1}, \rho_{A,x}\}).$$

Combining with Fano's inequality gives

$$n\chi(\{|X|^{-1}, \rho_{A,x}\}) \geq h(\delta) + (1 - \delta)\log(|X| - 1).$$

If we fix any constant δ, and we use that

$$\log(|X|) = \Omega(|A|) \qquad \text{and} \qquad \chi(\{|X|^{-1}, \rho_{A,x}\}) = \mathcal{O}\left(\varepsilon^2 \log \frac{|A|}{\varepsilon}\right)$$

we obtain

$$n = \Omega\left(\frac{|A|}{\varepsilon^2 \log(|A|/\varepsilon)}\right).$$

□

The logarithmic factor $\log(|A|/\varepsilon)$ is small compared to $|A|/\varepsilon$, and can perhaps be removed with a sharper proof.

10.5 The relative entropy

We end this lecture with a final entropic quantity, the *relative entropy*. It is not required for the remainder of the course, but for the sake of completeness we include a discussion (as it does play a major role in further developments of the theory of quantum information). We will see that it offers another reincarnation of strong subadditivity, this time as a *monotonicity* property.

The relative entropy differs from the quantities we have seen before in that it depends on *two* states (or probability distributions) rather than a single one. In the classical case, it is also known as the *Kullback-Leibler divergence*.

Definition 10.16. Suppose p, q are probability distributions on the same set of outcomes. Then the *relative entropy* of p and q is defined as

$$D(p\|q) := \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

if $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$ and otherwise $D(p\|q) = \infty$.

Clearly, if $p = q$ we have $D(p\|q) = 0$. One can also show that in general

$$D(p\|q) \geq 0$$

with equality if and only if $p = q$ as you may show in Exercise 10.8. The relative entropy can be thought of as a measure of how different p and q are. It is not a distance metric however (this is why it is called a *divergence* rather than a distance), as it is not symmetric in p and q, i.e. $D(p\|q)$ need not be the same as $D(q\|p)$.

We next define the relative entropy for quantum states. Note that

$$D(p\|q) = \sum_x p(x) \left(\log(p(x)) - \log(q(x))\right).$$

We define the relative entropy for positive operators, as it is sometimes useful to apply it to operators which do not have normalized trace.

Definition 10.17. For $\rho, \sigma \in \mathrm{PSD}(\mathcal{H})$, the *relative entropy* is defined as

$$D(\rho\|\sigma) = \operatorname{tr}[\rho(\log(\rho) - \log(\sigma))]$$

if $\operatorname{im}(\rho) \subseteq \operatorname{im}(\sigma)$ and otherwise $D(\rho\|\sigma) = \infty$.

Similar to the classical case, the reason for the case distinction is that if $\operatorname{im}(\rho) \subseteq \operatorname{im}(\sigma)$, then $\rho \log(\sigma)$ is a well-defined operator, and otherwise $\rho \log(\sigma)$ is infinite. It is easy to see that the relative entropy is invariant under isometries: if $V \in \mathrm{Isom}(\mathcal{H}, \mathcal{K})$

$$D(V\rho V^\dagger \| V\sigma V^\dagger) = D(\rho\|\sigma).$$

The relative entropy is a 'parent quantity' for other entropic quantities, in the sense that one can deduce the entropic quantities from last lecture as special cases of the relative entropy. For instance, it follows directly from the definition that for $\rho \in \mathrm{S}(\mathcal{H})$

$$H(\rho) = -D(\rho\|\mathbb{1}) = \log(d) - D(\rho\|\tau) \tag{10.8}$$

where $d = \dim(\mathcal{H})$ and $\tau = \frac{\mathbb{1}}{d}$ is the maximally mixed state. If $\rho_{AB} \in \mathrm{S}(AB)$ we may recover the conditional entropy and mutual information as

$$\begin{aligned} H(A|B)_\rho &= -D(\rho_{AB}\|\mathbb{1}_A \otimes \rho_B) \\ I(A:B)_\rho &= D(\rho_{AB}\|\rho_A \otimes \rho_B). \end{aligned} \tag{10.9}$$

In fact, what is also true is that

$$\begin{aligned} H(A|B)_\rho &= -\min_{\sigma_B} D(\rho_{AB}\|\mathbb{1}_A \otimes \sigma_B) \\ I(A:B)_\rho &= \min_{\sigma_A, \sigma_B} D(\rho_{AB}\|\sigma_A \otimes \sigma_B) \end{aligned} \tag{10.10}$$

where one minimizes over states $\sigma_A \in \mathrm{S}(A)$ and $\sigma_B \in \mathrm{S}(B)$.

The most important fact about the relative entropy is that it is monotonic under quantum channels.

Theorem 10.18 (Monotonicity of relative entropy)**.** *If $\Phi_{A\to B} \in \mathrm{C}(A,B)$, and $\rho_A, \sigma_A \in \mathrm{S}(A)$, then*

$$D(\rho_A\|\sigma_A) \geq D(\Phi_{A\to B}(\rho_A)\|\Phi_{A\to B}(\sigma_A)).$$

This fact can be derived from strong subadditivity, as we will see below. On the other hand, it is also easy to see that it implies strong subadditivity, since it directly implies the data processing inequalities for the conditional entropy and mutual information of Theorem 10.4 and Theorem 10.8. The intuition behind the statement is that (as in data processing) applying a quantum channel to both ρ_A and σ_A should never make it easier to distinguish the states, so they become 'closer' to each other and have smaller relative entropy. Before proving Theorem 10.18, let us see a few direct consequences.

Theorem 10.19. *(a) $D(\rho\|\sigma) \geq 0$ for $\rho, \sigma \in \mathrm{S}(\mathcal{H})$ with equality if and only if $\rho = \sigma$.*

(b) If $\Phi_{A\to B} \in \mathrm{C}(A, B)$ is a unital *channel, meaning that $\Phi_{A\to B}(\mathbb{1}_A) = \mathbb{1}_B$,*

$$H(\Phi_{A\to B}(\rho_A)) \geq H(\rho_A)$$

for all $\rho_A \in \mathrm{S}(A)$.

Proof. (a) By applying Theorem 10.18 using the channel which takes the trace over the whole Hilbert space we see that $D(\rho\|\sigma) \geq 0$. Now suppose that $D(\rho\|\sigma) = 0$. Consider an arbitrary measurement, and let p and q denote the outcome probability distributions when measuring respectively ρ and σ. Then by applying Theorem 10.18 with the measurement channel we see that $D(p\|q) = 0$, but we already saw that implied $p = q$. If two states have the same outcome probabilities for any measurement they must be the same, so $\rho = \sigma$.

(b) This follows directly from Eq. (10.8).

□

In order to prove Theorem 10.18, we will prove a relation between the relative entropy and the conditional entropy.

Lemma 10.20. *For any $\rho, \sigma \in \mathrm{S}(AB)$ we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} H(t\rho + (1-t)\sigma) = \operatorname{tr}[(\sigma - \rho)\log(\sigma)].$$

Proof. Suppose that $t \mapsto M(t)$ is a differentiable function from the real numbers to Hermitian matrices with derivative $M'(t)$. The chain rule implies that for a real-valued differentiable function f with derivative f'

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}[f(M(t))] = \operatorname{tr}[f'(M(t))M'(t)].$$

So, using the derivative of $f(x) = -x\log(x)$ which has derivative $-\log(x) - 1$ and $M(t) = t\rho + (1-t)\sigma$ which has derivative $\rho - \sigma$ we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(t\rho + (1-t)\sigma) &= \frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}[f(M(t))] \\ &= -\operatorname{tr}[(\rho - \sigma)\log(t\rho + (1-t)\sigma)] + \underbrace{\operatorname{tr}[\rho - \sigma]}_{=0}\end{aligned}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} H(t\rho + (1-t)\sigma) = \operatorname{tr}[(\sigma - \rho)\log(\sigma)].$$

□

We will prove Theorem 10.18 by appealing to the following fact which is, once again, an incarnation of strong subadditivity.

Theorem 10.21. *The function $\rho_{AB} \mapsto H(A|B)_\rho$ is a concave function on $\mathrm{S}(AB)$.*

Proof. Let p_X be a probability distribution, and let $\rho_{AB,x}$ be a collection of states, and let

$$\rho_{XAB} = \sum_x p_X(x)|x\rangle\langle x| \otimes \rho_{AB,x}.$$

Note that $\rho_{AB} = \sum_x p_X(x)\rho_{AB,x}$. Then, by data processing

$$H(A|XB)_\rho \leq H(A|B)_\rho$$

which may be rewritten using Lemma 9.5 as

$$\sum_x p_X(x) H(A|B)_{\rho_{AB,x}} \leq H(A|B)_{\rho_{AB}}$$

which means that the conditional entropy is concave. □

Proof of Theorem 10.18. By the invariance of relative entropy under isometries, by applying a Stinespring extension it suffices to show monotonicity for the partical trace. That is, we need to show that for $\rho_{AB}, \sigma_{AB} \in \mathrm{S}(AB)$

$$D(\rho_{AB}\|\sigma_{AB}) \geq D(\rho_A\|\sigma_A).$$

From Lemma 10.20 and simply writing out all terms one sees that

$$D(\rho_{AB}\|\sigma_{AB}) - D(\rho_A\|\sigma_A) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} H(B|A)_{t\rho+(1-t)\sigma} - H(B|A)_\rho + H(B|A)_\sigma. \tag{10.11}$$

By Theorem 10.21

$$H(B|A)_{t\rho+(1-t)\sigma} \geq tH(B|A)_\rho + (1-t)H(B|A)_\sigma$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} H(B|A)_{t\rho+(1-t)\sigma} \geq H(B|A)_\rho - H(B|A)_\sigma$$

so from Eq. (10.11) we conclude that

$$D(\rho_{AB}\|\sigma_{AB}) - D(\rho_A\|\sigma_A) \geq 0.$$

□

Outlook

The discussion of lower bounds for learning quantum states is based on [19]. We have given the bound for *pure* states. More generally, if one does not restrict to pure states but states of rank r then the argument can be adapted to give a sample complexity lower bound scaling with $r|A|/\varepsilon^2$ (ignoring logarithmic factors). In particular, without any rank constraint this gives a lower bound for the sample complexity of $|A|^2/\varepsilon^2$. In the same work the authors derive (almost) matching upper bounds by proposing a specific measurement.

Rényi entropies

The Shannon and von Neumann entropy are part of a bigger family of entropic quantities, the so called Rényi entropies.

Definition 10.22. For $\alpha \in [0,1) \cup (1,\infty)$ the α-th Rényi entropy is given by

$$H_\alpha(p) := \frac{1}{1-\alpha} \log\left(\sum_x p(x)^\alpha\right).$$

Moreover,

$$H_1(p) = H(p) \quad \text{and} \quad H_\infty(p) = -\log(\max_x p(x))$$

If $\rho \in \mathrm{S}(\mathcal{H})$, and $\alpha \in (0,1) \cup (1,\infty)$ we define

$$H_\alpha(\rho) = \frac{1}{1-\alpha} \log(\mathrm{tr}[\rho^\alpha])$$

and we define $H_1(\rho) = H(\rho)$ and $H_\infty(\rho) = -\log(\|\rho\|_\infty)$.

Note that this is consistent with $H_0(p) = \log(|\mathrm{supp}(p)|)$ and $H_0(\rho) = \log(\mathrm{rank}(\rho))$. We saw that for the task of compression, the asymptotic rate was computed by the Shannon or von Neumann entropy, whereas the one-shot optimal result was given by a (smoothed) Rényi entropy (note that indeed $\alpha = 0$). There are also information processing tasks where the asymptotic rate is for instance a conditional entropy, and in that case the one-shot task should be characterized by a Rényi conditional entropy. One could (naively) define the α-th conditional Rényi entropy as $H_\alpha(A|B) = H_\alpha(AB) - H_\alpha(B)$. It turns out however, that (even in the classical setting) this is not the right definition! Moreover, the 'correct' definition is not unique and depends on the task at hand!

A good way to define Rényi conditional entropies is by using the relative entropy as in Eq. (10.9). Of course, this just reduces the task to defining an appropriate Rényi version of the relative entropy. The classical Rényi relative entropy (or Rényi divergence) is defined as

$$D_\alpha(p\|q) = \frac{1}{\alpha-1} \log\left(\sum_x p(x)^\alpha q(x)^{1-\alpha}\right)$$

The nonuniqueness of generalizing to the quantum setting is that ρ and σ do not commute. Two possible generalizations (which are equivalent for commuting states) are the so-called *hypothesis testing Rényi divergence*

$$D_\alpha^{(h)}(\rho\|\sigma) = \frac{1}{\alpha-1} \log\bigl(\mathrm{tr}[\rho^\alpha \sigma^{1-\alpha}]\bigr)$$

and the *sandwiched Rényi divergence*

$$D_\alpha^{(s)}(\rho\|\sigma) = \frac{1}{\alpha-1} \log\Bigl(\mathrm{tr}[(\sigma^{\frac{1-\alpha}{2\alpha}} \rho \sigma^{\frac{1-\alpha}{2\alpha}})^\alpha]\Bigr).$$

Given a choice of Rényi divergence D_α (for instance $D_\alpha = D_\alpha^{(s)}$) a reasonable definition of a conditional Rényi entropy is then

$$\tilde{H}_\alpha(A|B)_\rho = D_\alpha(\rho_{AB}\|\mathbb{1}_A \otimes \rho_B).$$

However, based on Eq. (10.10) one often defines

$$H_\alpha(A|B)_\rho = -\min_{\sigma_B\in \mathrm{S}(B)} D_\alpha(\rho_{AB}\|\mathbb{1}_A\otimes\sigma_B). \tag{10.12}$$

We will not go into further detail on the properties of the zoo of one-shot quantities that can be obtained in this way and their interpretations. The only take-away messages of the above discussion is that there is a systematic approach to constructing various versions of the Rényi conditional entropy, and that which definition one should use may depend on the specific task. In other words, while asymptotically many information processing tasks are characterized by the same entropic quantities, in the one-shot regime one may have to consider different quantities even though the asymptotic answers are equal.

10.6 Exercises

10.1 **Computing mutual information and conditional entropies:** Let's practice computing some entropic quantities!

(a) Let $|\psi_{AB}\rangle$ be the two-qubit state $\frac{1}{\sqrt{2}}(|{+}{-}\rangle - |{-}{+}\rangle)$ shared between Alice and Bob. Compute $H(A|B)$ and $I(A:B)$.

(b) Alice and Bob measure in the basis $|+\rangle, |-\rangle$. What are the conditional entropy and mutual information of the probability distributions of their measurement outcomes?

(c) Now consider the 3-party GHZ state $\tau_{ABC} = |\mathrm{GHZ}\rangle\langle\mathrm{GHZ}|_{ABC}$, where

$$|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}\big(|000\rangle + |111\rangle\big)\ .$$

Compute $H(ABC)_\tau$, $H(A|B)_\tau$, and $H(A|BC)_\tau$.

10.2 **Data processing conditional entropy:** Consider the two-qubit maximally entangled state $\sigma_{AB} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|$.

(a) Consider the depolarizing channel $\mathcal{D}_p : \mathrm{Lin(A)} \to \mathrm{Lin(A)}$, given by

$$M_A \mapsto (1-p)M_A + p\,\mathrm{tr}[M_A]\frac{\mathbb{1}_A}{2}.$$

Compute $H(B|A)$ for the quantum state $\rho_{AB} = (\mathcal{D}_p\otimes\mathcal{I}_B)(\sigma_{AB})$ for $p = 0.25$.

(b) Consider the dephasing channel $\mathcal{P}_p : \mathrm{Lin(A)} \to \mathrm{Lin(A)}$, given by

$$M_A \mapsto (1-p)M_A + p\sum_{a\in\{0,1\}}\langle a|M_A|a\rangle|a\rangle\langle a|.$$

Compute $H(B|A)$ for the quantum state $\rho_{AB} = (\mathcal{P}_p\otimes\mathcal{I}_B)(\sigma_{AB})$ for $p = 0.25$.

(c) Consider the erasure channel $\mathcal{E}_p : \mathrm{Lin(A)} \to \mathrm{Lin(A')}$, where $\mathcal{H}'_A = \mathcal{H}_A\oplus\mathrm{span}\{|\perp\rangle\}$, given by

$$M_A \mapsto (1-p)M_A + p\,\mathrm{tr}[M_A]|\perp\rangle\langle\perp|.$$

Compute $H(B|A')$ for the quantum state $\rho_{A'B} = (\mathcal{E}_p\otimes\mathcal{I}_B)(\sigma_{AB})$ for $p = 0.25$.

(d) Conclude that the data processing inequality strictly holds for the conditional entropy in all three above cases.

10.3 **Information in product states:** Show that if $\rho_{AB} \in \mathrm{S}(AB)$ and $\sigma_{CD} \in \mathrm{S}(CD)$, for the product state $\rho_{AB} \otimes \sigma_{CD}$ the conditional entropy and mutual information are additive:

$$H(AC|BD)_{\rho\otimes\sigma} = H(A|B)_\rho + H(C|D)_\sigma$$

and

$$I(AC : BD)_{\rho\otimes\sigma} = I(A : B)_\rho + I(C : D)_\sigma.$$

10.4 **Negative conditional entropy:** Negative conditional entropy is associated with entanglement, as you will confirm in this exercise.

(a) Show that a *pure* state $\rho_{AB} \in \mathrm{S}(AB)$ has $H(A|B)_\rho < 0$ if and only if it is entangled.
(b) Let $\rho_{AB} \in \mathrm{S}(AB)$ be a separable state. Argue that there exists a classical-quantum state σ_{AX} such that ρ_{AB} is obtained by applying a quantum channel $\Phi_{X\to B}$ to σ_{AX}:

$$\rho_{AB} = (\mathcal{I}_A \otimes \Phi_{X\to B})(\sigma_{AX}).$$

(c) Use this to show that if $\rho_{AB} \in \mathrm{S}(AB)$ has negative conditional entropy $H(A|B)_\rho < 0$, the state ρ_{AB} must be entangled.
(d) Not all entangled states have negative conditional entropy. Give an example of a state ρ_{AB} that is entangled but has positive conditional entropy $H(A|B)_\rho > 0$. *Hint: use Exercise 10.3.*

10.5 **Quantum relative entropy:** Define the single-qubit states ρ, σ, τ by

$$\rho = \frac{2}{3}|0\rangle\langle 0| + \frac{1}{3}|1\rangle\langle 1| \ , \quad \sigma = |0\rangle\langle 0| \ , \quad \tau = \frac{1}{2}\mathbb{1} \ .$$

Compute $D(\rho\|\tau)$, $D(\sigma\|\tau)$, $D(\rho\|\sigma)$, and $D(\sigma\|\rho)$. Deduce that the relative entropy is not symmetric, and does not satisfy the triangle inequality.

10.6 **Concavity of conditional entropy:** In Theorem 10.21 we have shown that the conditional entropy is concave. In this exercise you will derive an upper bound to how concave the conditional entropy can be. Let p_X be a probability distribution, and let $\rho_{AB,x}$ be a collection of states, and let

$$\rho_{XAB} = \sum_x p_X(x)|x\rangle\langle x| \otimes \rho_{AB,x}.$$

Show that

$$H(A|B)_\rho \leq \sum_x p_X(x) H(A|B)_{\rho_{AB,x}} + H(p_X).$$

Hint: use Exercise 9.5.

10.7 **Properties of relative entropy:**

(a) Show that if $\rho_A, \sigma_A \in \mathrm{S}(A)$ and $\rho_B, \sigma_B \in \mathrm{S}(B)$ we have

$$D(\rho_A \otimes \rho_B \| \sigma_A \otimes \sigma_B) = D(\rho_A\|\sigma_A) + D(\rho_B\|\sigma_B).$$

(b) Verify Eq. (10.8) and Eq. (10.9).

(c) Suppose that ρ_{XA} and σ_{XA} are classical-quantum states, so

$$\rho_{XA} = \sum_x p_X(x)|x\rangle\langle x| \otimes \rho_{A,x} \qquad \sigma_{XA} = \sum_x p_X(x)|x\rangle\langle x| \otimes \sigma_{A,x}$$

for a probability distribution p_X and collections of states $\rho_{A,x}, \sigma_{A,x} \in \mathrm{S}(A)$. Show that

$$D(\rho_{XA}\|\sigma_{XA}) = \sum_x p_X(x) D(\rho_{A,x}\|\sigma_{A,x}).$$

(d) Show that the the relative entropy is *jointly convex*, meaning that for $\rho_1, \rho_2, \sigma_1, \sigma_2 \in \mathrm{S}(\mathcal{H})$ and $p \in [0,1]$ we have

$$D(p\rho_1 + (1-p)\rho_2 \| p\sigma_1 + (1-p)\sigma_2) \leq pD(\rho_1\|\sigma_1) + (1-p)D(\rho_2\|\sigma_2).$$

Hint: use monotonicity of the relative entropy.

10.8 **Positivity of classical relative entropy:** Show that for two probability distributions p and q on the same set of outcomes

$$D(p\|q) \geq 0$$

with equality if and only if $p = q$.

10.9 **Measurements increase entropy:** Show that if p is the distribution of outcomes upon measuring ρ using a basis measurement, then $H(\rho) \geq H(p)$.

10.10 **Strong subadditivity from monotonicity:** You have seen in lectures that the monotonicity of the relative entropy follows from strong subadditivity; here we consider the other direction.

Starting from the expression for the conditional entropy of a tripartite system,

$$H(A|BC)_\rho = -D(\rho_{ABC}\|\mathbb{1}_A \otimes \rho_{BC}) \ ,$$

use the monotonicity of D under quantum channels to derive strong subadditivity:

$$H(ABC)_\rho + H(B)_\rho \leq H(AB)_\rho + H(BC)_\rho \ .$$

10.11 **Different versions of strong subadditivity:** In this lecture, we deduced data processing inequalities from strong subadditivity of the von Neumann entropy. Show that conversely, the data processing inequality for the mutual information implies strong subadditivity.

10.12 **Rényi entropies:** You will prove some important properties of Rényi entropies in this exercise. Let $\rho \in \mathrm{S}(\mathcal{H})$.

(a) Show that

$$\lim_{\alpha \to 1} H_\alpha(\rho) = H(\rho).$$

(b) Show that

$$\lim_{\alpha \to \infty} H_\alpha(\rho) = -\log(\|\rho\|_\infty) = H_\infty(\rho).$$

(c) Show that for $0 \leq \alpha < \beta$ and $\rho \in \mathrm{S}(\mathcal{H})$ we have $H_\alpha(\rho) \geq H_\beta(\rho)$. In particular, this shows

$$-\log(\|\rho\|_\infty) \leq H_\alpha(\rho) \leq \log(\mathrm{rank}(\rho)).$$

10.13 **Continuity of conditional entropy:** The aim of this exercise is to prove Theorem 10.9. Let $\rho_{AB}, \sigma_{AB} \in S(AB)$ be states with trace distance $\epsilon = T(\rho_{AB}, \sigma_{AB})$.

(a) Prove that there exist states $\omega_{AB}, \rho'_{AB}, \sigma'_{AB}$ such that

$$(1+\epsilon)\omega_{AB} = \rho_{AB} + \epsilon\rho'_{AB} = \sigma_{AB} + \epsilon\sigma'_{AB}\ .$$

(b) Prove that

$$H(A|B)_\omega \leq h\left(\frac{\epsilon}{1+\epsilon}\right) + \frac{1}{1+\epsilon}H(A|B)_\rho + \frac{\epsilon}{1+\epsilon}H(A|B)_{\rho'}\ .$$

Hint: use Exercise 10.6.

(c) Use the concavity of the conditional entropy to show that

$$H(A|B)_\omega \geq \frac{1}{1+\epsilon}H(A|B)_\sigma + \frac{\epsilon}{1+\epsilon}H(A|B)_{\sigma'}\ .$$

(d) Conclude that

$$|H(A|B)_\rho - H(A|B)_\sigma| \leq (1+\epsilon)h\left(\frac{\epsilon}{1+\epsilon}\right) + 2\epsilon\log|A|\ .$$

10.14 **The Pinsker inequality:**

(a) Let $\mathbf{X}$ and $\mathbf{Y}$ be Bernoulli random variables with probabilities p and q respectively, where $p, q \in [0,1]$. Show that

$$f(p,q) := D(\mathbf{X}\|\mathbf{Y}) - \frac{1}{2}T(\mathbf{X},\mathbf{Y})^2 = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - 2(p-q)^2\ ,$$

where $T(\mathbf{X},\mathbf{Y}) = \sum_x |p_X(x) - p_Y(x)|$ is the statistical distance between $\mathbf{X}$ and $\mathbf{Y}$.

(b) Keeping p constant, show that f is a convex function of q which is minimised when $p = q$. Hence prove the classical Pinsker inequality:

$$D(\mathbf{X}\|\mathbf{Y}) \geq \frac{1}{2}T(\mathbf{X},\mathbf{Y})^2\ .$$

(c) Now let $\rho, \sigma \in S(A)$ be quantum states, and let $\{\mu_\rho, \mu_\sigma\}$ be a POVM which optimally distinguishes between ρ and σ. Now let $p = \mathrm{tr}[\mu_\rho\rho]$ and $q = \mathrm{tr}[\mu_\rho\sigma]$, with X and Y defined as before. Show that

$$T(\rho,\sigma) = T(\mathbf{X},\mathbf{Y})\ .$$

(d) Use the monotonicity of the relative entropy under the measurement channel (5.5) to deduce the quantum Pinsker inequality

$$D(\rho\|\sigma) \geq \frac{1}{2}T(\rho,\sigma)^2\ .$$

10.15 **Correlations in bipartite systems:**

(a) Let $\rho_{AB} \in S(AB)$ be a bipartite state. Use Exercise 10.14 to show that

$$I(A:B)_\rho \geq \frac{1}{2} T(\rho_{AB}, \rho_A \otimes \rho_B)^2 \ .$$

Deduce that $H(AB)_\rho = H(A)_\rho + H(B)_\rho$ if and only if ρ_{AB} is a product state.

(b) Let $\{\mu_{A,0}, \mu_{A,1}\}$, $\{\mu_{B,0}, \mu_{B,1}\}$ be two-outcome POVMs on the A and B systems respectively, whose measurement outcomes give rise to random variables $\mathbf{X}$ and $\mathbf{Y}$ in $\{0,1\}$. Show that

$$I(A:B)_\rho \geq \frac{1}{2} \mathrm{Cov}(\mathbf{X}, \mathbf{Y})^2 \ ,$$

where Cov is the *covariance* defined by

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}(\mathbf{X}\mathbf{Y}) - \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{Y}) \ .$$

Lecture 11

Symmetry and randomness

In information theory it is often the case that *random* constructions (chosen according to an appropriate distribution) perform well. For instance, if one wants to send information over a noisy classical channel, a *random code* performs well with high probability.

There is a close relation between randomness and symmetry, which essentially comes to down to the following: a uniformly random state is invariant under unitary transformations. In this lecture we will introduce two symmetries acting on quantum states. We relate this to notions of uniform randomness. As an application, we show how to perform optimal learning of an unknown pure quantum state.

11.1 Two symmetries

The mathematical framework for studying symmetries is *group theory*. We will not be too formal about group theory: a group is a set with a multiplication rule and which contains inverses. We need two main examples.

Example 11.1. Given $\mathcal{H} = \mathbb{C}^d$ we have the group of unitary matrices $\mathrm{U}(d)$. It is a group, since the product of unitary matrices is again unitary, and the inverse of a unitary matrix is also unitary.

Example 11.2. A *permutation* is a bijective (one-to-one) map

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}.$$

The collection of all permutations of $\{1, \ldots, n\}$ forms a group, known as the *symmetric group* S_n. The multiplication operation here is composition of maps, so $\sigma\pi$ is the map which sends $i \mapsto \sigma(\pi(i))$.

Quantum states live on Hilbert spaces; informally speaking a symmetry is a group that acts on a Hilbert space. The two groups above have natural actions on Hilbert spaces.

Clearly, the group $\mathrm{U}(A)$ acts on $\mathcal{H}_A$. If we take n copies of this quantum system, then we get a system A^n with Hilbert space $\mathcal{H}_A^{\otimes n}$. We have an action of the unitary group $\mathrm{U}(A)$ on this Hilbert space by

$$U \mapsto U^{\otimes n}$$

which just applies the same unitary to each copy of A. There is also a natural action of the symmetric group, which simply permutes the copies of $\mathcal{H}_A$ according to the permutation. For each $\pi \in S_n$ we define $R_\pi \in \mathrm{Lin}(A^n)$ by

$$R_\pi|\phi_1\rangle \dots |\phi_n\rangle = |\phi_{\pi^{-1}(1)}\rangle \dots |\phi_{\pi^{-1}(n)}\rangle.$$

The inverse is needed to make sure that this plays nice with composition of permutations:

$$\begin{aligned} R_\sigma R_\pi|\phi_1\rangle \dots |\phi_n\rangle &= R_\sigma|\phi_{\pi^{-1}(1)}\rangle \dots |\phi_{\pi^{-1}(n)}\rangle \\ &= |\phi_{\pi^{-1}(\sigma^{-1}(1))}\rangle \dots |\phi_{\pi^{-1}(\sigma^{-1}(n))}\rangle \\ &= |\phi_{(\sigma\pi)^{-1}(1)}\rangle \dots |\phi_{(\sigma\pi)^{-1}(n)}\rangle = R_{\sigma\pi}. \end{aligned}$$

These two actions (of $\mathrm{U}(A)$ and S_n) on A^n are in a way 'complementary'. First, we note that for any $U \in \mathrm{U}(A)$ and $\pi \in S_n$, the two actions commute: $[U^{\otimes n}, R_\pi] = 0$. In other words, first applying the same unitary to each copy and then permuting has the same effect as first permuting the copies and then applying the same unitary to each copy. There is a powerful fact that is a converse to this observation: if an operator commutes with all unitaries of the form $U^{\otimes n}$, then it must be a linear combination of permutations:

Theorem 11.3. *An operator $M \in \mathrm{Lin}(A^n)$ is such that*

$$[U^{\otimes n}, M] = 0 \qquad \textit{for all } U \in \mathrm{U}(A)$$

if and only if

$$M = \sum_{\pi \in S_n} \alpha_\pi R_\pi \qquad \alpha_\pi \in \mathbb{C}.$$

We will not prove Theorem 11.3. One way in which it can be used is the following: suppose that we are given M which is such that M commutes with $U^{\otimes n}$ for all unitaries U. Then we know it must be a linear combination of operators R_π. The set of operators R_π is closed under adjoints (since $R_\pi^\dagger = R_{\pi^{-1}}$). This means that using the Hilbert-Schmidt inner product, the coefficients α_π are completely determined by the numbers

$$\gamma_\sigma(M) = \mathrm{tr}[R_\sigma^\dagger M].$$

However, the operators R_π are *not* orthonormal with respect to the Hilbert-Schmidt inner product.

Here are the two simplest examples (corresponding $n = 1, 2$), which you can verify in Exercise 11.2. If $M \in \mathrm{Lin}(\mathcal{H})$ is such that $[M, U] = 0$ for all $U \in \mathrm{U}(A)$, then

$$M = \frac{\mathrm{tr}[M]}{|A|}\mathbb{1}. \tag{11.1}$$

For $n = 2$ there are two different permutations, the identity permutation and the swap permutation. The corresponding operators R_π acting on $\mathcal{H}_A^{\otimes 2}$ are the identity and the swap operator F, which acts as

$$F|\phi\rangle|\psi\rangle = |\psi\rangle|\phi\rangle.$$

Now suppose that $M \in \mathrm{Lin}(A^2)$ is such that $[M, U^{\otimes 2}] = 0$ for all $U \in \mathrm{U}(A)$, then

$$M = \alpha\mathbb{1} + \beta F \tag{11.2}$$

where

$$\alpha = \frac{1}{|A|^3 - |A|}(|A|\,\mathrm{tr}[M] - \mathrm{tr}[FM]) \qquad \beta = \frac{1}{|A|^3 - |A|}(|A|\,\mathrm{tr}[FM] - \mathrm{tr}[M]).$$

11.1.1 The symmetric subspace

One of the main reasons to consider the actions of $\mathrm{U}(A)$ and S_n described above is to analyze situations where we have many copies of a single state. This is relevant to study asymptotics (as in Lecture 8 when we studied the asymptotics of compression) or to analyze an information processing scenario with a finite number of copies (such as when we investigated the sample complexity of learning quantum states). Clearly, when we have a state $|\psi\rangle^{\otimes n}$ this is *invariant* when we apply any permutation R_π for $\pi \in S_n$. Similarly, if $\rho \in \mathrm{S}(A)$, we have $[R_\pi, \rho^{\otimes n}] = 0$.

It is useful to consider the subspace of all vectors $|\Phi\rangle \in \mathcal{H}_A^{\otimes n}$ which are left invariant by R_π; this subspace is known as the *symmetric subspace*:

$$\mathrm{Sym}_n(A) := \{|\Phi\rangle \in \mathcal{H}_A^{\otimes n} : R_\pi|\Phi\rangle = |\Phi\rangle \text{ for all } \pi \in S_n\}.$$

It is clear that this defines a subspace. We let Π_n denote the orthogonal projection onto $\mathrm{Sym}_n(A) \subseteq \mathcal{H}_A^{\otimes n}$.

Lemma 11.4. *The dimension of the symmetric subspace is*

$$\dim \mathrm{Sym}_n(A) = \binom{n+|A|-1}{n}.$$

The projection onto the symmetric subspace is

$$\Pi_n = \frac{1}{n!} \sum_{\pi \in S_n} R_\pi. \tag{11.3}$$

Proof. We start by proving Eq. (11.3). Let P be the right-hand side of Eq. (11.3). We note that for any $\sigma \in S_n$, the map $\pi \mapsto \sigma\pi$ defines a bijection of S_n. This implies that

$$R_\sigma P = \frac{1}{n!} \sum_{\pi \in S_n} \underbrace{R_\sigma R_\pi}_{=R_{\sigma\pi}} = \frac{1}{n!} \sum_{\pi' \in S_n} R_{\pi'} = P$$

and therefore

$$\begin{aligned} P^2 &= \frac{1}{n!} \sum_{\sigma \in S_n} R_\sigma P \\ &= \frac{1}{n!} \sum_{\sigma \in S_n} P = P. \end{aligned}$$

Similarly,

$$P^\dagger = \frac{1}{n!} \sum_{\pi \in S_n} R_\pi^\dagger = \frac{1}{n!} \sum_{\pi \in S_n} R_{\pi^{-1}} = P.$$

We conclude that P is a projection. Next, we observe that (by definition) if we have $|\Phi\rangle \in \mathrm{Sym}_n(A)$, we have $P|\Phi\rangle = |\Phi\rangle$ (so the image of P contains $\mathrm{Sym}_n(A)$). On the other hand, since $R_\sigma P = P$ for all $\sigma \in S_n$, the image of P is invariant under the action by S_n, so the image of P is contained in $\mathrm{Sym}_n(A)$. We conclude that P is the projection operator onto $\mathrm{Sym}_n(A)$.

Choose a basis $|a\rangle$ for $a = 0, \dots, |A|-1$ for A. The set

$$\Pi_n|a_1 \dots a_n\rangle$$

spans $\mathrm{Sym}_n(A)$. Because we symmetrize, this only depends on the collection $\{a_1, \ldots, a_n\}$ but not on its order. For that reason, we define for any $\vec{i} = (i_0, \ldots, i_{|A|-1})$ where $i_0 + \cdots + i_{|A|-1} = n$

$$|\vec{i}\rangle = \Pi_n|\underbrace{0\ldots0}_{i_0}\underbrace{1\ldots1}_{i_1}\ldots\rangle$$

These vectors are all nonzero and orthogonal, and they span $\mathrm{Sym}_n(A)$, so the dimension of $\mathrm{Sym}_n(A)$ is given by counting how many such $\vec{i}$ exist. It is a standard exercise in combinatorics to verify that this is $\binom{n+|A|-1}{n}$, which you can do in Exercise 11.3. □

11.2 Random states and unitaries

In information theory it is often useful to pick a *random quantum state* or *random unitary*. We would like to have a (continuous) probability distribution which corresponds to uniformly random states or unitaries. For uniformly random quantum states it is clear how to obtain this: choose $|\psi\rangle$ uniformly at random from the unit sphere in the Hilbert space $\mathcal{H}_A$. This is invariant under unitaries: we have any fixed unitary $U \in \mathrm{U}(A)$, then selecting $|\psi\rangle$ uniformly at random, and applying U to it gives a state $U|\psi\rangle$ which itself has a uniform distribution. Similarly, we would like uniformly random unitaries to be such that choosing U at random, and then applying fixed unitaries $V, W \in \mathrm{U}(A)$ to get a unitary VUW not change the distribution (VUW is again distributed uniformly).

Such a probability distribution exists; it is known as the Haar measure. We will not define the Haar measure very formally, but use the following theorem, which states that the Haar measure is the unique measure satisfying invariance. This serves as a stand-in for a concrete definition.

Theorem 11.5 (Haar measure)**.** *For any Hilbert space $\mathcal{H}$ there exists a unique measure, which we call the* Haar measure*, $\mathrm{d}U$ on $\mathrm{U}(\mathcal{H})$ such that for any continuous function $f : \mathrm{U}(\mathcal{H}) \to \mathbb{C}$ and any $V, W \in \mathrm{U}(\mathcal{H})$*

$$\int_{\mathrm{U}(\mathcal{H})} f(WUV)\mathrm{d}U = \int_{\mathrm{U}(\mathcal{H})} f(U)\mathrm{d}U$$

and which is normalized by

$$\int_{\mathrm{U}(\mathcal{H})} 1\mathrm{d}U = 1.$$

If f is some function on the unitary group we denote by $\mathbb{E}_U$ the expectation value with respect to the Haar measure:

$$\mathbb{E}_U f := \int_{\mathrm{U}(\mathcal{H})} f(U)\mathrm{d}U.$$

We can use the Haar measure to define uniformly random quantum states, simply by fixing some state $|0\rangle$ and applying a random unitary U to get a state $|\psi\rangle = U|0\rangle$. This gives a measure on pure states which is such that

$$\int_{\mathrm{S}(A)} f(U\psi)\mathrm{d}\psi = \int_{\mathrm{S}(A)} f(\psi)\mathrm{d}\psi$$

for any $U \in \mathrm{U}(A)$, and

$$\int_{\mathrm{S}(A)} \mathrm{d}\psi = 1.$$

As an application of integrating over random states we give an alternative integral expression for the projection onto the symmetric subspace.

Lemma 11.6. *The integral over $|\psi\rangle\langle\psi|^{\otimes n}$ is proportional to the projection onto* $\mathrm{Sym}_n(A)$*:*

$$\binom{n+d-1}{n} \int_{\mathrm{S}(A)} |\psi\rangle\langle\psi|^{\otimes n} \mathrm{d}\psi = \Pi_n.$$

Proof. Let

$$Q_n = \binom{n+d-1}{n} \int_{\mathrm{S}(A)} |\psi\rangle\langle\psi|^{\otimes n} \mathrm{d}\psi$$

By unitary invariance, for any $U \in \mathrm{U}(A)$

$$\begin{aligned} U^{\otimes n} Q_n &= \binom{n+d-1}{n} \int_{\mathrm{S}(A)} (U|\psi\rangle\langle\psi|)^{\otimes n} \mathrm{d}\psi \\ &= \binom{n+d-1}{n} \int_{\mathrm{S}(A)} (U|\psi\rangle\langle\psi|U^\dagger)^{\otimes n} \mathrm{d}\psi U^{\otimes n} \\ &= \binom{n+d-1}{n} \int_{\mathrm{S}(A)} |\psi\rangle\langle\psi|^{\otimes n} \mathrm{d}\psi U^{\otimes n} = Q_n U^{\otimes n} \end{aligned}$$

Since $[U^{\otimes n}, Q_n] = 0$ for all $U \in \mathrm{U}(A)$ Theorem 11.3 tells us that Q_n is a linear combination of the operators R_π

$$Q_n = \sum_{\pi \in S_n} \alpha_\pi R_\pi.$$

We need to compute the coefficients α_π. It is clear that $R_\pi Q_n = Q_n$ for all $\pi \in S_n$. This means that

$$\gamma_\pi(Q_n) = \mathrm{tr}[R_\pi^\dagger Q_n] = \mathrm{tr}[Q_n]$$

is constant (independent of π). The same is true for Π_n: since $R_\pi \Pi_n = \Pi_n$ for all $\pi \in S_n$

$$\gamma_\pi(\Pi_n) = \mathrm{tr}[R_\pi^\dagger \Pi_n] = \mathrm{tr}[\Pi_n].$$

Since these values uniquely determine the operator, it now suffices to check that $\mathrm{tr}[Q_n] = \mathrm{tr}[\Pi_n]$. The trace of Q_n is given by

$$\mathrm{tr}[Q_n] = \binom{n+d-1}{n} \int_{\mathrm{S}(A)} \underbrace{\mathrm{tr}\left[|\psi\rangle\langle\psi|^{\otimes n}\right]}_{=1} \mathrm{d}\psi = \binom{n+d-1}{n}$$

by the normalization of the integral. This equals the dimension of the symmetric subspace by Lemma 11.4 and hence the trace of Π_n. □

11.3 Optimal learning of pure quantum states

Recall the task of *learning a quantum state* from Lecture 10: given n copies of a state $|\psi\rangle \in \mathcal{H}_a$ we would like to perform a measurement and return an estimate $|\hat{\psi}\rangle$ which is close to $|\psi\rangle$. We will now describe a concrete measurement which performs this task. It will be a measurement with a continuous set of outcomes, so in a short intermezzo we will review how to extend the notions of measurements to have outcomes in a continuum of values. Recall that we defined a measurement on A with outcomes in X to be a collection of positive operators $\mu(x) \in \mathrm{PSD}(A)$ such that $\sum_x \mu(x) = \mathbb{1}_A$. If X is a continuous set of outcomes with an integration measure we define a measurement to be a collection $\mu(x) \in \mathrm{PSD}(A)$ such that $x \to \mu(x)$ is integrable and

$$\int_X \mu(x) \mathrm{d}x = \mathbb{1}_A.$$

In the case with finite outcomes, the probability of getting outcome x when measuring a state ρ_A was given by $p(x) = \mathrm{tr}[\mu(x)\rho_A]$. In the continuous case, we get a probability distribution where the probability density is given by $p(x) = \mathrm{tr}[\mu(x)\rho_A]$. This means that the probability that $x \in \Omega \subset X$ is given by

$$\Pr(x \in \Omega) = \int_\Omega \mathrm{tr}[\mu(x)\rho_A]\mathrm{d}x$$

and the expectation value of some function f of the outcome x is given by

$$\mathbb{E}f = \int_X f(x)\, \mathrm{tr}[\mu(x)\rho_A]\mathrm{d}x.$$

The idea is that for the state learning problem we take measurement operators proportional to $|\phi\rangle\langle\phi|^{\otimes n}$, so the outcomes range over all pure states. This makes sense, since we would like the measurement to return an estimate of the state. If we let

$$\mu(\phi) = \binom{n + |A| + 1}{n} |\phi\rangle\langle\phi|^{\otimes n} \tag{11.4}$$

then by Lemma 11.6

$$\int \mu(\phi) \mathrm{d}\phi = \Pi_n$$

is the projection Π_n onto the symmetric subspace. This means that we may define a measurement which is given by the operators $\mu(\phi)$ together with $\mathbb{1} - \Pi_n$ (the projection onto the complement of the symmetric subspace) which we assign outcome $\perp$.

Theorem 11.7. *Using the measurement defined by Eq. (11.4) with outcome ϕ on a state $|\psi\rangle^{\otimes n}$, the expected value of the squared overlap is*

$$\mathbb{E}|\langle\phi|\psi\rangle|^2 \geq 1 - \frac{|A|}{n}.$$

Proof. First, observe that we never obtain outcome $\perp$, since $|\psi\rangle^{\otimes n}$ is in the symmetric subspace. The probability distribution of obtaining outcome ϕ is given by

$$\mathrm{tr}[\mu(\phi)|\psi\rangle\langle\psi|^{\otimes n}] = \langle\psi|^{\otimes n}\mu(\phi)|\psi\rangle^{\otimes n}$$

$$= \binom{n+|A|-1}{n} |\langle\psi|\phi\rangle|^{2n}.$$

That means that

$$\mathbb{E}|\langle\phi|\psi\rangle|^2 = \binom{n+|A|-1}{n} \int_{\mathrm{S}(A)} |\langle\phi|\psi\rangle|^{2n+2} \mathrm{d}\phi \tag{11.5}$$

$$\tag{11.6}$$

Now, by the normalization condition for $n+1$ copies we have

$$\begin{aligned}\binom{n+|A|}{n+1} \int_{\mathrm{S}(A)} |\langle\phi|\psi\rangle|^{2n+2} \mathrm{d}\phi &= \binom{n+|A|}{n+1} \int_{\mathrm{S}(A)} \mathrm{tr}[|\phi\rangle\langle\phi|^{\otimes(n+1)} |\psi\rangle\langle\psi|^{\otimes(n+1)}] \mathrm{d}\phi \\ &= \mathrm{tr}[\Pi_{n+1} |\psi\rangle\langle\psi|^{\otimes(n+1)}] = 1.\end{aligned}$$

Combining with Eq. (11.5) we get

$$\begin{aligned}\mathbb{E}|\langle\phi|\psi\rangle|^2 &= \binom{n+|A|}{n+1}^{-1} \binom{n+|A|-1}{n} \\ &= \frac{(n+|A|-1)!(n+1)!(|A|-1)!}{(n+|A|)!n!(|A|-1)!} = \frac{n+1}{n+|A|} = 1 - \frac{|A|-1}{n+|A|} \\ &\geq 1 - \frac{|A|}{n}.\end{aligned}$$

□

Corollary 11.8. *Let $\varepsilon, \delta > 0$. There exists a measurement μ on $n = \mathcal{O}(\frac{|A|}{\varepsilon^2})$ copies of A such that given $|\psi^{\otimes n}\rangle$, the measurement returns with probability at least $1-\delta$ an estimate $|\hat{\psi}\rangle$ for which the states $\rho_A = |\psi\rangle\langle\psi|$ and $\hat{\rho}_A = |\hat{\psi}\rangle\langle\hat{\psi}|$ are ε-close:*

$$T(\rho_A, \hat{\rho}_A) \leq \varepsilon.$$

Proof. We choose the measurement as in Lemma 11.6 and choose the estimate as $|\hat{\psi}\rangle = |\phi\rangle$, the outcome of the measurement. By Theorem 11.7 we have

$$\mathbb{E}\left(1 - |\langle\hat{\psi}|\psi\rangle|^2\right) \leq \frac{|A|}{n}.$$

Fix $\varepsilon, \delta > 0$ and let $n \geq \frac{|A|}{\delta\varepsilon^2}$. By Markov's inequality Lemma B.3 we have

$$\Pr\left(1 - |\langle\hat{\psi}|\psi\rangle|^2 \geq \varepsilon^2\right) \leq \frac{|A|}{n\varepsilon^2} \leq \delta.$$

Since $T(\rho_A, \hat{\rho}_A) = \sqrt{1 - |\langle\hat{\psi}|\psi\rangle|^2}$ this implies that with probability at least $1-\delta$ the estimate is close in trace distance: $T(\rho_A, \hat{\rho}_A)$. □

By Theorem 10.15 the number of copies required scales in the optimal way with $|A|$ and ε (ignoring log-factors in the sample complexity).

11.4 Random unitaries and decoupling

The most basic information about a random variable $\mathbf{X}$ is its expectation value (or mean) $\mathbb{E}\mathbf{X}$. To get an idea of how 'spread out' the values of $\mathbf{X}$ around the mean are one may compute the variance

$$\operatorname{Var}(\mathbf{X}) = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})^2 = \mathbb{E}\mathbf{X}^2 - (\mathbb{E}\mathbf{X})^2.$$

More generally, we can extract *all* information[1] about the random variable from the *moments* $\mathbb{E}\mathbf{X}^n$ for $n \in \mathbb{N}$. A quantum analog is the following: given $M_A \in \mathrm{S}(A)$ we can apply a random unitary to obtain

$$M_A(U) = U M_A U^\dagger$$

and the analog of its moments are the operators

$$\mathbb{E}_U M_A(U)^{\otimes n} = \int_{\mathrm{U}(A)} (U M_A U^\dagger)^{\otimes n} \mathrm{d}U \in \operatorname{Lin}(A^n).$$

We may easily check that these operators commute with the action of $\mathrm{U}(A)$: for any $V \in \mathrm{U}(A)$

$$\begin{aligned} V^{\otimes n} \mathbb{E}_U M_A(U)^{\otimes n} &= \int_{\mathrm{U}(A)} (V U M_A U^\dagger)^{\otimes n} \mathrm{d}U \\ &= \int_{\mathrm{U}(A)} (V U M_A (VU)^\dagger)^{\otimes n} \mathrm{d}U V^{\otimes n} = \mathbb{E}_U M_A(U)^{\otimes n} V^{\otimes n} \end{aligned}$$

by the invariance property of the Haar measure. By Theorem 11.3 this implies that

$$\mathbb{E}_U M_A(U)^{\otimes n} = \sum_{\pi \in S_n} \alpha_\pi R_\pi$$

for some $\alpha_\pi \in \mathbb{C}$. Since $U^{\otimes n}$ commutes with R_π and by cyclicity of the trace

$$\operatorname{tr}[R_\pi^\dagger \mathbb{E}_U M_A(U)^{\otimes n}] = \operatorname{tr}[R_\pi^\dagger M_A^{\otimes n}].$$

This means that from Eq. (11.1) we get

$$\mathbb{E}_U M_A(U) = \frac{\operatorname{tr}[M_A]}{|A|} \mathbb{1}_A. \tag{11.7}$$

We can interpret Eq. (11.7) as saying that the quantum channel in which one applies a uniformly random unitary U to a quantum system is the completely depolarizing channel.

Eq. (11.2) can be used in the same way to compute the second moment. Here we may use that $\operatorname{tr}[M_A^{\otimes 2}] = \operatorname{tr}[M_A]^2$ and that for the swap operator F we have (Exercise 11.1)

$$\operatorname{tr}\left[F M_A^{\otimes 2}\right] = \operatorname{tr}\left[M_A^2\right]. \tag{11.8}$$

This leads to

$$\mathbb{E}_U M_A(U)^{\otimes 2} = \alpha \mathbb{1} + \beta F \tag{11.9}$$

where

$$\alpha = \frac{1}{|A|^3 - |A|}(|A| \operatorname{tr}[M_A]^2 - \operatorname{tr}\left[M_A^2\right]) \qquad \beta = \frac{1}{|A|^3 - |A|}(|A| \operatorname{tr}\left[M_A^2\right] - \operatorname{tr}[M_A]^2). \tag{11.10}$$

We won't need any higher moments, but the same logic applies.

[1] This is true as long as $\mathbf{X}$ satisfies mild conditions, which are always satisfied if $\mathbf{X}$ takes bounded values.

11.4.1 The decoupling inequality

We will use the above computations to show that random unitaries have a 'decoupling' property. Consider quantum systems $A = A_1 A_2$ and R. What we will show is that if A_2 is sufficiently large, then if we start with a state ρ_{AR} and apply a random unitary on A, the resulting state $\sigma_{AR} = (U_A \otimes \mathbb{1}_R)\rho_{AR}(\mathrm{U}_A^\dagger \otimes \mathbb{1}_R)$ will be such that upon discarding the subsystem A_2, the state $\sigma_{A_1 R}$ is approximately a product state between A_1 and R. Moreover, the reduced state σ_{A_1} is close to maximally mixed. The proof of Theorem 11.10 is a bit longer than most other results in these notes (although all the steps are essentially straightforward). We will be rewarded for this hard work next lecture, where we will see that it can be used to perform many quantum information protocols.

We need the following estimate for the trace norm, the proof of which will be Exercise 11.6.

Lemma 11.9. *Let $M \in \mathrm{Lin}(\mathcal{H})$, then*

$$\|M\|_1 \leq \sqrt{\mathrm{rank}(M)}\|M\|_2.$$

Theorem 11.10. *Let $A = A_1 A_2$ and let $\rho_{AR} \in \mathrm{PSD}(AR)$. Then*

$$\mathbb{E}_{U_A}\left\|\mathrm{tr}_{A_2}\left[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right] - \tau_{A_1} \otimes \rho_R\right\|_1^2 \leq \frac{|A_1||R|}{|A_2|}\,\mathrm{tr}[\rho_{AR}^2]$$

where $\tau_{A_1} = \frac{\mathbb{1}_{A_1}}{|A_1|}$ is the maximally mixed state.

Proof. By Lemma 11.9

$$\begin{aligned}&\left\|\mathrm{tr}_{A_2}\left[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right] - \tau_{A_1} \otimes \rho_R\right\|_1^2 \\ &\quad\leq |A_1||R|\left\|\mathrm{tr}_{A_2}\left[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right] - \tau_{A_1} \otimes \rho_R\right\|_2^2.\end{aligned} \tag{11.11}$$

If we let

$$X = \mathrm{tr}_{A_2}\left[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right]$$

then

$$\begin{aligned}\mathbb{E}_U X &= \mathrm{tr}_{A_2}\left[\mathbb{E}_U (U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right] \\ &= \mathrm{tr}_{A_2}\left[\frac{1}{|A|}\mathbb{1}_A \otimes \rho_R\right] = \tau_{A_1} \otimes \rho_R\end{aligned}$$

using that by Eq. (11.7) applying a random unitary corresponds to a completely depolarizing channel. That means that we may estimate from Eq. (11.11) as

$$\mathbb{E}_U\left\|\mathrm{tr}_{A_2}\left[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right] - \tau_{A_1} \otimes \rho_R\right\|_1^2 \leq |A_1||R|\mathbb{E}_U\,\mathrm{tr}\big[(X - \mathbb{E}_U X)^2\big]$$

We expand the variance as

$$\mathbb{E}_U\,\mathrm{tr}\big[(X - \mathbb{E}_U X)^2\big] = \mathrm{tr}\big[\mathbb{E}_U X^2\big] - \mathrm{tr}\big[(\mathbb{E}_U X)^2\big]$$

The second term is given by

$$\operatorname{tr}\big[(\mathbb{E}_U X)^2\big] = \operatorname{tr}\big[\tau_{A_1}^2\big] \operatorname{tr}\big[\rho_R^2\big] = \frac{1}{|A_1|} \operatorname{tr}\big[\rho_R^2\big] \tag{11.12}$$

Let us now focus on the first term $\mathbb{E}_U \operatorname{tr}\big[X^2\big]$. We rewrite it using Eq. (11.8) and cyclicity of the trace as

$$\begin{aligned}\mathbb{E}_U \operatorname{tr}\big[X^{\otimes 2} F_{A_1A_1RR}\big] &= \mathbb{E}_U \operatorname{tr}\Big[(U_A \otimes \mathbb{1}_R)^{\otimes 2} \rho_{AR}^{\otimes 2} (U_A^\dagger \otimes \mathbb{1}_R)^{\otimes 2} (F_{A_1A_1} \otimes \mathbb{1}_{A_2A_2} \otimes F_{RR})\Big] \\ &= \mathbb{E}_U \operatorname{tr}\Big[\rho_{AR}^{\otimes 2}\big((U_A^\dagger)^{\otimes 2}(F_{A_1A_1} \otimes \mathbb{1}_{A_2A_2}) U_A^{\otimes 2}\big) \otimes F_{RR}\Big]\end{aligned}$$

Now we may use Eq. (11.9) to compute

$$\mathbb{E}_U \, (U_A^\dagger)^{\otimes 2}(F_{A_1A_1} \otimes \mathbb{1}_{A_2A_2}) U_A^{\otimes 2} = \alpha \mathbb{1}_{AA} + \beta F_{AA}$$

where

$$\begin{aligned}\alpha &= \frac{1}{|A|^3 - |A|}(|A| \operatorname{tr}[F_{A_1A_1} \otimes \mathbb{1}_{A_2A_2}] - \operatorname{tr}[(F_{A_1A_1} \otimes \mathbb{1}_{A_2A_2})F_{AA}]) \\ &= \frac{1}{|A|^3 - |A|}(|A||A_1||A_2|^2 - |A_1|^2|A_2|)\end{aligned}$$

using that $(F_{A_1A_1} \otimes \mathbb{1}_{A_2A_2})F_{AA} = \mathbb{1}_{A_1A_1} \otimes F_{A_2A_2}$. This simplifies, using $|A| = |A_1||A_2|$, to

$$\alpha = \frac{|A||A_2| - |A_1|}{|A|^2 - 1} \leq \frac{1}{|A_1|}$$

By a similar computation,

$$\beta = \frac{|A||A_1| - |A_2|}{|A|^2 - 1} \leq \frac{1}{|A_2|}.$$

All in all this gives

$$\begin{aligned}\alpha \operatorname{tr}\big[\rho_{AR}^{\otimes 2}(\mathbb{1}_{AA} \otimes F_{RR})\big] + \beta \operatorname{tr}\big[\rho_{AR}^{\otimes 2}(F_{AA} \otimes F_{RR})\big] &= \alpha \operatorname{tr}\big[\rho_R^2\big] + \beta \operatorname{tr}\big[\rho_{AR}^2\big] \\ &\leq \frac{1}{|A_1|} \operatorname{tr}\big[\rho_R^2\big] + \frac{1}{|A_2|} \operatorname{tr}\big[\rho_{AR}^2\big]\end{aligned}$$

Note that the first term matches Eq. (11.12). When the dust settles, we conclude that

$$\begin{aligned}&\mathbb{E}_U \Big\| \operatorname{tr}_{A_2}\Big[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\Big] - \tau_{A_1} \otimes \rho_R \Big\|_1^2 \\ &\quad \leq |A_1||R| \left(\frac{1}{|A_1|} \operatorname{tr}\big[\rho_R^2\big] + \frac{1}{|A_2|} \operatorname{tr}\big[\rho_{AR}^2\big] - \frac{1}{|A_1|} \operatorname{tr}\big[\rho_R^2\big]\right) = \frac{|A_1||R|}{|A_2|} \operatorname{tr}\big[\rho_{AR}^2\big].\end{aligned}$$

□

Outlook

The proof of Theorem 11.3 is based on *representation theory*, which is the mathematical study of symmetries. The representation theory of the unitary group and the symmetric group S_n is a very useful tool in quantum information theory. Theorem 11.3 is the basis for *Schur-Weyl duality* which relates the representation theory of the symmetric group S_n and the unitary group $\mathrm{U}(d) := \mathrm{U}(\mathbb{C}^d)$.

11.5 Exercises

11.1 **The swap operator:** If F is the swap operator on $\mathcal{H}$, show that

$$\operatorname{tr}[M^{\otimes 2}F] = \operatorname{tr}[M^2].$$

11.2 **Unitarily invariant operators:** Let $F \in \operatorname{Lin}(A^2)$ be the swap operator.

(a) Show that $\operatorname{tr}[F] = |A|$.

(b) Verify Eq. (11.1) and Eq. (11.2) using Exercise 11.1. Derive the values of α and β in Eq. (11.10).

11.3 **Dimension symmetric subspace**. Verify the claim in Lemma 11.4 about the dimension of the symmetric subspace.

11.4 **Inner products between random vectors:** Let $|\psi\rangle$ and $|\phi\rangle$ be randomly chosen vectors $\mathcal{H} = \mathbb{C}^d$. More precisely, we mean that

$$|\psi\rangle = U_1|0\rangle \ , \quad |\phi\rangle = U_2|0\rangle \ ,$$

where $|0\rangle \in \mathcal{H}$ is fixed, and U_1, U_2 are both uniformly distributed according to the Haar measure. Their inner product defines a random variable

$$\mathbf{X} = |\langle\psi|\phi\rangle|^2 \ .$$

(a) Show that the expected value of $\mathbf{X}$ is given by

$$\mathbb{E}\mathbf{X} = \frac{1}{d} \ .$$

Hint: You should start by using the invariance of the Haar measure to argue that, without loss of generality, you can take $U_2 = \mathbb{1}$.

(b) Show that the variance of $\mathbf{X}$ is given by

$$\operatorname{Var}\mathbf{X} = \frac{1}{d^2} \ .$$

11.5 **Typical states are highly entangled:** In this exercise you will show that if one has a bipartite system AB with Hilbert spaces of sufficiently large dimension, a *random* pure state will be close to maximally entangled with high probability. We can pick a random pure state by choosing $|\psi\rangle$ for a uniformly random ψ on the unit sphere of $\mathcal{H}_A \otimes \mathcal{H}_B$.

(a) Argue that this is equivalent to picking some initial fixed state $|0_{AB}\rangle$ and applying a random unitary $U_{AB} \in \mathrm{U}(AB)$, so $|\psi_{AB}\rangle = U_{AB}|0_{AB}\rangle$.

(b) Show that if we let $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$, we have

$$\operatorname{tr}[\rho_A^2] = \operatorname{tr}\left[\rho_{AB}^{\otimes 2} F_{AA} \otimes \mathbb{1}_{BB}\right]$$

where F_{AA} is the swap operator on two copies of A.

(c) Show that

$$\mathbb{E}_U \rho_{AB}^{\otimes 2} = \frac{1}{|AB|(|AB|+1)} \left(\mathbb{1}_{AABB} + F_{AABB}\right).$$

and hence

$$\mathbb{E}_U \operatorname{tr}[\rho_A^2] \leq \frac{1}{|A|} + \frac{1}{|B|}.$$

(d) Use this to show that

$$\mathbb{E}_U H_2(\rho_A) \geq -\log(\frac{1}{|A|} + \frac{1}{|B|})$$

Hint: use Jensen's inequality.

(e) Suppose that $|A| \leq |B|$. Show that

$$\log(|A|) - \log\left(1 + \frac{|A|}{|B|}\right) \leq \mathbb{E}_U H(\rho_A) \leq \log(|A|)$$

Hint: you may use the results of Exercise 10.12.

(f) Conclude that for *large* $|A|$ and $|B|$ we have

$$\mathbb{E}_U H(\rho_A) \approx \min(\log(|A|), \log(|B|)).$$

Remark: By Markov's inequality to the positive quantity $\min(\log(|A|), \log(|B|)) - H(\rho_A)$ this actually means that we have $\mathbb{E}_U H(\rho_A) \approx \min(\log(|A|), \log(|B|))$ with high probability. In other words, a random state is close to maximally entangled with high probability!

11.6 **Trace norm estimate:** The goal of this exercise will be to prove a (sharper version of) Lemma 11.9. So, as in Lemma 11.9 we let $M \in \mathrm{Lin}(\mathcal{H})$. We let $\omega \in \mathrm{PD}(\mathcal{H})$. *Hint: it is probably helpful to recall some properties of the trace norm in Lecture 6. You will need your favorite inequality: the Cauchy-Schwartz inequality!*

(a) Show that

$$\|M\|_1 = \max_{U \in \mathrm{U}(\mathcal{H})} \left| \mathrm{tr}\left[\left(\omega^{\frac{1}{4}} U \omega^{\frac{1}{4}} \right) \left(\omega^{-\frac{1}{4}} M \omega^{-\frac{1}{4}} \right) \right] \right|.$$

(b) Next, show that

$$\|M\|_1 \leq \sqrt{\max_{U \in \mathrm{U}(\mathcal{H})} \left| \mathrm{tr}\left[\omega^{\frac{1}{2}} U \omega^{\frac{1}{2}} U^\dagger \right] \right| \mathrm{tr}\left[\omega^{-\frac{1}{4}} M^\dagger \omega^{-\frac{1}{2}} M \omega^{-\frac{1}{4}} \right]}.$$

(c) Argue that

$$\max_{U \in \mathrm{U}(\mathcal{H})} \left| \mathrm{tr}\left[\omega^{\frac{1}{2}} U \omega^{\frac{1}{2}} U^\dagger \right] \right| = \mathrm{tr}[\sigma].$$

and conclude that

$$\|M\|_1 \leq \sqrt{\mathrm{tr}[\omega]} \left\| \omega^{-\frac{1}{4}} M \omega^{-\frac{1}{4}} \right\|_2$$

(d) Prove Lemma 11.9 by taking ω to be the projection onto the image of M.

11.7 **Measuring purity with randomness:** This exercise looks at some different ways to measure the purity of a quantum state $\rho_A \in \mathrm{S}(A)$, defined by $P(\rho_A) = \mathrm{tr}[\rho_A^2]$.

(a) Use Exercise 11.1 to deduce that $P(\rho_A)$ can be estimated, given two simultaneous copies of ρ_A, i.e. by measuring the state $\rho_A^{\otimes 2}$.

(b) In fact we can estimate $P(\rho)$ using only one copy of ρ_A at a time. Let $|\psi_A\rangle \in \mathcal{H}_A$ be a normalised state, and $U \in \mathrm{U}(A)$ be a unitary operator. Define $m_U(\psi_A) = \mathrm{tr}[\rho_A U |\psi_A\rangle\langle\psi_A| U^\dagger]$, the probability that we return "1" if we measure ρ_A along the axis of $U|\psi_A\rangle$. Show that

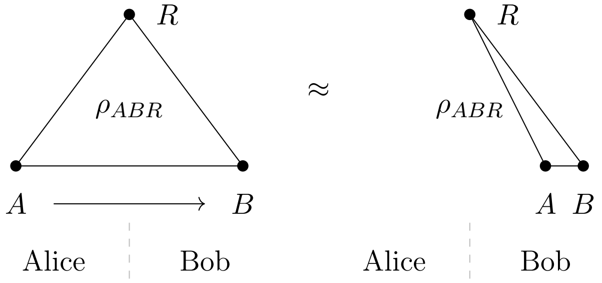
$$P(\rho_A) = |A|(|A| + 1)\mathbb{E}_U[m_U(\psi_A)^2] - 1 \,,$$

and hence describe how $P(\rho_A)$ can be estimated.

Lecture 12

Quantum state merging

Last lecture we saw that under appropriate conditions we can use a (random) unitary to achieve decoupling. In this lecture we will see an application of this tool: state merging. State merging is the following problem: Alice and Bob share a quantum state ρ_{AB} and they want to transfer Alice's part of the state to Bob. Moreover, they want to do so in a way preserving external correlations. That is, we should consider a purification ρ_{ABR} where a third party Robin holds the system R. The goal of the protocol is that Bob ends up with the AB systems by interacting with Alice, and the joint state with Robin is (approximately) ρ_{ABR}.



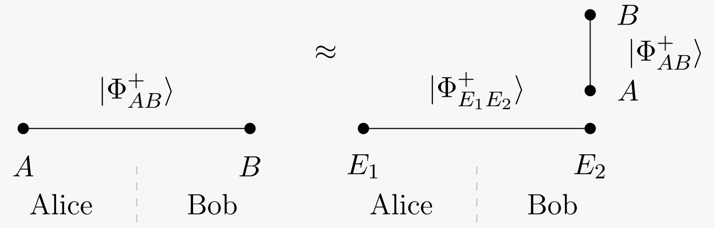
The key question is what the required resources are for this task! There are two possible questions:

- Alice can send qubits to Bob. How many qubits does she need to send?
- Alice can send classical bits to Bob, and they may consume some amount of pre-shared maximally entangled states. How much classical communication do they need, and how much entanglement is required?

As for compression, we can define this task in the situation where there is a single copy of the state, and we want to perform (approximate) state merging. We will define and investigate the asymptotic version of the problem, where we have many copies $\rho_{ABR}^{\otimes n}$, and we want to transfer A^n to Bob, and we would like to know what the *rates* of the required resources are.

To give a rough upper bound on the required resources, note that state merging can be achieved if Alice just sends the full system over to Bob, i.e. she sends over $\log(|A|)$ qubits. In fact, based on our knowledge of compression, we see that it in fact suffices that she compresses her part of the state and sends over $H_0^\varepsilon(A)_\rho$ qubits in the one-shot case, or sends over qubits at a rate of $H(A)_\rho$ in the asymptotic scenario. Also, by teleportation, if Alice can merge by sending over r qubits, then she can also use teleportation to send over these qubits, using $2r$ classical bits and r maximally entangled pairs. However, we will see that we can do better than this in general! Not only can the required number of qubits be lower, but it is also possible that after the protocol we have actually *generated* additional entanglement between Alice and Bob.

Example 12.1. To give a (silly but instructive) example: suppose that Alice and Bob share a maximally entangled state (and therefore are uncorrelated with Robin). Then Bob can just locally prepare a maximally entangled state, and he has achieved state merging. At the same time, Alice and Bob still share the maximally entangled state. Therefore, state merging has achieved at *zero* cost, and we are in fact left with a maximally entangled state between Alice and Bob!



In this figure, the original systems AB are simply relabeled to E_1E_2, representing the 'leftover' entanglement, and Bob locally prepared a maximally entangled state.

12.1 The decoupling principle

The key technical ingredient for constructing a state merging protocol will be *decoupling.* If we have a quantum system to which we apply a channel, we would like to know whether we can recover the information in the initial quantum system. If we understand how this works, we can use this to protect quantum information against errors. Let $\Phi_{A\to B} \in \mathrm{C}(A, B)$ be a quantum channel. We can *recover* the state from this channel if there exists a recovery channel $\mathcal{R}_{B\to A}$ which is such that for a purification ρ_{AR} of ρ_A

$$((\mathcal{R}_{B\to A} \circ \Phi_{A\to B}) \otimes \mathcal{I}_R)(\rho_{AR}) = \rho_{AR}.$$

An example of this is given by compression, where $\Phi_{A\to B}$ is the encoding channel and $\mathcal{R}_{B\to A}$ is the decoding channel. More generally, we would like to know when we can recover, given some channel $\Phi_{A\to B}$ and a state ρ_A. If we consider a Stinespring extension $V_{BE} \in \mathrm{Isom}(A, BE)$ of the channel $\Phi_{A\to B}$, then the idea is that 'all information being preserved in B' is equivalent to 'no information gets transferred to E'. That is, if we let σ_{BRE} be the state obtained after applying the Stinespring isometry V then we can recover if and only if there are no correlations between R and E, so

$$\sigma_{RE} = \sigma_R \otimes \sigma_E.$$

Why can we recover if σ decouples between E and R? We can see this from the uniqueness of purifications. On the one hand, σ_{BRE} is a purification of σ_{RE}. On the other hand we can pick and arbitrary purification τ_{EF} of σ_E, and let $\rho_{AR} \otimes \tau_{EF}$ be a purification of σ_{RE}. Then there must be an isometry $W \in \mathrm{Isom}(B, AF)$ such that

$$\rho_{AR} \otimes \tau_{EF} = (W \otimes \mathbb{1}_{RE})\sigma_{BRE}(W^\dagger \otimes \mathbb{1}_{RE}).$$

The isometry W only acts on the B system, and we can take it to be the Stinespring extension of a recovery channel

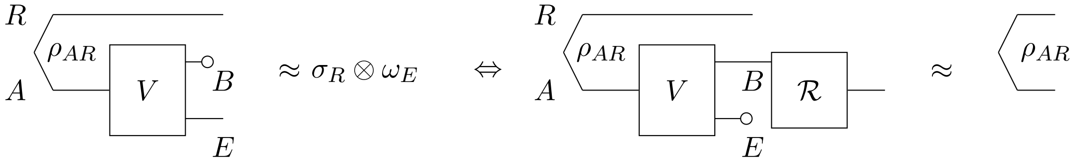
$$\mathcal{R}_{B\to A}[M_B] = \mathrm{tr}_F[W M_B W^\dagger]. \tag{12.1}$$

This is such that it maps σ_{BR} to $\mathrm{tr}_{EF}[\rho_{AR} \otimes \tau_{EF}] = \rho_{AR}$. Finally, if Alice does *not* trace out the system F, she shares the pure state τ_{EF} with the environment.

There is also an *approximate* version: we say we can (approximately) *recover* ρ_A with error $\varepsilon > 0$ if there exists some recovery channel $\mathcal{R}_{B\to A}$ such that

$$P_E(\mathcal{R}_{B\to A} \circ \Phi_{A\to B}, \rho_A) \leq \varepsilon.$$

Recall that we defined the entanglement purified distance by choosing a purification $\rho_{AR} = |\phi_{AR}\rangle\langle\phi_{AR}|$ for ρ_A. The approximate recovery property is equivalent to σ_{RE} being close to a product state $\sigma_R \otimes \omega_E$:



This is made precise in the following result, which is based on Uhlmann's theorem.

Lemma 12.2. *Let $\rho_{AR} \in \mathrm{S}(AR)$ be pure. Let $V \in \mathrm{Isom}(A, BE)$ be a Stinespring extension of a channel $\Phi_{A\to B}$. Denote by $\sigma_{BRE} = |\psi_{BRE}\rangle\langle\psi_{BRE}|$ the state obtained from applying V to ρ_{AR}. Then the following are equivalent:*

(a) The state ρ_A can be recovered with error ε from $\Phi_{A\to B}(\rho_A)$, i.e. there exists a channel $\mathcal{R}_{B\to A}$ such that

$$P_E(\mathcal{R}_{B\to A} \circ \Phi_{A\to B}, \rho_A) \leq \varepsilon.$$

(b) The reduced state σ_{RE} is close to a product state: there exists $\omega_E \in \mathrm{S}(E)$ such that

$$P(\sigma_{RE}, \sigma_R \otimes \omega_E) \leq \varepsilon.$$

Proof. Suppose that $P(\sigma_{RE}, \sigma_R \otimes \omega_E) \leq \varepsilon$. Note that by construction, $|\psi_{BER}\rangle$ is a purification of σ_{ER}. On the other hand, we may pick an arbitrary purification ω_{EF} of ω_E, and then $\rho_{AR} \otimes \omega_{EF}$ is a purification of $\sigma_R \otimes \omega_E$. By Uhlmann's theorem there must be an isometry $W \in \mathrm{Isom}(B, AF)$ such that

$$\varepsilon \geq P(\sigma_{RE}, \sigma_R \otimes \omega_E) = P((W \otimes \mathbb{1}_{RE})\sigma_{BRE}(W^\dagger \otimes \mathbb{1}_{RE}), \rho_{AR} \otimes \omega_{EF}).$$

But now we may simply take W to be the Stinespring isometry of our recovery channel and define $\mathcal{R}_{B\to A}$ as in Eq. (12.1). Using monotonicity of the purified distance we see that

$$\begin{aligned} P_E(\mathcal{R}_{B\to A} \circ \Phi_{A\to B}, \rho_{AR}) &= P((\mathcal{R}_{B\to A} \otimes \mathcal{I}_R)(\sigma_{BR}), \rho_{AR}) \\ &\leq P((W \otimes \mathbb{1}_{RE})\sigma_{BRE}(W^\dagger \otimes \mathbb{1}_{RE}), \rho_{AR} \otimes \omega_{EF}) \\ &\leq \varepsilon. \end{aligned}$$

For the converse, suppose that there exists a recovery channel $\mathcal{R}_{B\to A}$ such that we can recover with error ε, so $P_E(\mathcal{R}_{B\to A} \circ \Phi_{A\to B}, \rho_A) \leq \varepsilon$. Let

$$\tilde{\rho}_{ARE} = (\mathcal{R}_{B\to A} \otimes \mathcal{I}_{RE})(\sigma_{BRE})$$

then this is such that $\mathrm{tr}_E[\tilde{\rho}_{ARE}] = (\mathcal{R}_{B\to A} \circ \Phi_{A\to B} \otimes \mathcal{I}_R)(\rho_{AR})$. On the other hand, since ρ_{AR} is pure, any extension ρ_{ARE} must be of the form $\rho_{AR} \otimes \omega_E$. Therefore, by Uhlmann's theorem there exists ω_E such that

$$\varepsilon \geq P((\mathcal{R}_{B\to A} \circ \Phi_{A\to B} \otimes \mathcal{I}_R)(\rho_{AR}), \rho_{AR}) = P(\tilde{\rho}_{ARE}, \rho_{AR} \otimes \omega_E).$$

Now, note that $\mathrm{tr}_A[\tilde{\rho}_{ARE}] = \sigma_{RE}$ since $\tilde{\rho}_{ARE} = (\mathcal{R}_{B\to A} \otimes \mathcal{I}_{RE})(\sigma_{BRE})$, and $\mathrm{tr}_A[\rho_{AR} \otimes \omega_E] = \rho_R \otimes \omega_E = \sigma_R \otimes \omega_E$. By monotonicity of the purified distance

$$P(\sigma_{RE}, \sigma_R \otimes \omega_E) = P(\tilde{\rho}_R \otimes \omega_E, \rho_{RE}) \leq P(\tilde{\rho}_{ARE}, \rho_{AR} \otimes \omega_E) \leq \varepsilon.$$

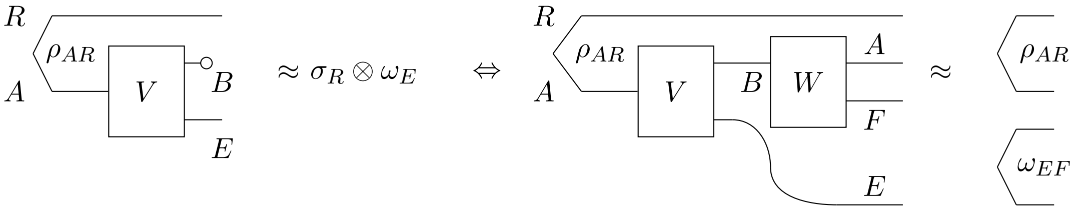
□

Again, if one does not trace out the F system when applying W, one is left with the state ω_{EF}:

Corollary 12.3. *Let $\rho_{AR} \in \mathrm{S}(AR)$ be pure, let $V \in \mathrm{Isom}(A, BE)$ and denote by σ_{BRE} the state obtained from applying V to ρ_{AR}. Then if σ_{RE} is close to a product state, so $P(\sigma_{RE}, \sigma_R \otimes \omega_E) \leq \varepsilon$, there exists an isometry $W \in \mathrm{Isom}(B, AF)$ such that when applying W to σ_{BRE}, we obtain a state τ_{AREF} which is such that*

$$P(\tau_{AREF}, \rho_{AR} \otimes \omega_{EF}) \leq \varepsilon.$$

As a diagram:



The decoupling inequality from Theorem 11.10 now tells us the following: given $A = A_1 A_2$ and a (pure) state ρ_{AR} with sufficiently large $|A_2|$ (such that $|A_2| \gg |A_1||R|$) the channel

$$M_A \mapsto \mathrm{tr}_{A_2}[U_A M_A U_A^\dagger]$$

where we apply a random unitary is such that with high probability (with respect to the choice of U_A) we can (approximately) recover ρ_A from A_2. Moreover, since the reduced state on A_1 is maximally mixed, the recovery channel may additionally produce a maximally entangled state purifying τ_{A_1}.

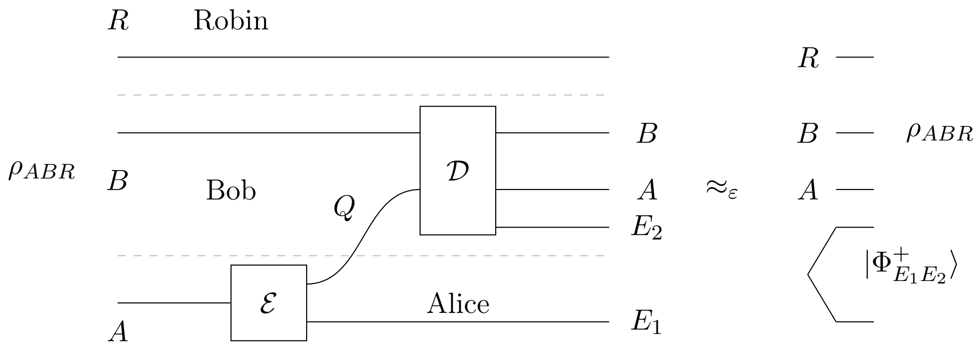
12.2 The state merging task

State merging is a task where Alice and Bob share a state ρ_{AB}, and want to get the system A to Bob. As discussed at the start of this lecture, they should do so in a way that preserves correlations with a reference system R held by Robin. Such a protocol consists of the following steps:

- Alice applies an encoding channel $\mathcal{E}$ to her system. She sends a quantum system Q to Bob and keeps a system E_1 herself.
- Bob applies a decoding channel to the system B and the system Q he received from Alice. He is left with systems AB and a system E_2

This should be such that on R and Bob's systems AB we are left with the state ρ_{ABR}, and on the systems $E_1 E_2$, shared between Alice and Bob, we have a maximally entangled state. We want this to be such that the number of qubits in Q (the qubits that are communicated between

Alice and Bob) is as small as possible, and the number of maximally entangled qubits in E_1E_2 is as large as possible. We allow a small error ε. It is probably easiest to see what this means in the following diagram:



We can also write down a formal definition:

Definition 12.4. For $\rho_{ABR} \in \mathrm{S}(ABR)$, a *state merging protocol* with error ε, quantum communication cost q and entanglement gain e consists of quantum channels $\mathcal{E} \in \mathrm{C}(A, QE_1)$ and $\mathcal{D} \in \mathrm{C}(QB, ABE_2)$ where Q, E_1 and E_2 are quantum systems with

$$\log(|Q|) \leq q \qquad \log(|E_1|) = \log(|E_2|) \geq e$$

which are such that if we let

$$\sigma_{ABRE_1E_2} = (\mathcal{D} \otimes \mathcal{I}_{E_1R})((\mathcal{E} \otimes \mathcal{I}_{BR})(\rho_{ABR}))$$

and $\omega_{E_1E_2} = |\Phi^+_{E_1E_2}\rangle\langle\Phi^+_{E_1E_2}|$ a maximally entangled state, then

$$P(\sigma_{ABRE_1E_2}, \rho_{ABR} \otimes \omega_{E_1E_2}) \leq \varepsilon.$$

Then, we will investigate the *rate* at which state transfer is possible, meaning that Alice and Bob share many copies of ρ_{AB}. They want to perform state merging for $\rho_{ABR}^{\otimes n}$ for large n using as few as possible qubits of communication per copy.

We let $Q^\varepsilon(A : B : R)_\rho$ denote the smallest q such that there exists a state merging protocol with error at most ε and quantum communication q for ρ_{ABR}.

For the asymptotic question we consider state transfer for $\rho_{ABR}^{\otimes n}$, and let the error vanish as n goes to infinity. We are then interested in how many qubits of communication we require per copy of ρ_{ABR}:

$$q(A : B : R)_\rho = \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} Q^\varepsilon(A^n : B^n : R^n)_{\rho^{\otimes n}}$$

We will show below that state merging can be achieved at a certain rate. Let us first sketch the idea of the argument.

i) Alice applies a random unitary to her system. Denote the resulting state by σ.

ii) She splits up her system into Q and E_1 in such a way that E_1 is decoupled from R.

iii) Now we have a state which is decoupled, so $\sigma_{E_1R} \approx \tau_{E_1} \otimes \sigma_R$. On the one hand, this has σ_{QE_1BR} as a purification. On the other hand, $|\Phi^+_{E_1E_2}\rangle\langle\Phi^+_{E_1E_2}| \otimes \rho_{ABR}$ is a purification of $\tau_{E_1} \otimes \sigma_R$. This implies by an application of Uhlmann's theorem that Bob can recover ρ_{ABR} from the B and Q systems by applying an isometry, moreover establishing a maximally entangled state between E_1 and E_2.

iv) Finally, we note that in the decoupling theorem (since $|A| = |Q||E_1|$), the appropriate condition will be

$$|Q|^2 \gg |A||R| \operatorname{tr}[\rho_{AR}^2]$$

This also creates a state which approximately maximally entangled on a system of size

$$|E| = \frac{|A|}{|Q|}$$

v) We will actually apply this procedure to n copies of ρ_{ABR}. By a typical subspace argument we will see that we can approximate the decoupling condition in the asymptotic limit by

$$\begin{aligned}\frac{1}{n}\log(|Q|) &\geq \frac{1}{2}\left(H(A) + H(R) - H(AR)\right) + \dots \\ &= \frac{1}{2}I(A:R) + \dots\end{aligned}$$

where there is an error term which goes to zero as n goes to infinity. The amount of entanglement created becomes

$$\begin{aligned}\frac{1}{n}\log(|E|) &= H(A) - \frac{1}{2}I(A:R) + \dots \\ &= \frac{1}{2}(H(A) + H(AR) - H(R)) + \dots \\ &= \frac{1}{2}I(A:B) + \dots\end{aligned}$$

using that ρ_{ABR} is pure.

We summarize our conclusions in a theorem, and we will make the argument precise.

Theorem 12.5. *State merging can be achieved by sending over qubits at rate $q(A:B:R)_\rho \leq \frac{1}{2}I(A:R)$, generating maximally entangled qubit states at a rate of $e(A:B:R)_\rho \geq \frac{1}{2}I(A:B)$.*

To prove this result, we first prove a one-shot version:

Lemma 12.6. *For $\rho_{ABR} \in \mathrm{S}(ABR)$ there exists a state merging protocol with error ε, quantum communication cost and entanglement gain e for any q such that*

$$q \geq \frac{1}{2}\left(\log(|A| + \log(|R|)) + \log(\operatorname{tr}[\rho_{AR}^2])\right) + 2\log\left(\frac{1}{\varepsilon}\right)$$

with entanglement gain

$$e = \log(|A|) - q.$$

Proof. First, note that if the bound in the statement of the lemma gives $q \geq \log(|A|)$ we can simply take $Q = A$ and send over the full system. Otherwise, we divide the system A into systems Q and E_1 such that

$$\frac{|E_1||R|}{|Q|}\operatorname{tr}[\rho_{AR}^2] \leq \varepsilon^4.$$

Since $|A| = |Q||E_1|$ this is equivalent to

$$|Q|^2 \geq \frac{|A||R|}{\varepsilon^4} \operatorname{tr}[\rho_{AR}^2]$$

and hence

$$q = \log(|Q|) \geq \frac{1}{2}\left(\log(|A| + \log(|R|)) + \log(\operatorname{tr}[\rho_{AR}^2])\right) + 2\log\left(\frac{1}{\varepsilon}\right).$$

By Theorem 11.10, when we apply a random unitary to A and trace out Q

$$\mathbb{E}_{U_A}\left\|\operatorname{tr}_Q\left[(U_A \otimes \mathbb{1}_R)\rho_{AR}(U_A^\dagger \otimes \mathbb{1}_R)\right] - \tau_{E_1} \otimes \rho_R\right\|_1^2 \leq \frac{|E_1||R|}{|Q|} \operatorname{tr}[\rho_{AR}^2] \leq \varepsilon^4.$$

Since the average error is at most ε^2, there must exist at least *some* unitary U_A such that

$$\|\rho_{E_1R} - \tau_{E_1} \otimes \rho_R\|_1 \leq \varepsilon^2$$

where ρ_{QE_1BR} is the state after applying U_A to ρ_{ABR}. This is the unitary Alice chooses as encoding channel. By Eq. (6.9)

$$P(\rho_{E_1R}, \tau_{E_1} \otimes \rho_R) \leq \sqrt{2T(\rho_{E_1R}, \tau_{E_1} \otimes \rho_R)} \leq \varepsilon.$$

This means that by Corollary 12.3 there exists an isometry $W \in \operatorname{Isom}(Q, AE_2)$ that Bob can apply such that after applying it they have a state $\sigma_{ABRE_1E_2}$ for which

$$P(\sigma_{ABRE_1E_2}, \rho_{ABR} \otimes \omega_{E_1E_2}) \leq \varepsilon.$$

Note that Bob holds the systems A, B and E_2 and Alice has kept the system E_1. The state $\omega_{E_1E_2}$ is a purification of the maximally mixed state τ_{E_1} and is therefore maximally entangled. We conclude that Alice and Bob have achieved state merging with error ε. □

Next, we apply this result to $\rho_{ABR}^{\otimes n}$. The (straightforward) idea is that we compress each of the systems to the typical subspace, in which case the dimensions of the systems of Alice, Bob and Robin are approximately given by $2^{nH(A)}$, $2^{nH(B)}$ and $2^{nH(R)}$ respectively. That means that we slightly deform to a state $\tilde{\rho}_{A'B'R'}$ to live purely on the typical subspaces A', B' and R'. We can perform a merging protocol with $\log(|A'|) \approx nH(A)$, $\log(|R'|) \approx nH(R)$, and since the eigenvalues of the reduced state on B' are approximately $2^{-nH(B)}$, $\operatorname{tr}[\rho_{\tilde{AR}}{}^2] = \operatorname{tr}[\tilde{\rho}_B^2] \approx 2^{-nH(B)}$. We conclude that by using the one-shot version in Lemma 12.6, we can merge using approximately $\frac{n}{2}(H(A) + H(R) - H(B)) = \frac{n}{2}(H(A) + H(R) - H(AR))$ qubits ($H(B) = H(AR)$ using that ρ_{ABR} is pure). In conclusion, the *rate* at which we need to send qubits is given by $\frac{1}{2}(H(A) + H(B) - H(AB)) = \frac{1}{2}I(A : R)$. To turn this in a rigorous proof we have to keep track of all the error terms. We will do so carefully below; while the details are perhaps slightly painful you should keep in mind the above simple high-level idea.

We start with a lemma based on the properties of typical subspaces.

Lemma 12.7. *Let $\rho_{ABC} \in \mathrm{S}(ABC)$ be the pure state $|\psi_{ABC}\rangle$. Choose $\varepsilon, \delta > 0$, and let $\Pi_{A,n,\delta}$, $\Pi_{B,n,\delta}$ and $\Pi_{C,n,\delta}$ be the typical subspace projectors onto the typical subspaces $S_{n,\delta}(\rho_A)$, $S_{n,\delta}(\rho_B)$ and $S_{n,\delta}(\rho_C)$. Then there exists an integer N such that for all $n \geq N$ the pure state $\tilde{\rho}_{A^nB^nC^n}$ given by*

$$|\tilde{\psi}_{A^nB^nC^n}\rangle = \frac{(\Pi_{A,n,\delta} \otimes \Pi_{B,n,\delta} \otimes \Pi_{C,n,\delta})|\psi_{ABC}\rangle^{\otimes n}}{\|(\Pi_{A,n,\delta} \otimes \Pi_{B,n,\delta} \otimes \Pi_{C,n,\delta})|\psi_{ABC}\rangle^{\otimes n}\|}$$

is such that

(a) $\tilde{\rho}_{A^nB^nC^n}$ is close to $\rho_{ABC}^{\otimes n}$: $P(\tilde{\rho}_{A^nB^nC^n}, \rho_{ABC}^{\otimes n}) \leq \varepsilon$.

(b) The rank of $\tilde{\rho}_{A^n}$ is at most $2^{n(H(A)_\rho - \delta)}$ and similar for B and C.

(c) $\mathrm{tr}[\tilde{\rho}_{A^n}^2] \leq 2^{-n(H(A)_\rho - 3\delta)+1}$ and similar for B and C.

Proof. First, note that by Lemma 8.12, when we measure whether we are in the typical subset on the A^n system (the measurement $\{\Pi_{A,n,\delta}, \mathbb{1} - \Pi_{A,n,\delta}\}$) for $\rho_A^{\otimes n}$ the probability of being in the typical subspace goes to 1 as $n \to \infty$. If we measure whether we are in the typical subsets for both A, B and C, then the probability of being in the typical subspace for all three systems goes to 1:

$$\begin{aligned} p_n &:= \|(\Pi_{A,n,\delta} \otimes \Pi_{B,n,\delta} \otimes \Pi_{C,n,\delta})|\psi_{ABC}\rangle^{\otimes n}\|^2 \\ &= \mathrm{tr}[(\Pi_{A,n,\delta} \otimes \Pi_{B,n,\delta} \otimes \Pi_{C,n,\delta})\rho_{ABC}^{\otimes n}] \underset{n\to\infty}{\to} 1. \end{aligned}$$

In this situation, the post-measurement state is $\tilde{\rho}_{A^nB^nC^n}$, so by the gentle measurement lemma (Lemma 6.20)

$$\lim_{n\to\infty} P(\tilde{\rho}_{A^nB^nC^n}, \rho_{ABC}^{\otimes n}) = 0.$$

This means that we can choose N such that for all $n \geq N$ we have $p_n \geq 1 - \varepsilon^2$ and $P(\tilde{\rho}_{A^nB^nC^n}, \rho_{ABC}^{\otimes n}) \leq \varepsilon$. Next, we prove (b) and (c), for the A system; the same argument applies with A replaced by the system B or C. The state $\tilde{\rho}_{A^nB^nC^n}$ by construction lives on the typical subspaces, so the rank of $\tilde{\rho}_{A^n}$ is at most the $\dim(S_{n,\delta}(\rho_A) \leq 2^{n(H(A)_\rho - \delta)}$ by Lemma 8.12, proving (b). Finally, we use the following facts, which you may prove in Exercise 12.1 If P, Q are positive operators, then

$$P \leq Q \Rightarrow \mathrm{tr}[P^2] \leq \mathrm{tr}[Q^2] \tag{12.2}$$

If $M_{AB} \in \mathrm{Lin}(AB)$ and $0 \leq \Pi_A \leq \mathbb{1}_A$,

$$\tilde{M}_B = \mathrm{tr}_A[(\Pi_A \otimes \mathbb{1}_B)M_{AB}(\Pi_A \otimes \mathbb{1}_B)] \leq M_B. \tag{12.3}$$

This implies that for

$$\tilde{\rho}_{A'B'R'} = \frac{1}{p_n}(\Pi_{A,n,\delta} \otimes \Pi_{B,n,\delta} \otimes \Pi_{R,n,\delta})\rho_{ABR}^{\otimes n}(\Pi_{A,n,\delta} \otimes \Pi_{B,n,\delta} \otimes \Pi_{R,n,\delta})$$

we get

$$\mathrm{tr}[\tilde{\rho}_{A^n}^2] \leq \frac{\mathrm{tr}[(\Pi_{B,n\delta}\rho_B^{\otimes n}\Pi_{B,n\delta})^2]}{p_n^2}.$$

We may assume without loss of generality that $p_n^2 \geq \frac{1}{2}$. By Lemma 8.12 the eigenvalues of $\Pi_{B,n\delta}\rho_B^{\otimes n}\Pi_{B,n\delta}$ are at most $2^{-n(H(B)_\rho-\delta)}$ and the number of nonzero eigenvalues is at most $|B'| \leq 2^{n(H(B)_\rho+\delta)}$, so we can bound

$$\operatorname{tr}[(\Pi_{B,n\delta}\rho_B^{\otimes n}\Pi_{B,n\delta})^2] \leq |B'|(2^{-n(H(B)_\rho-\delta)})^2 \leq 2^{-n(H(B)_\rho-3\delta)}$$

proving (c). □

Proof of Theorem 12.5. Let $\rho_{ABR} = |\phi_{ABR}\rangle\langle\phi_{ABR}|$. Choose $\varepsilon, \delta > 0$, let $\tilde{\rho}_{A^nB^nR^n}$ be as in Lemma 12.7 for sufficiently large n. The state $\tilde{\rho}_{A^nB^nR^n}$ is a state which lives on the typical subspaces; we denote these quantum systems by A', B' and R'. We now apply the merging protocol from Lemma 12.6, using the reduced systems A', B' and R'. This merging protocol needs to send

$$q = \log(|Q|) = \lceil\frac{1}{2}\left(\log(|A'| + \log(|R'|)) + \log(\operatorname{tr}[\tilde{\rho}_{A'R'}^2])\right) + 2\log\left(\frac{1}{\varepsilon}\right)\rceil$$

qubits to merge $\tilde{\rho}_{A'B'R'}$ with error ε. In this expression

$$\log(|A'|) \leq n(H(A)_\rho + \delta) \qquad \log(|R'|) \leq n(H(R)_\rho + \delta)$$

By Lemma 12.7

$$\operatorname{tr}[\tilde{\rho}_{A'R'}^2] = \operatorname{tr}[\tilde{\rho}_{B'}^2] \leq 2^{-n(H(B)-3\delta)+1}.$$

The first equality comes from the fact that $\tilde{\rho}_{A'B'R'}$ is pure, so $\tilde{\rho}_{A'R'}$ and $\tilde{\rho}_{B'}$ have the same nonzero spectrum. This means that we need at most

$$\frac{n}{2}(H(A) + H(R) - H(B) + 5\delta) + 2\log\left(\frac{1}{\varepsilon}\right) + 2$$

qubits. Denote by $\tilde{\sigma}_{A^nB^nR^nE_1E_2}$ the state resulting from merging $\tilde{\rho}_{A^nB^nR^n}$. When we apply the protocol to the original state $\rho_{ABR}^{\otimes n}$, we see that for the final state $\sigma_{A^nB^nR^n}$, by applying the triangle inequality,

$$\begin{aligned} P(\sigma_{A^nB^nR^nE_1E_2}, \rho_{ABR}^{\otimes n} \otimes \omega_{E_1E_2}) \leq\ & P(\sigma_{A^nB^nR^nE_1E_2}, \tilde{\sigma}_{A^nB^nR^nE_1E_2}) \\ & + P(\tilde{\sigma}_{A^nB^nR^nE_1E_2}, \tilde{\rho}_{A^nB^nR^n} \otimes \omega_{E_1E_2}) \\ & + P(\tilde{\rho}_{A^nB^nR^n} \otimes \omega_{E_1E_2}, \rho_{ABR}^{\otimes n} \otimes \omega_{E_1E_2}). \end{aligned}$$

The first term is at most $P(\rho_{ABR}^{\otimes n}, \tilde{\rho}_{A^nB^nR^n}) \leq \varepsilon$ (by monotonicity). The second term is at most ε, since the protocol is a state merging protocol with error ε for $\tilde{\rho}_{A^nB^nR^n}$. The final term equals $P(\rho_{ABR}^{\otimes n}, \tilde{\rho}_{A^nB^nR^n}) \leq \varepsilon$. In conclusion, the error in the state merging protocol is at most 3ε. This shows that

$$Q^{3\varepsilon}(A^n : B^n : R^n)_\rho \leq \frac{n}{2}(H(A) + H(R) - H(B) + 5\delta) + 2\log(\varepsilon) + 2$$

Since ε and δ were arbitrary

$$\begin{aligned} q(A : B : R)_\rho &= \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}Q^\varepsilon(A^n : B^n : R^n)_\rho \leq \frac{1}{2}(H(A) + H(R) - H(B)) \\ &= \frac{1}{2}I(A : R)_\rho \end{aligned}$$

using that $H(B)_\rho = H(AR)_\rho$ since ρ_{ABR} is pure. The entanglement gain (as per Lemma 12.6) is at least

$$e = \log(|A'|) - q \geq n(H(A)_\rho - \delta) - \frac{n}{2}(H(A) + H(R) - H(B) - 5\delta) - 2\log\left(\frac{1}{\varepsilon}\right) - 2$$

using that $|A'| = \dim(S_{A,n,\delta}) \geq 2^{n(H(A)_\rho - \delta)}$. Taking the asymptotic limit and letting δ go to zero, we gain entangled qubits at rate

$$e(A:B:R)_\rho \geq \frac{1}{2}(H(A)_\rho + H(B)_\rho - H(R)_\rho) = I(A:R)_\rho$$

again using that ρ_{ABR} is pure. □

Corollary 12.8. *State merging can be achieved by sending over classical bits at rate*

$$I(A:R).$$

and at entanglement cost of rate

$$H(A|B).$$

In particular, if $H(A|B)$ is negative, maximally entangled qubits are created at rate $|H(A|B)|$.

Proof. This is a direct consequence of using the protocol of Theorem 12.5, but sending over the qubits using teleportation. This requires communicating classical bits at rate $I(A:R)$ (since each qubit requires two classical bits), and consuming maximally entangled qubits at the same rate. However, the protocol of Theorem 12.5 also *creates* entanglement at rate $\frac{1}{2}I(A:B)$, so the total required entanglement cost

$$\begin{aligned}\frac{1}{2}(I(A:R) - I(A:B)) &= \frac{1}{2}(H(A) + H(R) - H(AR) - H(A) - H(B) + H(AB)) \\ &= H(AB) - H(B) = H(A|B)\end{aligned}$$

using that $H(R) = H(AB)$ and $H(AR) = H(B)$ since ρ_{ABR} is pure. □

Applications of state merging

State merging encompasses a number of special cases, and our results for state merging directly provide achievability bounds for a number of important tasks!

(a) **Compression:** consider the case where there system B is trivial. In that case the goal is to get A to Bob (preserving the correlations with R) sending over as few qubits as possible. This is precisely the task of compression. For a pure state ρ_{AR} we have $I(A:R)_\rho = 2H(A)_\rho$, so the rate of state merging $\frac{1}{2}I(A:R)_\rho$ indeed recovers our asymptotic optimal rate $H(A)_\rho$ for compression.

(b) **Entanglement distillation:** In general we see that if Alice and Bob share (many copies of) ρ_{AB}, they can use the state merging protocol (the version of Corollary 12.8) to obtain maximally entangled qubits at rate $-H(A|B)_\rho$ (if this quantity is positive) using an LOCC protocol. In other words, state merging defines an *entanglement distillation* protocol. If the state ρ_{AB} is pure we find that $H(A|B) = -H(A)$ so we can distill entanglement at rate $H(A)$. This is consistent with our hands-on approach in Exercise 8.8. Note also that for pure ρ_{AB} the rate of classical communication is given by $I(A:R) = 0$ since ρ_{AR} is a product state. So we need only a sublinear amount of classical communication!

Let us define the notion of *entanglement distillation* more formally now, as well as the *entanglement cost*. Entanglement distillation deals with the following question: if ρ_{AB} is a state shared between Alice and Bob, how many maximally entangled states can Alice and Bob extract from ρ_{AB} using LOCC? On the other hand, the entanglement cost is how many maximally entangled qubit states are needed in order to prepare ρ_{AB} using LOCC operations. We will study the asymptotic version of this question (so we have many copies of ρ_{AB} and we would like to know at which rate we can do the conversions). To formally define this we first define the one-shot version. We let $\Phi^+_{r,A'B'}$ denote a maximally entangled state of dimension r between Alice and Bob on systems A' and B'. The question is what the minimal value of r is if we want to prepare ρ_{AB} with small error using LOCC, or conversely, what is the largest r so that we can approximately distill $|\Phi_{r,A'B'}\rangle$ from ρ_{AB} using LOCC. This is formalized by the following definition.

Definition 12.9. Let $\rho_{AB} \in \mathrm{S}(AB)$ and $\varepsilon > 0$. Then the entanglement cost of preparing ρ_{AB} with error at most ε is defined as

$$\begin{aligned} E_C^\varepsilon(\rho_{AB}) = \min\{&\log(r) \text{ such that there exists an LOCC channel} \\ &\Phi_{A'B'\to AB}, P(\Phi_{A'B'\to AB}(|\Phi^+_{r,A'B'}\rangle\langle\Phi^+_{r,A'B'}|), \rho_{AB}) \leq \varepsilon\}. \end{aligned}$$

The entanglement distillation from ρ_{AB} with error at most ε is defined as

$$\begin{aligned} E_D^\varepsilon(\rho_{AB}) = \max\{&\log(r) \text{ such that there exists an LOCC channel} \\ &\Phi_{AB\to A'B'}, P(\Phi_{AB\to A'B'}(\rho_{AB}), |\Phi^+_{r,A'B'}\rangle\langle\Phi^+_{r,A'B'}|) \leq \varepsilon\}. \end{aligned}$$

For the asymptotic version, we proceed as usual and we look at the *rate* at which we can perform the conversions if we have many copies of ρ_{AB}.

Definition 12.10. Let $\rho_{AB} \in \mathrm{S}(AB)$, then we define the (asymptotic) entanglement cost of ρ_{AB} as

$$E_C(\rho_{AB}) = \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}E_C^\varepsilon(\rho_{AB}^{\otimes n})$$

and the (asymptotic) entanglement distillation as

$$E_D(\rho_{AB}) = \lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}E_D^\varepsilon(\rho_{AB}^{\otimes n}).$$

In these definitions, ρ_{AB} is allowed to be a mixed state. For general mixed states it is difficult to compute $E_C(\rho_{AB})$ and $E_D(\rho_{AB})$ (or its one-shot versions). One fact which makes intuitive sense is that $E_C(\rho_{AB}) \geq E_D(\rho_{AB})$. If $E_D(\rho_{AB})$ were strictly larger than $E_C(\rho_{AB})$, one could first use entanglement at rate $E_C(\rho_{AB})$ to create copies of ρ_{AB} and then distill maximally entangled states at rate $E_D(\rho_{AB})$ which would give rise to an LOCC protocol generating additional entanglement, which is not possible.

Lemma 12.11. *For any state $\rho_{AB} \in \mathrm{S}(AB)$*

$$E_C(\rho_{AB}) \geq E_D(\rho_{AB})$$

The proof, in which you should make the above intuition into a rigorous argument, is Exercise 12.4. The state

Theorem 12.12. *If $\rho_{AB} \in \mathrm{S}(AB)$,*

$$E_C(\rho_{AB}) \leq \min(H(A)_\rho, H(B)_\rho).$$

Proof. We may use the following protocol:

i) Alice locally prepares $\rho_{AB}^{\otimes n}$.

ii) She compresses the B-systems to qubits at rate $H(B)_\rho$.

iii) Alice teleports these qubits to Bob, using maximally entangled qubit states at rate $H(B)_\rho$.

iv) Bob applies the decoder of the compression protocol.

By definition of the compression code, at the end of the protocol Alice and Bob share a state which is a good approximation to $\rho_{AB}^{\otimes n}$ (and the error can be taken to go to zero as n goes to infinity). □

As we observed above, state merging gives a protocol for entanglement distillation. However, this is a protocol that potentially has to 'borrow' some initial entangled qubits (it consumes entangled qubits at rate $\frac{1}{2}I(A:R)$ and distills them at rate $\frac{1}{2}I(A:B)$), which gives a net gain of entanglement if $H(A|B) < 0$. One can prove that even if one does not allow initially borrowing, one can distill entanglement at rate $-H(A|B)$.

Theorem 12.13. *For any state $\rho_{AB} \in \mathrm{S}(AB)$*

$$E_D(\rho_{AB}) \geq -H(A|B).$$

For *pure* states we can say more!

Theorem 12.14. *If $\rho_{AB} \in \mathrm{S}(AB)$ is pure,*

$$E_D(\rho_{AB}) = E_C(\rho_{AB}) = H(A)_\rho = H(B)_\rho.$$

Proof. For pure ρ_{AB} we have $-H(A|B)_\rho = H(A)_\rho$ and we get from Lemma 12.11, Theorem 12.12 and Theorem 12.13

$$E_C(\rho_{AB}) \leq H(A)_\rho \leq E_D(\rho_{AB}).$$

However, by Lemma 12.11 $E_D(\rho_{AB}) \leq E_C(\rho_{AB})$ so we must have $E_D(\rho_{AB}) = E_C(\rho_{AB}) = H(A)_\rho$. □

This establishes the entropy $H(A)_\rho$ as a good operational measure for entanglement for pure states ρ_{AB}.

12.3 State redistribution

We can also consider a situation where Alice only wants to send over a subsystem of her part of the quantum state. This is known as *state redistribution*. To formalize this, consider a state $\rho_{ABCR} \in \mathrm{S}(ABCR)$. Initially Alice holds systems A and C, and Bob holds system B. The goal is to get system A to Bob, while Alice keeps C. As usual, one can formulate a one-shot version of the task and an asymptotic version. We will focus on the asymptotic quantum state redistribution task.

This can be achieved by sending qubits at a rate given by the *conditional mutual information*. Given $\rho_{ABC} \in \mathrm{S}(ABC)$ conditional mutual information of A and B, conditioned on C, is defined as

$$I(A:B|C) := I(A:BC) - I(A:C).$$

Writing out in terms of individual entropies yields

$$I(A:B|C) = H(AC) + H(BC) - H(ABC) - H(C).$$

The fact that $I(A:B|C) \geq 0$ is equivalent to strong subadditivity! Moreover, $I(A:B|C) = I(B:A|C)$, and

$$I(A:B|C) = H(A|C) - H(A|BC) = H(B|C) - H(B|AC) \leq 2\min\{\log(|A|), \log(|B|)\} \quad (12.4)$$

by Lemma 10.3.

Theorem 12.15. *State redistribution can be achieved by sending over qubits at rate* $\frac{1}{2}I(A:R|B)$ *and consuming entanglement at rate* $\frac{1}{2}\left(I(A:C) - I(A:B)\right)$.

This can be proven by a combination of two appropriate state merging protocols, we will not give details here. Note that if the C system is trivial, this reduces to state merging, and the rates in Theorem 12.15 coincide with those in Theorem 12.15, as you can check in Exercise 12.3

This result gives an alternative proof of strong subadditivity! Recall that compression gave an operational proof of subadditivity: jointly compressing AB is at least as efficient as separately compressing A and B. In Theorem 12.15 the quantity $\frac{1}{2}I(A:R|B)$ clearly represents the quantum communication cost, which should be a non-negative number, so

$$I(A:R|B) = I(A:RB) - I(A:B) \geq 0$$

which is equivalent to data processing for the mutual information (and hence to strong subadditivity).

12.4 Converse bound

At this point we have shown that we can achieve state merging and state redistribution at certain rates. We will now show that these rates are, in fact, optimal! We will do so by proving optimality of the rate for state redistribution, as it has state merging as a special case.

Let us first sketch the idea, for convenience pretending we are redistributing a single copy. Let ω denote the state received by Bob, and σ the final state of the protocol (so $\sigma_{ABR} \approx \rho_{ABR}$). The idea is that when sending Q we have a bound for the conditional mutual information

$$I(R:Q|B) \leq 2\log(|Q|)$$

while on the other hand by data processing

$$I(R:QB)_\omega \geq I(R:AB)_\sigma \approx I(R:AB)_\rho$$

using that the final state $\sigma_{ABR} \approx \rho_{ABR}$. Furthermore, $I(R:B)_\omega = I(R:B)_\rho$ (since $\omega_{RB} = \rho_{RB}$). This means that (up to a small error)

$$2\log(|Q|) \geq I(R:QB)_\omega - I(R:B)_\omega \geq I(R:AB)_\rho - I(R:B)_\rho = I(R:A|B)_\rho.$$

The key ingredients we use in the proof are data processing for the mutual information and a continuity estimate for the mutual information in order to control the errors.

Theorem 12.16. *State redistribution is not possible at rates smaller than $\frac{1}{2}I(A:R|B)$. In particular, for quantum state merging*

$$q(A:B:R)_\rho \geq \frac{1}{2}(A:R)_\rho$$

Proof. Suppose that q is an achievable rate for state redistribution. Then for any ε and δ there must exist n and a protocol which redistributes $\rho_{ABCR}^{\otimes n}$ using at most $n(q+\delta)$ qubits of communication, with error at most ε. Let $\sigma_{A^nB^nC^nR^n}$ denote the state after the state redistribution, which by assumption satisfies

$$T(\rho_{ABCR}^{\otimes n}, \sigma_{A^nB^nC^nR^n}) \leq P(\rho_{ABCR}^{\otimes n}, \sigma_{A^nB^nC^nR^n}) \leq \varepsilon.$$

By the continuity of the mutual information in Eq. (10.5) this implies

$$|I(R^n:A^nB^n)_\sigma - I(R^n:A^nB^n)_{\rho^{\otimes n}}| \leq \underbrace{4\varepsilon\log(|R^n|)}_{=4n\varepsilon\log(|R|)} + \frac{2}{1+\varepsilon}h(\frac{\varepsilon}{1+\varepsilon}).$$

On the other hand, we obtained σ by applying quantum channels and sending over at most $n(q+\delta)$ qubits. Let $\omega_{QB^nR^n}$ denote the state received by Bob; note that Bob and Robin have not done anything so $\omega_{B^nR^n} = \rho_{B^nR^n}$. By data processing,

$$I(R^n:A^nB^n)_\sigma \leq I(R^n:QB^n)_\omega.$$

By Eq. (12.4)

$$\begin{aligned}2n(q+\delta) &\geq 2\log(|Q|) \geq I(R^n:Q|B^n)_\omega = I(R^n:QB^n)_\omega - I(R^n:B^n)_\omega \\ &\geq I(R^n:A^nB^n)_\sigma - I(R^n:B^n)_\rho \\ &\geq I(R^n:A^nB^n)_\rho - I(R^n:B^n)_\rho - 4n\varepsilon\log(|R|) - \frac{2}{1+\varepsilon}h(\frac{\varepsilon}{1+\varepsilon})\end{aligned}$$

We conclude that

$$\begin{aligned}2q &\geq \frac{1}{n}I(R^n:A^n|B^n)_\rho - 4\varepsilon\log(|R|) - \delta - \frac{2}{n(1+\varepsilon)}h(\frac{\varepsilon}{1+\varepsilon}) \\ &= I(R:A|B)_\rho - f(\varepsilon,\delta,n)\end{aligned}$$

where $f(\varepsilon,\delta,n)$ goes to zero as we let ε,δ go to zero. We conclude that

$$q \geq \frac{1}{2}I(R:A|B)_\rho = \frac{1}{2}I(A:R|B)_\rho.$$

□

Note that in the proof it was important that the dependence of the continuity for the entropy depended *logarithmically* on the dimension and hence *linearly* on the number of qubits; hopefully this makes you appreciate the continuity bound in Theorem 10.9!

12.5 Exercises

12.1 **Partial trace gymnastics:** Verify Eq. (12.2) and Eq. (12.3).

12.2 **Managing entanglement:** Alice and Bob share many copies of the states ρ_{ABR} and σ_{ABR}, which have some correlations with the environment system $\mathcal{H}_R$.

Alice and Bob aim to perform the quantum state merging protocol many times, to merge n_ρ copies of the ρ_{ABR} state and n_σ copies of the σ_{ABR} state with Bob's system.

(a) Consider the case where Alice and Bob share no other entanglement, but have access to unlimited classical communication. Show that the value (n_ρ, n_σ) is achievable if

$$n_\rho H(A|B)_\rho + n_\sigma H(A|B)_\sigma \leq 0 \ .$$

(b) Comment on what happens in the case $H(A|B)_\rho = H(A|B)_\sigma = 0$.

(c) Now suppose that Alice's internet service provider has imposed a limit of $N \gg 1$ bits of classical communication with Bob. What constraint does this impose on the achievable values of (n_ρ, n_σ)?

(d) Assume $d_A = d_B = d_R = 2$. Let

$$\rho_{ABR} = \frac{3}{4}|\Phi^+_{AB}\rangle\langle\Phi^+_{AB}| \otimes \frac{1}{2}\mathbb{1}_R + \frac{1}{4}|\Phi^+_{AR}\rangle\langle\Phi^+_{AR}| \otimes \frac{1}{2}\mathbb{1}_B \ ,$$

where $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle$, and let $\sigma_{ABR} = \frac{1}{8}\mathbb{1}_{ABR}$. Sketch a diagram of the achievable values for (n_ρ, n_σ) in this case.

12.3 **From state redistribution to merging:** Show that for a pure state ρ_{ABR}, we have $I(A:R|B) = I(A:R)$.

12.4 **Entanglement cost and distillation:**

(a) Show that $|\Phi^+_{AB}\rangle$ is a maximally entangled state of dimension d. Show that if $\rho_{AB} \in \mathrm{S}(AB)$ has entanglement rank at most r, then

$$F\big(\rho_{AB}, |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|\big) \leq \frac{r}{d}.$$

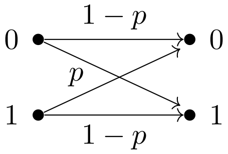
(b) Prove Lemma 12.11. *Hint: suppose that $E_D(\rho_{AB}) > E_C(\rho_{AB})$ and derive a contradiction.*

12.5 **Conversion between arbitrary states:** Show that Alice and Bob can asymptotically convert a pure state ρ_{AB} into a pure state σ_{AB} at rate $H(\rho_A)/H(\sigma_A)$ using LOCC.

Lecture 13

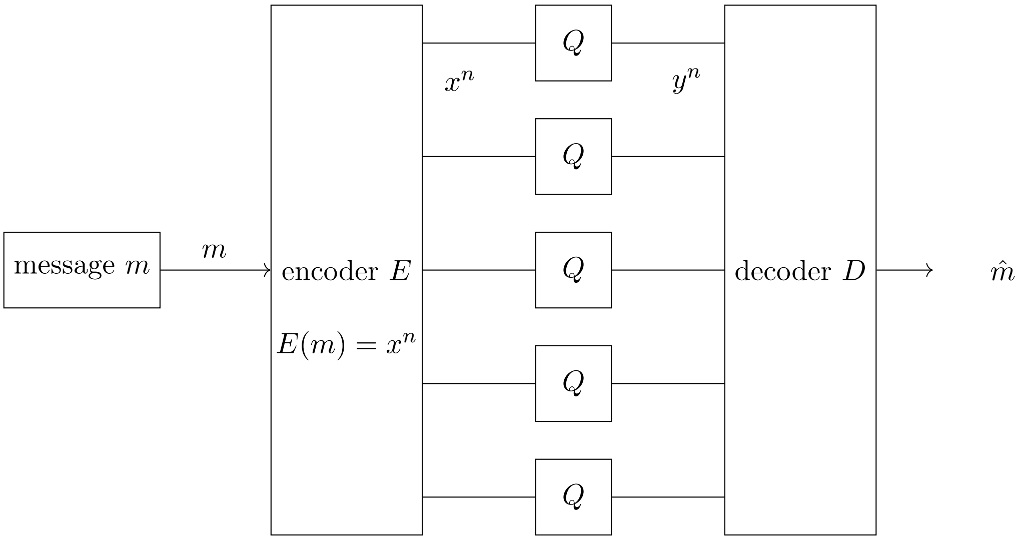
Quantum capacity

One of the basic tasks of information theory is to *reliably transfer information over a noisy channel.* In the classical setting this task has been achieved with spectacular success in practical applications: modern communication technology is able to reliable transfer information at very high rates using (for example) electromagnetic waves. As a very simple (but useful) model, consider the binary symmetric channel from Example 4.3 which with some probability p flips the value of a single bit:



This models a noisy communication channel, which sends over a single bit, but with probability p introduces an error. The trick to use *multiple* uses of this channel to reliably send information is to introduce *redundancy* in the information you send over. The most basic example is the following: we want to send a single bit x which is 0 or 1. We use the channel three times and *encode* the message by repeating it three times: $x \mapsto xxx$. On the receiving side we obtain $y = y_1y_2y_3$. We assume that $p < \frac{1}{2}$. The receiver would like to know the bit x. What is their best guess, based on y? The intuitive answer is that the best guess is simply a majority vote for the bits of y. For example, if $y = 010$, then we guess that the original message was 0, and there has been a single bit flip on 000. It is clear that if we were only allowed to use the channel once, the probability of error is p. You can show (Exercise 13.1) that using the channel three times as described above will improve the error probability.

The above scheme is known as a *repetition code*. It is an example of *error correcting codes*. An error correcting code for a (classical) channel Q from X to Y consists of an encoding map, encoding messages m into code words of length n on X^n, and a decoding map, which associates a guess $\hat{m}$ to an outcome in Y^n. This is described by the following diagram:



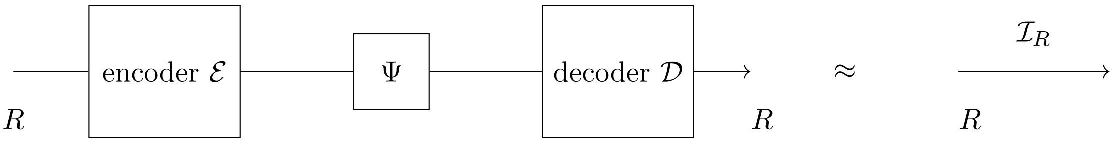
Good error correcting codes are such that

(a) The probability of error is small: $\Pr(\hat{m} \neq m)$ is small.

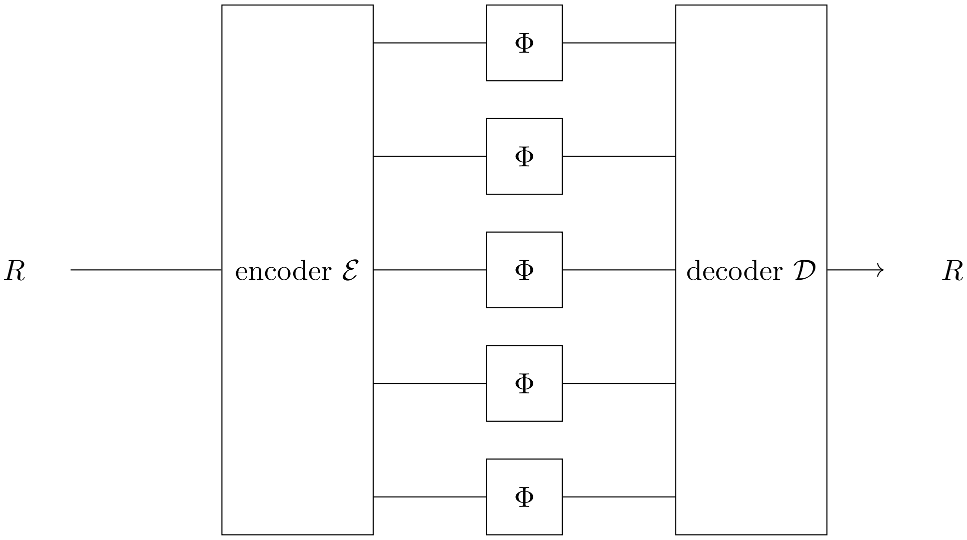
(b) The amount of redundancy is not overly large. The amounts of bits we send is $N = \log(|M|)$, and we would like the number of channel uses per bit of information transferred, $\frac{n}{N}$, to be as small as possible.

In this lecture we will study an analogous situation, where Alice tries to send *quantum information* to Bob, by using a *quantum channel* $\Phi_{A\to B}$. We will compute the optimal rate at which Alice can send qubits to Bob by using the channel $\Phi_{A\to B}$ many times. Let us start by defining what a quantum error correcting code should be.

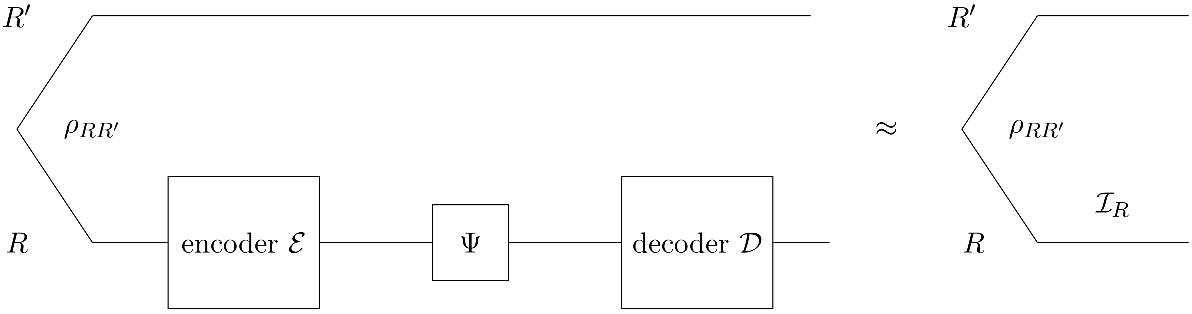
The set-up will be in close analogy to the classical situation. What we would like to achieve is that Alice and Bob can simulate an identity channel on a system R from Alice to Bob, by using some channel $\Psi_{A\to B}$. That means that there should exist an encoding channel $\mathcal{E} \in \mathrm{C}(R, A)$ and a decoding channel $\mathcal{D} \in \mathrm{C}(B, R)$ such that $\mathcal{D} \circ \Psi_{A\to B} \circ \mathcal{E} \approx \mathcal{I}_R$



To capture the rate at which we can transfer information over a channel $\Phi_{A\to B}$, we apply this scenario to the channel $\Psi_{A^n\to B^n} = \Phi_{A\to B}^{\otimes n}$



We characterize closeness to the identity channel using the entanglement purified distance. For every state ρ_R and purification $\rho_{RR'}$ (where we take R' to be a copy of R) we should have



which we capture in the following definition.

Definition 13.1. An (r, ε)-error correcting code for $\Psi_{A\to B}$ consists of quantum channels $\mathcal{E} \in \mathrm{C}(R, A)$ and $\mathcal{D} \in \mathrm{C}(B, R)$ for a system R with $\log(|R|) \geq r$ such that

$$P_E(\mathcal{D} \circ \Psi_{A\to B} \circ \mathcal{E}, \rho_R) \leq \varepsilon \text{ for all } \rho_R \in \mathrm{S}(R).$$

We denote by $Q^\varepsilon(\Psi)$ the optimal number of qubits we can send using $\Psi_{A\to B}$ with error at most ε:

$$Q^\varepsilon(\Psi) = \max_r \{r\colon \text{there exists an } (r, \varepsilon)\text{-error correcting code for } \Psi_{A\to B}\ \}.$$

If Alice and Bob are using an (r, ε)-error correcting code for $\Psi_{A\to B}$, and Alice starts a maximally entangled state on RR', then by assumption Alice and Bob end up with a state which is close to maximally entangled. This means that Alice and Bob can use the channel $\Psi_{A\to B}$ to establish entanglement. We may also study *entanglement generating codes* for a quantum channel $\Psi_{A\to B}$.

Definition 13.2. An (r, ε)-entanglement generating code for $\Psi_{A\to B}$ consists of quantum channels $\mathcal{E} \in \mathrm{C}(R, A)$ and $\mathcal{D} \in \mathrm{C}(B, R)$ for a system R with $\log(|R|) \geq r$ such that

$$P_E(\mathcal{D} \circ \Psi_{A\to B} \circ \mathcal{E}, \tau_R)$$

where τ_R is the maximally mixed state.

We denote by $Q^{\varepsilon}_{\mathrm{EG}}(\Psi)$ the optimal number of maximally entangled qubits we can generate using $\Psi_{A\to B}$ with error at most ε:

$$Q^{\varepsilon}_{\mathrm{EG}}(\Psi) = \max_r\{r\colon \text{ there exists an } (r, \varepsilon)\text{-entanglement generating code for } \Psi_{A\to B}\ \}.$$

It is clear from the definition that every (r, ε)-error correcting code is also an (r, ε)-entanglement generating code, so

$$Q^{\varepsilon}_{\mathrm{EG}}(\Psi) \geq Q^{\varepsilon}(\Psi) \tag{13.1}$$

There is a converse to this. If we have an entanglement generating code, it must accurately transfer half of a maximally entangled state $|\Phi^+_{RR'}\rangle$. It is possible to show that if one has an entanglement generating code, then one can find a subsystem (which is only one qubit smaller) such that when Alice and Bob restrict to this subsystem, *every* state is transmitted reliably, and they have an error correcting code. This is captured by the following result, which we will not prove.

Theorem 13.3. *Let $\Psi_{A\to B} \in \mathrm{C}(A, B)$, let $\varepsilon > 0$ and let $\delta = \sqrt{8\varepsilon}$. Then*

$$Q^{\delta}(\Psi) \geq Q^{\varepsilon}_{\mathrm{EG}}(\Psi) - 1.$$

See 19.1.2 in [38] for a proof. The consequence is that when we consider asymptotic scenarios, generating entanglement is equivalent to sending over quantum information.

Given a quantum channel $\Phi_{A\to B}$, we would like to know the optimal rate of quantum communication. We are allowed to use the channel many times: using the channel n times corresponds to the channel $\Phi^{\otimes n}_{A\to B}$. The rate of communication is the number of qubits we can communicate per channel use, as we allow arbitrarily small error and let the number of channel uses go to ∞:

Definition 13.4. The quantum capacity and the entanglement generating quantum capacity of a quantum channel $\Phi_{A\to B} \in \mathrm{C}(A, B)$ are defined as

$$Q(\Phi) := \lim_{\varepsilon\to 0} \lim_{n\to\infty} \frac{1}{n} Q^{\varepsilon}(\Phi^{\otimes n})$$

and

$$Q_{\mathrm{EG}}(\Phi) := \lim_{\varepsilon\to 0} \lim_{n\to\infty} \frac{1}{n} Q^{\varepsilon}_{\mathrm{EG}}(\Phi^{\otimes n})$$

respectively.

By Theorem 13.3 and Eq. (13.1) we have

$$Q(\Phi) = Q_{\mathrm{EG}}(\Phi),$$

and we will simply refer to this quantity as the *quantum capacity* of the channel $\Phi_{A\to B}$.

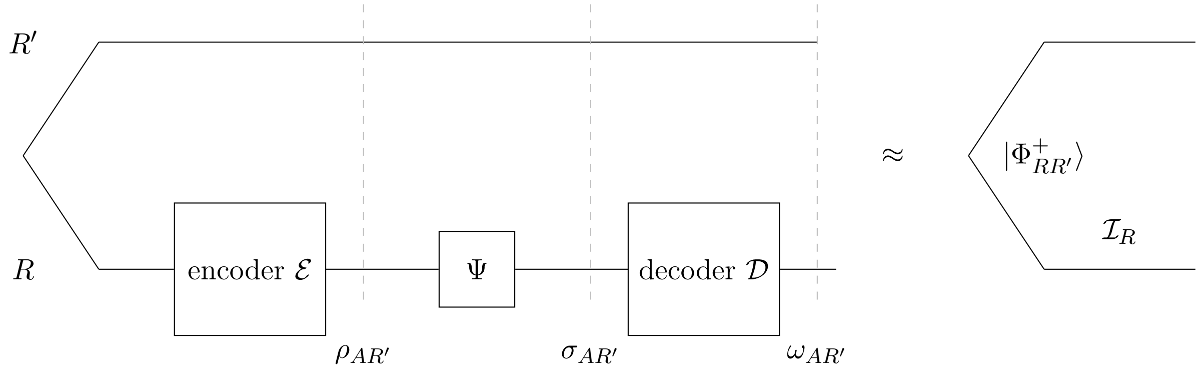
13.1 The coherent information

For quantum information processing tasks we encountered before we saw that the asymptotic rate that characterized the task was computed by an entropic quantity. For compression this was simply the entropy; for state merging we encountered the mutual information (and variations depending on the task). Can we also find an entropic quantity that characterizes the quantum capacity?

There is indeed such a quantity. It is easiest to find it as a *upper bound* (i.e. we see that the rate can at most be a certain quantity). We will first give an informal derivation, ignoring the asymptotics and error terms. It is, once again, based on a data processing inequality (returning to the theme that strong subadditivity is the fundamental fact that bounds information processing protocols).

Suppose that Alice and Bob have an entanglement generating code for $\Phi_{A\to B}$ and are able to generate a state $\omega_{RR'}$ which is close to a maximally entangled state. Let $\rho_{AR'}$ be the state that Alice prepares, and $\sigma_{BR'}$ the state that Bob receives. By data processing (and continuity of the conditional entropy)

$$H(R'|A)_\rho \leq H(R'|B)_\sigma \leq H(R'|R)_\omega \approx -\log(|R|)$$



In conclusion, we must have that the number of qubits r

$$r \leq -H(R'|B). \tag{13.2}$$

We will make this argument precise below in Theorem 13.6. This discussion motivates the introduction of the *coherent information* of a quantum channel.

Definition 13.5. Given a quantum channel $\Phi_{A\to B}$, the coherent information of $\Phi_{A\to B}$ is defined as

$$I_c(\Phi) = \max_{\rho_{AR'}} -H(R'|B)_\sigma \tag{13.3}$$

where the maximum is over all pure states $\rho_{AR'} = |\phi_{AR'}\rangle\langle\phi_{AR'}| \in \mathrm{S}(AR')$ and $\sigma_{BR'} = (\Phi_{A\to B} \otimes \mathcal{I}_{R'})(\rho_{AB})$.

Note that in this maximum we may restrict to a system which is a copy of A, so $R' = A'$. We also could have maximized over states which are not necessarily pure; data processing implies that in that case the maximum can be attained by a pure state (so assuming the state is pure isom not really a restriction). To get more intuition for this quantity, note that if we let $V \in \mathrm{Isom}(A, BE)$

be a Stinespring extension of $\Phi_{A\to B}$ and $\sigma_{BER'}$ the pure state $(V \otimes \mathbb{1}_{R'})|\phi_{AR'}\rangle$ (with $|\phi_{AR'}\rangle$ the state which maximizes Eq. (13.3)) then

$$I_c(\Phi) = -H(R'|B)_{\sigma_{BR'}} = H(B)_\sigma - H(E)_\sigma$$

using that $H(BR')_\sigma = H(E)_\sigma$ for the pure state $\sigma_{BER'}$. In other words, the coherent information is the amount of information arriving at B, minus the amount of information that is lost to the environment E (optimized over choice of initial state).

We saw informally that there must be some state $\rho_{AR'}$ such that $-H(R'|B)_\sigma$ is an upper bound on the capacity. However, that argument did *not* take into account the fact that to bound the *rate* we need to apply our argument to many copies of the channel! While it is true that

$$H(R'^n|B^n)_{\sigma^{\otimes n}} = nH(R'|B)_\sigma$$

we do not need to start with a product state $\rho_{AR'}^{\otimes n}$. In other words, when we compute $I_c(\Phi^{\otimes n})$, it is not clear that

$$I_c(\Phi^{\otimes n}) \overset{?}{=} nI_c(\Phi).$$

It turns out that in general we do *not* have equality here. The relevant quantity for the rate will then be

$$\lim_{n\to\infty} \frac{1}{n} I_c(\Phi^{\otimes n}).$$

Let us now make this more precise and prove an upper bound on the quantum capacity.

Theorem 13.6. *For $\Phi_{A\to B} \in \mathrm{C}(A, B)$ the quantum capacity is upper bounded as*

$$Q(\Phi) \leq \lim_{n\to\infty} \frac{1}{n} I_c(\Phi^{\otimes n}).$$

Proof. Let $\varepsilon, \delta > 0$. There must exist N such that for all $n > N$ there exists an (r, ε)-error correcting code for $\Phi_{A\to B}^{\otimes n}$ with

$$r = n(Q(\Phi) - \delta)$$

with encoder $\mathcal{E} \in \mathrm{C}(R, A^n)$ and decoder $\mathcal{D} \in \mathrm{C}(B^n, R)$. Let $\omega_{RR'}$ be a maximally entangled state, let $\rho_{A^nR'} = (\mathcal{E} \otimes \mathbb{1}_{R'})(\omega_{RR'})$. Denote by $\sigma_{B^nR'}$ the result of applying $\Phi_{A\to B}^{\otimes n}$ to $\rho_{A^nR'}$ and let $\tilde{\omega}_{RR'}$ be the state that we get when Bob applies the decoder $\mathcal{D}$ to $\sigma_{B^nR'}$. By assumption, $P(\tilde{\omega}_{RR'}, \omega_{RR'}) \leq \varepsilon$. Now, by data processing and by the definition of the coherent information of $\Phi_{A\to B}$

$$-H(R'|R)_{\tilde{\omega}} \leq -H(R'|B^n)_\sigma \leq I_c(\Phi^{\otimes n}).$$

On the other hand, as $T(\tilde{\omega}_{RR'}, \omega_{RR'}) \leq P(\tilde{\omega}_{RR'}, \omega_{RR'}) \leq \varepsilon$ by the Fuchs-van de Graaf inequality Eq. (6.9), Theorem 10.9 implies

$$\left| H(R'|R)_{\tilde{\omega}} - H(R'|R)_\omega \right| \leq 2\varepsilon \log(|R|) + (1+\varepsilon) h\left(\frac{\varepsilon}{1+\varepsilon}\right).$$

Since $\omega_{RR'}$ is maximally entangled, $H(R'|R)_\omega = -\log(|R|)$, so

$$\log(|R|) \leq \frac{1}{1+2\varepsilon} \left(-H(R'|R)_{\tilde{\omega}} + (1+\varepsilon) h\left(\frac{\varepsilon}{1+\varepsilon}\right) \right).$$

Combining these observations,

$$\begin{aligned} n(Q(\Phi)-\delta) &\le \log(|R|) \\ &\le \frac{1}{1+2\varepsilon}\left(-H(R'|R)_{\tilde{\omega}} + (1+\varepsilon)h\left(\frac{\varepsilon}{1+\varepsilon}\right)\right) \\ &\le \frac{1}{1+2\varepsilon}\left(I_c(\Phi^{\otimes n}) + (1+\varepsilon)h\left(\frac{\varepsilon}{1+\varepsilon}\right)\right). \end{aligned}$$

We conclude that as we let $n \to \infty$

$$Q(\Phi) \le \frac{1}{1+2\varepsilon} \lim_{n\to\infty} \frac{1}{n} I_c(\Phi^{\otimes n}) + \delta.$$

Since ε and δ can be taken to be arbitrarily small,

$$Q(\Phi) \le \lim_{n\to\infty} \frac{1}{n} I_c(\Phi^{\otimes n}).$$

□

13.2 Random coding and decoupling

Our next goal is to argue that there exists a matching lower bound for the quantum capacity. To this end, we have to argue that there exist error correcting codes with good rates. The approach will be very similar to the one in previous chapter!

The fact that random unitaries are decoupling keeps on giving! We start with a variation on Theorem 11.10 for the coding problem. This will serve as the main ingredient for a one-shot protocol for entanglement generation. To conclude, we show that the one-shot protocol gives a rate which matches the upper bound from Theorem 13.6 and is optimal.

Let $|\phi_{AA'}\rangle$ be pure quantum state, where A' is a copy of A. Let $V \in \mathrm{Isom}(A, BE)$ be a Stinespring extension of a channel $\Psi_{A\to B}$. Let

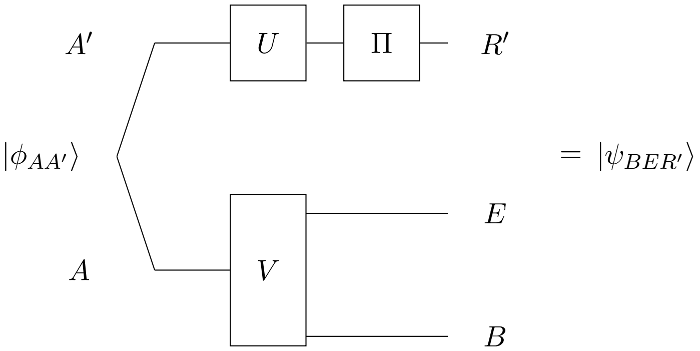
$$|\psi_{BEA'}\rangle = (V \otimes \mathbb{1}_{A'})|\phi_{AA'}\rangle \tag{13.4}$$

be the result of applying the Stinespring extension.

Choose a unitary $U \in \mathrm{U}(A)$ uniformly at random and let Π be a projection onto a subspace $\mathcal{H}_R \subseteq \mathcal{H}_A$. Then we can consider

$$|\theta_{BER'}\rangle = \sqrt{\frac{|A|}{|R|}}(\mathbb{1}_{BE} \otimes (\Pi U))|\phi_{BEA'}\rangle, \tag{13.5}$$

which is a not necessarily normalized state.

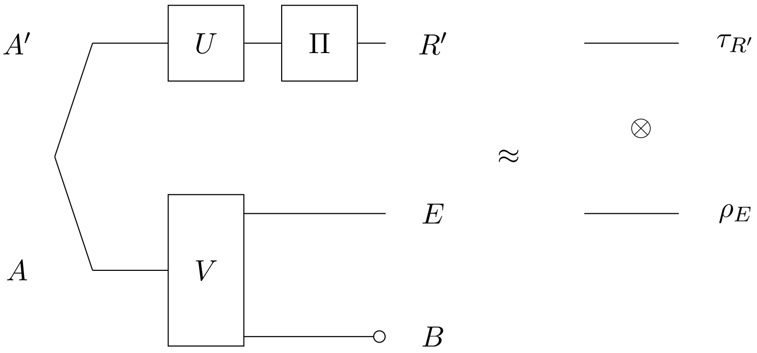


One can also read this as applying a (uniformly) random projection onto a subspace of dimension $|R|$. The factor $\sqrt{\frac{|A|}{|R|}}$ guarantees that on average the state is normalized. We have the following decoupling result:

Theorem 13.7. *Let $\rho_{BEA'} = |\psi_{BEA'}\rangle\langle\psi_{BEA'}|$ and $\sigma_{BER'} = |\theta_{BER'}\rangle\langle\theta_{BER'}|$ be defined as above. Then*

$$\mathbb{E}_U \|\sigma_{ER'} - \rho_E \otimes \tau_{R'}\|_1^2 \leq |R||E| \operatorname{tr}[\rho_{EA'}^2].$$

The proof is closely analogous to that of Theorem 11.10. The result may be visualized as



This decoupling result implies a one-shot coding theorem, which is in spirit very close to the one-shot state merging result of Lemma 12.6. Note that the states $\sigma_{BER'} = |\theta_{BER'}\rangle\langle\theta_{BER'}|$ (which depend on the random unitary) need not be normalized. However, we have the following fact (Exercise 13.2): if $\rho_A \in \mathrm{S}(A)$ and $\sigma_A \in \mathrm{PSD}(A)$, then

$$\|\rho_A - \frac{\sigma_A}{\operatorname{tr}[\sigma_A]}\|_1 \leq 2\|\rho_A - \sigma_A\|_1. \tag{13.6}$$

The following is a direct consequence of Lemma 12.2:

Lemma 13.8. *Let $\Psi_{A\to B}$ be a quantum channel with Stinespring extension $V \in \mathrm{Isom}(A, BE)$ and let $|\phi_{AR'}\rangle$ be a pure state. Let $\rho_{BER'} = |\psi_{BER'}\rangle\langle\psi_{BER'}|$ be defined by*

$$|\psi_{BER'}\rangle = (V \otimes \mathbb{1}_{R'})|\phi_{AR'}\rangle.$$

If there exists some state $\sigma_E \in \mathrm{S}(E)$ such that

$$P(\rho_{ER'}, \sigma_E \otimes \tau_{R'}) \leq \varepsilon$$

then there exists an (r, ε)-entanglement generating code for $\Psi_{A\to B}$ with $r = \log(|R'|)$.

Proof. It suffices to find a decoder channel on B which maps $\sigma_{BR'}$ to an ε-approximation of a maximally entangled state. Let R be a copy of R', let $\omega_{RR'}$ be a maximally entangled state and let $\sigma_{EE'}$ be a purification of σ_E. By the decoupling principle Corollary 12.3 we find an isometry $W \in \mathrm{Isom}(B, AR')$ such that the state

$$(W \otimes \mathbb{1}_{ER'})\rho_{BER'}(W^\dagger \otimes \mathbb{1}_{ER'})$$

is ε-close in purified distance. If we define the decoder as

$$\mathcal{D}[M_B] = \operatorname{tr}_{E'}[W M_B W^\dagger]$$

we see that by monotonicity of the purified distance

$$P((\mathcal{D}\otimes\mathcal{I}_{R'})(\rho_{BR'}),\omega_{RR'}).$$

□

We apply this one-shot coding theorem to the situation where we apply n copies of a channel $\Phi_{A\to B}$. The quantum capacity will arise from a two-layer argument: we will first show that the capacity is at least $I_c(\Phi)$. Next, we apply this result to the channel $\Phi_{A\to B}^{\otimes n}$ to see that for every n, the capacity is at least $\frac{1}{n}I_c(\Phi^{\otimes n})$.

Theorem 13.9. *Let* $\Phi_{A\to B}\in \mathrm{C}(A,B)$. *Then*

$$Q(\Phi)\geq I_c(\Phi).$$

Proof. Let $|\phi_{AA'}\rangle$ be such that $I_c(\Phi)=-H(A'|B)_\rho$ for $\rho_{BA'}=(\Phi_{A\to B}\otimes\mathcal{I}_{A'})(\rho_{AA'})$. We let $V\in\mathrm{Isom}(A,BE)$ be a Stinespring extension of $\Phi_{A\to B}$, so

$$|\psi_{BEA'}\rangle=(V\otimes\mathbb{1}_{R'})|\phi_{AA'}\rangle$$

is a purification of $\rho_{BA'}$. Let $\varepsilon,\delta>0$ be arbitrary. By Lemma 12.7 there is an N such that for $n\geq N$, the pure state $\tilde\rho_{B^nE^nR'^n}$ defined by applying typical subspace projections and normalizing

$$|\tilde\psi_{B^nE^nA'^n}\rangle:=\frac{(\Pi_{B,n,\delta}\otimes\Pi_{E,n,\delta}\otimes\Pi_{A',n,\delta})|\psi_{BEA'}\rangle^{\otimes n}}{\|(\Pi_{B,n,\delta}\otimes\Pi_{E,n,\delta}\otimes\Pi_{A',n,\delta})|\psi_{BEA'}\rangle^{\otimes n}\|}$$

is such that $P(\rho_{BEA'}^{\otimes n},\tilde\rho_{B^nE^nA'^n})\leq\varepsilon$ and

$$\mathrm{tr}[\tilde\rho_{E^nR'^n}^2]\leq 2^{-n(H(B)-3\delta)+1}. \tag{13.7}$$

We choose R to be a system of

$$\log(|R|)=\lfloor n(H(B)-H(E)-4\delta)-2\log(\varepsilon)-1\rfloor \tag{13.8}$$

qubits. We now let U be random unitary on $S_{n,\delta}(\rho_A)$ and Π be a projection onto a subspace $\mathcal{H}_R\subseteq S_{n,\delta}(\rho_A)$, and define $\tilde\sigma_{B^nE^nR'}$ as in Eq. (13.5) as $|\tilde\theta_{B^nE^nR'}\rangle\langle\tilde\theta_{B^nE^nR'}|$

$$|\tilde\theta_{B^nE^nR'}\rangle=\sqrt{\frac{|S_{n,\delta}(\rho_A)|}{|R|}}(\mathbb{1}_{B^nE^n}\otimes(\Pi U))|\tilde\psi_{B^nE^nA'^n}\rangle.$$

This need not be normalized. Using that the operator has support on the typical subspace, by Theorem 13.7 when we average over U

$$\begin{aligned}\mathbb{E}_U\|\tilde\sigma_{E^nR'}-\tilde\rho_{E^n}\otimes\tau_{R'}\|_1^2&\leq|R||S_{n,\delta}(\rho_E)|\,\mathrm{tr}[\tilde\rho_{EA'}^2]\\&\leq|R|2^{n(H(E)-H(B)+4\delta)+1}\\&\leq\varepsilon^2.\end{aligned} \tag{13.9}$$

We used Eq. (13.7), the fact that $\tilde\rho_{E^n}$ is supported on the typical subspace and our choice of number of qubits in R' by Eq. (13.8). The state $\tau_{R'}$ is a maximally mixed state. We now let $\sigma_{B^nE^nR'}$ be $|\theta_{B^nE^nR'}\rangle\langle\theta_{B^nE^nR'}|$

$$|\theta_{B^nE^nR'}\rangle=\sqrt{\frac{|S_{n,\delta}(\rho_A)|}{|R|}}(\mathbb{1}_{B^nE^n}\otimes(\Pi U))|\psi_{B^nE^nA'^n}\rangle.$$

(so the difference with $\tilde{\sigma}_{B^nE^nR'}$ is that we did not apply the typical subspace projections). Again, $\sigma_{B^nE^nR'}$ need not be normalized. We now compute average over the choice of random unitary and apply the triangle inequality:

$$\begin{aligned}\mathbb{E}_U\|\sigma_{E^nR'} - \tilde{\rho}_{E^n} \otimes \tau_{R'}\|_1 \leq \mathbb{E}_U\|\sigma_{E^nR'} - \tilde{\sigma}_{E^nR'}\|_1 \\ + \mathbb{E}_U\|\tilde{\sigma}_{E^nR'} - \tilde{\rho}_{E^n} \otimes \tau_{R'}\|_1.\end{aligned}$$

The first term is given by

$$\mathbb{E}_U\|\sigma_{E^nR'} - \tilde{\sigma}_{E^nR'}\|_1 \leq \|\rho_{E^nR'} - \tilde{\rho}_{E^nR'}\|_1 \leq \varepsilon \tag{13.10}$$

by Exercise 13.3. By Eq. (13.9) and Jensen's inequality the second term is bounded as

$$\begin{aligned}\mathbb{E}_U\|\tilde{\sigma}_{E^nR'} - \tilde{\rho}_{E^n} \otimes \tau_{R'}\|_1 &\leq \sqrt{\mathbb{E}_U\|\tilde{\sigma}_{E^nR'} - \tilde{\rho}_{E^n} \otimes \tau_{R'}\|_1} \\ &\leq \varepsilon.\end{aligned}$$

Since

$$\mathbb{E}_U\|\sigma_{E^nR'} - \tilde{\rho}_{E^n} \otimes \tau_{R'}\|_1 \leq 2\varepsilon$$

we may choose *some* unitary U such that for this choice of unitary $\|\sigma_{E^nR'} - \tilde{\rho}_{E^n} \otimes \tau_{R'}\|_1 \leq 2\varepsilon$. By the Fuchs-van de Graaf inequalities and Eq. (13.6) when we normalize $\sigma_{B^nE^nR'}$ we have found a state such that

$$P(\sigma_{E^nR'}, \tilde{\rho}_{E^n} \otimes \tau_{R'}) \leq \sqrt{2\varepsilon}.$$

By Lemma 13.8 we conclude that we have found a $(\log(|R|), \sqrt{2\varepsilon})$-code for $\Phi_{A\to B}^{\otimes n}$, and

$$Q^{\sqrt{2\varepsilon}}(\Phi) \geq \frac{1}{n}\log(|R|) \geq H(B) - H(E) - 4\delta - \frac{1}{n}(2\log(\varepsilon) - 2).$$

Since ε, δ were arbitrary, we conclude that

$$Q(\Phi) \geq H(B) - H(E) = H(B) - H(A') = H(B|A') = I_c(\Phi).$$

□

After all this hard work, we can finally state a complete characterization of the quantum capacity!

Theorem 13.10. *The quantum capacity of a quantum channel is given by*

$$Q(\Phi) = \lim_{n\to\infty} \frac{1}{n} I_c(\Phi^{\otimes n}).$$

Proof. By Theorem 13.6 we have

$$Q(\Phi) \leq \lim_{n\to\infty} \frac{1}{n} I_c(\Phi^{\otimes n}).$$

On the other hand, it is clear that

$$Q(\Phi) = \lim_{\varepsilon\to 0}\lim_{n\to\infty} \frac{1}{n} Q^\varepsilon(\Phi^{\otimes n}) = \lim_{\varepsilon\to 0}\lim_{m\to\infty} \frac{1}{nm} Q^\varepsilon(\Phi^{\otimes nm})$$

$$
\begin{aligned}
&= \frac{1}{n} \lim_{\varepsilon \to 0} \frac{1}{m} Q^{\varepsilon}((\Phi^{\otimes n})^{\otimes m}) \\
&\geq \frac{1}{n} I_c(\Phi^{\otimes n})
\end{aligned}
$$

using Theorem 13.9 in the last inequality. Taking $n \to \infty$ we conclude that

$$Q(\Phi) = \lim_{n \to \infty} \frac{1}{n} I_c(\Phi^{\otimes n}).$$

□

While this is a beautiful theorem, quantum information theorists sometimes feel unhappy about this result. The reason is that it does not directly give an easy-to-compute quantity: $I_c(\Phi^{\otimes n})$ involves an optimization problem over a space exponentially large in n, and one has to take a limit of n to infinity... For certain channels, the coherent information has the property that $I_c(\Phi^{\otimes 2}) = 2I_c(\Phi)$ (in which case one says that the capacity is *additive*). In this situation we immediately see that we get the much nicer formula $Q(\Phi) = I_c(\Phi)$. However, there exist examples of channels for which $I_c(\Phi^{\otimes 2}) > 2I_c(\Phi)$ and this simplification does not occur.

13.3 Exercises

13.1 **Repetition code:** [MW: Missing.]

13.2 **Trace distance and normalization:** Let $\rho_A \in \mathrm{S}(A)$ and let $P_A \in \mathrm{PSD}(A)$. Suppose that $\|\rho_A - P_A\|_1 \leq \varepsilon$. Show that $|\mathrm{tr}[P_A] - 1| \leq \varepsilon$, and use that to show that for

$$\sigma_A = \frac{P_A}{\mathrm{tr}[P_A]}, \qquad \|\rho_A - \sigma_A\| \leq \varepsilon.$$

13.3 **Averaged trace norm:** Suppose $X_A \in \mathrm{Lin}(A)$ is Hermitian, $P_A \in \mathrm{Lin}(A)$, and U_A is drawn from the Haar distribution on $\mathrm{U}(A)$. We would like to compute

$$\mathbb{E}_U \|P_A U_A X_A U_A^\dagger P_A^\dagger\|_1.$$

(a) Show that there exist positive operators X_A^+ and X_A^- such that

$$\|X_A\|_1 = \mathrm{tr}[X_A^+] + \mathrm{tr}[X_A^-]$$

and

$$\|P_A U_A X_A U_A^\dagger P_A^\dagger\|_1 = \mathrm{tr}[P_A U_A X_A^+ U_A^\dagger P_A^\dagger] + \mathrm{tr}[P_A U_A X_A^- U_A^\dagger P_A^\dagger].$$

(b) Show that

$$\mathbb{E}_U \|P_A U_A X_A U_A^\dagger P_A^\dagger\|_1 = \frac{\mathrm{tr}[P_A^\dagger P_A]}{|A|} \|X_A\|_1.$$

(c) Verify Eq. (13.10). Note that the random unitary acts on the typical subspace.

Lecture 14

Quantum key distribution

We will now shift gears to one of the most important applications of quantum information theory: quantum key distribution. This is a powerful application of quantum information theory to cryptography. The most basic scenario of cryptography deals with two parties want to communicate some message, which they wish to keep secret from any other parties. However, they can only use means of communications where there is the risk that someone intercepts their message without them knowing it (perhaps they send a letter, and someone could open the letter halfway towards its destination). If the message is *encrypted* in such a way that only the sender and the legitimate receiver are able to understand the content this does not pose a problem. Throughout history, a wide variety of cryptosystems has been invented which are such that the encoded message (at least at first glance) looks unintelligible. For instance, the sender and receiver could have some secret dictionary and have a code word for each real word. A problem with such an approach is that if there is a sufficiently large amount of encoded data available to the eavesdropper, she will at some point be able to see patterns in the encoded messages and start to learn the actual encoding. Indeed, one of the birthplaces of information theory and computing theory has been the Polish and British effort to break German cryptography during World War II. Large-scale secure cryptography is a cornerstone of modern digital technology: we want to be able to send private and sensitive information over digital communication channels. We will give a short (and superficial) introduction to some relevant concepts from cryptography, and then describe how quantum information theory can help!

14.1 Cryptography

A central concept in cryptography is the notion of an *adversary*, who is trying to discover the secret message. We will call her Eve, for eavesdropper. As usual, we furthermore have Alice and Bob, who are honest parties and want to communicate a secret message. We will assume that Eve can do *whatever she wants* with the communication she intercepts. For instance, if there is some noise on the communication channel, we must assume that Eve controls the noise and all information leaking into the environment (so Eve is both *eavesdropper* and *environment*).[1] Eve also is assumed to know precisely the details of the *protocol* Alice and Bob implement. This is an important principle in cryptography: one could do cryptography by keeping the encoding and decoding procedures secret (as opposed to using for example open source software), but this means that the secrecy could be compromised when information about your protocols is leaked. So, for security of a protocol, we should know that the eavesdropper knows the protocol (e.g. she

[1] This is a fundamentally different perspective than the one usual in physics, where in Einstein's words one might believe that "Raffiniert ist der Herr Gott, aber boshaft ist er nicht".

has access to the same open source software). What can Alice and Bob achieve in this scenario?

The one-time pad

Fortunately, there is a simple and perfectly secure way for Alice and Bob to communicate a secret message, the so-called *one-time pad*. Suppose that Alice wants to send a message to Bob, which she has encoded in a bitstring m of length m. Furthermore, suppose that Alice and Bob share a *key*, which is a string k of m uniformly random bits, and which is such that the key k is not known to Eve. Then Alice can simply send over the encrypted message $c = m \oplus k$. Here $\oplus$ denotes bitwise addition modulo 2 (also known as the *parity*). Bob can simply decode by $m = c \oplus k$.

Example 14.1. Suppose that Alice has message $m = (010010)$ and key $k = (101011)$. Then the encrypted message c is given by:

m	0	1	0	0	1	0
k	1	0	1	0	1	1
$c = m \oplus k$	1	1	1	0	0	1
$c \oplus k$	0	1	0	0	1	0

so Bob indeed decodes the message m correctly.

This is perfectly secure: if the key is chosen uniformly random, then it is easy to check that for any message m, the encrypted message c is uniformly random. Indeed, let c be an arbitrary bitstring, then $k = m \oplus c$ is the unique key such that if we encode with k we get c, $m \oplus k = c$. Since the key is uniformly random, the code word is also uniformly random. This implies that if the key is unknown to Eve

$$I(M : C) = 0,$$

so she obtains no information about m from c. Note that Eve is assumed to *know* precisely the encoding and decoding protocol Alice and Bob use, but she does not know the value of the key. One issue is that the *length* of the key has to be equal to the information content of the source (as you may show in Exercise 14.1). So, it is for instance not secure to reuse the same key multiple times (this is why it is called a one-time pad), as in that case Eve can start to extract information from the code word about the key (and hence the message).

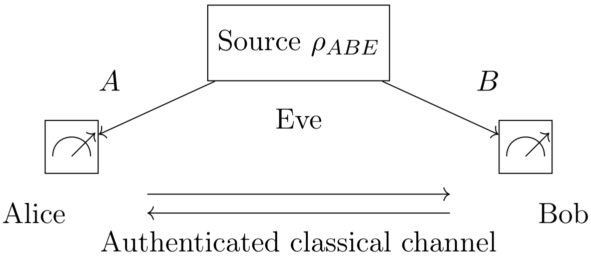
Key generation protocols

The above discussion shows that the task of secure communication reduces to the task of establishing a good key. A significant challenge is that the amount of key we have to establish equals the length of the actual message we want to send. It may seem like we did not make much progress: if we can not communicate securely, how do we generate the key? Of course, one option would be to somehow pre-establish the key (say, Alice physically goes over to Bob and leaves him some large amount of key). This is possible but often not practical. What we would ideally like is *public key distribution*, where Alice and Bob are able to use a public channel to communicate (so Eve can see what they send) and they are nevertheless able to establish a secret key. Fortunately and perhaps surprisingly, there are known methods for doing so. The known *classical* methods for public key distribution are based on *computational hardness* of certain problems. In this case, the key generation is only secure if we assume that Eve is restricted to have only limited computational power. In practice this is often reasonable (i.e. we have computational problems which can not be solved by a supercomputer with the best known algorithms within a reasonable time). Nevertheless, it can be problematic! For instance, suppose that one has a secret that needs to stay secret for multiple decades. It is quite possible that in the meantime there are significant advances in either algorithms or hardware, allowing Eve to break the secret. A dramatic example is given by quantum computation. Many of the methods currently used for public key distribution are based on the hardness of number-theoretic problems, and in particular on the hardness of factoring large numbers which are the product of two primes. Quantum computation would be able to efficiently solve these problems using Shor's algorithm, breaking a substantial part of current cryptography. There do exist candidates for computational problems which can be used for public key distribution for which we do not know efficient quantum algorithms. This so-called post-quantum cryptography will not concern us here. We will discuss another direction, where quantum information theory offers us a fantastic possibility: generating key which is *information-theoretically secure*! This is known to be impossible using only classical communication.

There are two frameworks for quantum key distribution. The first is consists of *entanglement-based* protocols. Here, the set-up is as follows:

- Alice and Bob receive a state ρ_{AB} from some *untrusted source.*
- Alice and Bob can communicate through a classical channel, which is a *public authenticated channel*[2]. This means that anyone can see the information transmitted over the channel (it is public) but it can not be altered by Eve (authenticated).

Alice and Bob then perform some appropriate measurements, and distill a key from the measurement results.

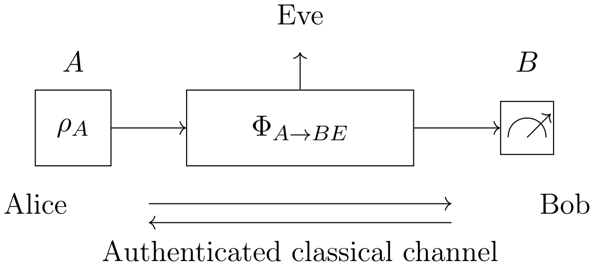


Alternatively, Alice and Bob can use a *prepare-and-measure* protocol, in which

- Alice and Bob can communicate using a quantum communication channel, allowing them to send qubits over to each other. This channel is untrusted, in the sense that we assume that Eve has access to the channel and can do with it whatever she wants.
- Alice and Bob can communicate over a classical channel, which as before is public and authenticated (so Eve knows what is being communicated but does not change the messages).

The idea behind the protocol is that Alice sends over qubits in different bases, and only announces the choice of basis after Bob has received and measured the qubits.

[2] There are cryptographic means to make an untrusted channel (which can possibly be manipulated by Eve) authenticated. This requires a relatively small amount of key. Quantum key distribution therefore requires a small initial amount of key.



The problem for Eve is that she does not know in which basis she has to measure in order to discover the information Alice sends over. If Eve eavedrops and she chooses to measure the communicated qubits and measures in the wrong basis, she will disturb the state. This will allow Alice and Bob to detect her presence as soon as she gains significant information about the key, in which case they abort the protocol. One of the quantum principle that ensures the possibility of quantum key distribution is the no cloning theorem (Theorem 5.4): Eve is not able to intercept the communicated qubits and make a copy of them and store them. Quantum key distribution achieves secure public key distribution (which is classically not possible) and this is one of the most important practical applications of quantum information theory. Additionally, the prepare-and-measure protocols are not too complicated and only require the ability to send single qubits from Alice to Bob, for instance encoded in photons, making quantum key distribution practically feasible with current technology.

Requirements for quantum key distribution

Let us now make formal what information-theoretic security means for a quantum key distribution protocol. We will make three natural requirements, stating that the protocol gives a correct key, that the key is secret, and that in the absence of Eve the protocol is not aborted. In these three requirements we allow a small probability of error.

Correctness: If we denote by k_A and k_B the key generated by Alice and Bob respectively, then we should have $k_A = k_B$. We allow the possibility that the protocol is incorrect with small probability. We say that a key generation protocol is $\varepsilon_{\mathrm{corr}}$-correct if

$$\Pr(k_A \neq k_B) \leq \varepsilon_{\mathrm{corr}}.$$

Secrecy: The second demand is that the key is secret (so Eve does not know what it is). Again, we allow a small error (so Eve could potentially learn a very small amount of information about the key). If Eve knows nothing about the key, then the state at the end of the protocol must be a product state between her and Alice and Bob. Moreover, the key should be a (nearly) uniformly random choice from the set of possible keys (which is for instance bitstrings of length m). That is, the key, say on the side of Alice, should be a maximally mixed state

$$\tau_A = \sum_{k\in\mathcal{K}} \frac{1}{|\mathcal{K}|} |k_A\rangle\langle k_A|.$$

We say that the protocol is $\varepsilon_{\mathrm{sec}}$-secret if

$$(1-p^{\perp})T(\rho_{AE}, \tau_A \otimes \rho_E) \leq \varepsilon_{\mathrm{sec}}$$

where ρ_{AE} is conditioned on not aborting the protocol and $p^\perp$ is the probability of aborting the protocol. The reason that the factor $1 - p^\perp$ is present is that in the situation where the protocol aborts with probability very close to 1, it may be possible for Eve to learn the key. It would be too restrictive to demand that if one conditions on the unlikely event that the protocol does not abort, Eve still does not learn anything about the key.

Robustness: There is a final requirement, which is essentially that the protocol should work well in the ideal case where Eve (or any noise) is absent. So, we say that the protocol, in the absence of noise or Eve should be such that the probability of aborting $p^\perp$ is small: the protocol is δ-robust, if $p^\perp \leq \delta$ in the absence of Eve (or noise).

A quantum key distribution protocol is secure if we have a family of protocols of increasing length n, such that for sufficiently large n, the values $\varepsilon_{\mathrm{corr}}$, ε_{secr} and δ can be made arbitrarily small.

14.2 Entanglement-based protocol

In an entanglement-based protocol, Alice and Bob receive a state from an untrusted source. The source of the state can be realized in different ways. For instance, it could be that Alice simply prepares a state locally, and sends part of it over to Bob over some quantum channel. Note that in this situation we can not assume that the state Bob receives really is the state as prepared by Alice, as it can have been manipulated by Eve.

To get an idea of how we may achieve a secret key, we will first describe two protocols which are incorrect, but which illustrate the basic ideas underlying quantum key distribution.

Shared entanglement can generate key

First of all, supposed that Alice and Bob share a maximally entangled qubit state $\rho_{AB} = |\Phi^+_{AB}\rangle$. Then, since the state is pure, they know that they must be uncorrelated with Eve, i.e. the total state ρ_{ABE} must be given by

$$\rho_{ABE} = |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}| \otimes \rho_E.$$

Now, if they simply measure in the standard basis they get a correlated random bit which is uncorrelated with Eve, i.e. they precisely get one bit of key. So, if Alice and Bob share a maximally entangled state, they can generate a perfectly correct and secret key! However, the source is of course assumed to *untrusted.* If the protocol is just measuring in the standard basis, but the state was not a maximally entangled state Alice and Bob could be deceived by Eve, so this simple approach does not give security.

Next, let as assume that Alice and Bob receive a state of n qubits, and the state is IID, so they receive a state $\rho_{AB}^{\otimes n}$. Again, this is not a good assumption since it strongly limits the possibility of what Eve may do. If we nevertheless assume this, Alice and Bob can simply perform the following protocol:

(a) Alice and Bob use LOCC on the first $N < n$ qubits in order to accurately determine which state ρ_{AB} they have n copies of.

(b) They then apply an entanglement distillation protocol (as we have seen before), extracting maximally entangled states at rate $-H(A|B)$.

(c) Finally, they measure to obtain a secret key.

Note that Alice and Bob use the authenticated channel to send classical information to learn the state and to perform entanglement distillation. This classical communication is not secret, but Eve can also not mislead them by changing the information they send. Under the restrictive assumption that the starting state is of the form $\rho_{AB}^{\otimes n}$, this gives a secure protocol!

Verifying entanglement

This is where the second idea comes in. We need a way for Alice and Bob to *verify* that the shared state is maximally entangled, or close to it. We can use the ideas of Lecture 3 for this! Suppose that Alice and Bob receive a state on n qubits each. In the ideal version, this state would be maximally entangled. What Alice and Bob do, is that they randomly pick a N subsystems and play the CHSH game on these copies of the state, using the public authenticated channel. If they win with probability close to $\frac{1}{2}(1+\frac{1}{\sqrt{2}})$ the state on these N qubits must have been the maximally entangled state (or close to it). Since they chose the N qubits randomly, this implies that the whole state must have been close to a maximally entangled state. They can use the other $n-N$ qubits to generate a secret key. Note that the measurement outcomes on the N qubits used in the CHSH game can not be used as key, since the measurement outcomes have been communicated over a public channel. If Alice and Bob do not win the CHSH game with high enough probability, they abort the protocol. To make such a protocol concrete, consider the following qubit basis, depending on an angle θ

$$|0^{(\theta)}\rangle = \cos(\theta)|0\rangle + \sin(\theta)|1\rangle \qquad |1^{(\theta)}\rangle = -\sin(\theta)|0\rangle + \cos(\theta)|1\rangle. \tag{14.1}$$

These are such that in the CHSH game, as discussed in Lecture 3, Alice measures either using $\theta = 0$ (the standard, or Z-basis $|0\rangle, |1\rangle$) or $\theta = \frac{\pi}{4}$ (the X-basis $|+\rangle, |-\rangle$) and Bob measures using $\theta = \frac{\pi}{8}$ or $\theta = -\frac{\pi}{8}$. Now, the protocol will be that for each qubit Alice randomly takes θ to be one of $0, \frac{\pi}{8}, \frac{\pi}{4}$ and measures in the corresponding basis, i.e. she chooses an angle $\theta_{A,i}$ for $i = 1, \dots, n$. Similarly, Bob randomly chooses an angle $\theta_{B,i}$ from $\frac{\pi}{8}, 0, \frac{\pi}{8}$ for the i-th qubit and measures in that basis. Alice and Bob then publicly communicate their choices of measurement basis $\theta_{A,i}$, $\theta_{B,i}$ for $i = 1, \dots, n$ (but not the outcomes). They then *sift* the qubits into three sets

$$\{1, \dots, n\} = I_1 \cup I_2 \cup I_3$$

The first set I_1 consists of qubits for which they chose the same basis, so either $\theta_{A,i} = \theta_{B,i} = 0$ or $\theta_{A,i} = \theta_{B,i} = \frac{\pi}{8}$ for $i \in I_1$. The second set I_2 consists of those qubits for which they chose a measurement basis corresponding to the CHSH game, i.e. $\theta_{A,i} \in \{0, \frac{\pi}{4}\}$ and $\theta_{B,i} \in \{-\frac{\pi}{8}, \frac{\pi}{8}\}$ for $i \in I_3$. Finally, I_3 is what is left over, and is discarded. Alice and Bob then publicly communicate the outcome of the measurements on I_2 and use these to 'play' the CHSH game. Note that in this setting there is no external referee, but Alice and Bob treat their own random choice of measurement bases as the questions of the game. If the fraction of qubits for which Alice and Bob win is close to the optimal quantum winning probability $\omega^*(\mathrm{CHSH}) = \frac{1}{2}(1+\frac{1}{\sqrt{2}})$, Alice and Bob know that their state on I_2 was closely to maximally entangled. However, we really want to know something about the qubits in I_1, which we are going to use to generate the key! The reason that we do in fact also learn something about I_1 is because the sets I_1 and I_2 are random. When distributing the source, Eve does not know which qubits are going to be the 'check' qubits in I_2 and which ones are going to be the 'key' qubits of E_1. In order to pass the test, the states in E_1 have to be close to maximally entangled states, but because of the random choice of subsystem this can only be achieved if *all* qubits are sufficiently close to a maximally entangled state.

Classical post-processing

We conclude from the previous that if Alice and Bob win the CHSH game with probability close to optimal, they will a state which is close to maximally entangled. The would like to use the measurement outcomes from I_1 as the secret key. Let $l = |I_1|$ and denote by $x^l = (x_1, \ldots, x_l)$ and $y^l = (y_1, \ldots, y_l)$ the bitstrings of measurement outcomes of Alice and Bob. Alice and Bob will get *mostly* the same outcomes when they measure in the same basis on I_1, so $x_i = y_i$. However, the state they share may not be exactly maximally entangled, so there can be some errors and there will be i such that $x_i \neq y_i$. In order to address these errors, Alice and Bob have to perform a procedure called *reconciliation*. Such a procedure is necessary for the *correctness* of the protocol. The idea is that Alice and Bob use error correction to reduce the errors, at the cost of making the key shorter. For example, Alice could take the first two bits and send over their parity $x_1 \oplus x_2$. If $x_1 \oplus x_2 \neq y_1 \oplus y_2$ Alice and Bob discard these bits. If they are the same, then they may discard x_2 and y_2, and use x_1 and y_2 as key bits. Note that from the parity $x_1 \oplus x_2$ alone Eve can not learn the value of x_1. Also, we now have $x_1 = y_1$ *unless* we had both $x_1 \neq y_1$ and $x_2 \neq y_2$, so this reduces the probability of error. The process described here is rather inefficient, as it reduces the key length by at least a factor of two. There are more complicated error correcting codes, which lead to more efficient reconciliation, reducing the errors at lower overhead.

Finally, as the CHSH game on I_2 only gave us a guarantee that the state was close to maximally entangled, there can be some (relatively small) amount of correlation with Eve. In other words, the key is not yet perfectly secure. Fortunately, there is a procedure called *privacy amplification* which takes as input a partially secret key and produces a shorter but more secure key. This is required for the *secrecy* of the protocol. This is a very important aspect of quantum key distribution, for now take it for given.

The E91 protocol

The whole discussion gives the so-called E91 protocol, named after its inventor Ekert and year of invention, 1991.

(a) Alice and Bob receive a state $\rho_{A^n B^n}$ on n pairs of qubits.

(b) Alice measures each of her qubits. She does so by choosing, for each qubits, randomly from one of the bases in Eq. (14.1) with $\theta = 0, \frac{\pi}{8}, \frac{\pi}{4}$, which gives a bitstring $x^n = (x_1, \ldots, x_n)$

(c) Similarly, Bob measures each of his qubits, using random bases with $\theta = -\frac{\pi}{8}, 0, \frac{\pi}{8}$ giving a bitstring of outcomes $y^n = (y_1, \ldots, y_n)$

(d) Alice and Bob communicate, publicly, their choice of basis (but not the measurement outcomes) and sift the qubits in the three sets I_1, I_2 and I_3.

(e) They communicate publicly the measurement outcomes on the set of qubits in I_2. On this set of outcomes they verify that they win the CHSH game with winning probability close to the optimal value. If this is not the case they abort the protocol.

(f) They now consider the set I_1 of outcomes on which they measured in the same basis, and they use an error correcting code to perform reconciliation on $(x_i, y_i)_{i \in I_1}$. With high probability this leaves Alice and Bob with an equal bitstring.

(g) Alice and Bob now share a key, and they use privacy amplification to distill a secure key k.

While the whole procedure is called quantum key 'distribution' it would be perhaps be more accurate to call it quantum key *generation*, as it is not the case that Alice can determine what the key is at the beginning of the protocol and then transmit it to Bob, but rather they end up with some random key at the end of the protocol.

Example 14.2. Here we give an example execution of the E91 protocol.

Qubit: i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Angle Alice: $\theta_{A,i}$	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{\pi}{8}$	0	0	$\frac{\pi}{4}$	0	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{\pi}{8}$	0	$\frac{\pi}{8}$	0	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{\pi}{4}$
Angle Bob: $\theta_{B,i}$	0	0	$\frac{\pi}{8}$	$\frac{\pi}{8}$	$-\frac{\pi}{8}$	$\frac{\pi}{8}$	0	0	$-\frac{\pi}{8}$	$\frac{\pi}{8}$	0	0	$\frac{\pi}{8}$	$-\frac{\pi}{8}$	0	$-\frac{\pi}{8}$
Outcomes Alice: x_i	0	0	**1**	0	1	0	**1**	0	0	**0**	**1**	0	1	1	0	1
Outcomes Bob: y_i	1	0	**1**	0	0	0	**1**	1	1	**1**	**1**	1	1	0	0	0

Alice and Bob communicate their angles publicly. The outcomes in I_1, where they measure in the same basis, are made bold in the above table and the outcomes in I_3 which are discarded are grey. Let us now focus on the check qubits I_2. Recall that Alice and Bob win the CHSH game either if they have the angles $\frac{\pi}{4}$ and $-\frac{\pi}{8}$ and get a different bit, so $x_i \oplus y_i = 1$, or otherwise they should get the same bit so $x_i \oplus y_i = 0$. We get the following:

Angle Alice: $\theta_{A,i}$	0	0	$\frac{\pi}{4}$	$\frac{\pi}{4}$	0	$\frac{\pi}{4}$
Angle Bob: $\theta_{B,i}$	$\frac{\pi}{8}$	$-\frac{\pi}{8}$	$\frac{\pi}{8}$	$-\frac{\pi}{8}$	$\frac{\pi}{8}$	$-\frac{\pi}{8}$
Outcomes Alice: x_i	0	1	0	0	1	1
Outcomes Bob: y_i	0	0	0	1	1	0
Win CHSH?	✓	✗	✓	✓	✓	✓

While the size is really too small to give a good estimate of the winning probability, we will pretend for now that this is indeed sufficient evidence that they win the CHSH game with high enough probability. Then, on the qubits in I_1, Alice and Bob perform one step of error correction:

Outcomes Alice: x_i	1	1	0	1
Outcomes Bob: y_i	1	1	1	1
Parities Alice: $x_{2i-1} \oplus x_{2i}$	0		1	
Parities Bob: $y_{2i-1} \oplus y_{2i}$	0		0	
Parity check?	✓		✗	
Key k:	1			

Here, we used the error correction procedure where Alice and Bob check for pairs of bits whether they have same parity. If they do not, they discard the pair. If they have the same parity they take the first bit of the pair as the key bit. All together, we have established one bit of key! Of course, this should really be done with a (much) larger number of qubits, and should be followed by a final privacy amplification step to ensure security.

In general, if we use this protocol, starting with n qubits, the expected number of qubits in I_1 is $\frac{2}{9}$, so the check qubits (for playing the CHSH game) are only a small overhead. The final remaining number of bits of key depends on the estimate how close the state was to maximally entangled (the larger the error, the more expensive the reconciliation and privacy amplification are).

14.3 Prepare-and-measure protocol

We will now describe the famous *BB84 protocol*, named after its inventors Bennett and Brassard, and the year of its invention, 1984. We now assume that Alice and Bob have access to an (unreliable) quantum communication channel, as well as a public authenticated classical channel. There are various protocols possible in this setting, the BB84 protocol proceeds as follows:

(a) Alice generates two uniformly random bitstrings $a^n = (a_1, \ldots, a_n)$ and $x^n = (x_1, \ldots, x_n)$ and for $i = 1, \ldots, n$, Alice sends over

$$H^{a_i}|x_i\rangle \quad \text{where} \qquad H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

That is, a_i determines whether she sends a bit in the $|0\rangle, |1\rangle$ basis or in the $|+\rangle, |-\rangle$ basis and x_i determines which basis state she sends.

(b) Bob also generates a uniformly random bit string $b^n = (b_1, \ldots, b_n)$. If $b_i = 0$ he measures in the standard basis, if $b_i = 1$ Bob measures in the $|\pm\rangle$ basis, giving a bitstring of outcomes $y^n = (y_1, \ldots, y_n)$.

(c) Alice and Bob now publicly communicate their choices of basis a^n and b^n. They discard the outcomes for which $a_i \neq b_i$ (the ones in which they measured in a different basis). This leaves them with the *sifted* bitstrings $x^{n'}$ and $y^{n'}$.

(d) Alice and Bob now randomly divide the sifted indices into I_1 and I_2, and they communicate the outcomes x_i and y_i for $i \in I_2$ publicly, in order to estimate the error rate. If the error rate is too high they abort the protocol.

(e) Alice and Bob use an error correcting code to perform reconciliation on $(x_i, y_i)_{i \in I_1}$ to make sure the key is correct.

(f) Alice and Bob use privacy amplification to improve the security of the resulting key k.

At what rate does the protocol establish key? The sifting approximately halves the length of their bitstring. Alice and Bob will also use half of the sifted bitstrings to estimate the error rate. So, after (d) Alice and Bob are left with approximately bitstrings of length $\frac{n}{4}$. The length of the final key depends on the amount of error: if the error is very small they only have to do little reconciliation and privacy amplification.

Example 14.3. Let us see an example execution of the BB84 protocol.

Qubit: i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Basis choice Alice: a_i	1	1	1	0	0	1	0	1	0	1	0	1	1	0	1	1
State choice Alice: x_i	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	0
Qubit sent over	$\|+\rangle$	$\|-\rangle$	$\|+\rangle$	$\|0\rangle$	$\|0\rangle$	$\|+\rangle$	$\|1\rangle$	$\|0\rangle$	$\|+\rangle$	$\|+\rangle$	$\|1\rangle$	$\|+\rangle$	$\|-\rangle$	$\|0\rangle$	$\|-\rangle$	$\|+\rangle$
Basis choice Bob: b_i	0	1	1	1	1	1	0	0	1	1	1	0	1	0	1	0
Outcome Bob: y_i	0	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1
Same basis: $a_i = b_i$?	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	✗

Of the remaining qubits Alice and Bob random choose half of the remaining indices I_1 to represent the key generating bits with which we have made bold. The remaining bits in I_2 are check bits.

Bits Alice: x_i	**1**	0	0	**1**	**0**	1	**0**	1
Bits Bob: y_i	**1**	1	0	**1**	**0**	1	**1**	1
Same outcome check bits?		✗	✓			✓		✓

Alice and Bob find that there is a single error in the check bits (that is of course already quite a lot on four bits, but let us assume for now that this is good enough to proceed). They are left with bits in I_1 and they perform a round of reconciliation to correct remaining errors.

Outcomes Alice: x_i	1	1	0	0
Outcomes Bob: y_i	1	1	0	1
Parities Alice: $x_{2i-1} \oplus x_{2i}$	0		1	
Parities Bob: $y_{2i-1} \oplus y_{2i}$	0		0	
Parity check?	✓		✗	
Key k:	1			

So, they are again left with a single bit of key. As in Example 14.2 Alice and Bob should really perform the protocol with a much larger number of qubits, and they will perform a privacy amplification step at the end.

The intuition behind the BB84 protocol is that *information gain implies disturbance*. Indeed, suppose that we take the perspective of Eve. One approach that she could take to eavesdrop is to try to distinguish the states $|0\rangle$, $|1\rangle$, $|+\rangle$ and $|-\rangle$.

Lemma 14.4 (Information gain implies disturbance). *Suppose $\rho_A = |\phi_A\rangle\langle\phi_A|$ and $\sigma_A = |\psi_A\rangle\langle\psi_A|$ are non-orthogonal pure states on a Hilbert space $\mathcal{H}_A$. Consider a channel $\Phi_{A\to AX}$ which is such that*

$$\mathrm{tr}_X[\Phi_{A\to AX}(\rho_A)] = \rho_A \quad \text{and} \quad \mathrm{tr}_X[\Phi_{A\to AX}(\sigma_A]) = \sigma_A$$

then

$$\mathrm{tr}_A[\Phi_{A\to AX}(\rho_A)] = \mathrm{tr}_A[\Phi_{A\to AX}(\sigma_A)].$$

Proof. Consider a Stinespring isometry $V \in \mathrm{Isom}(A, AXE)$ of $\Phi_{A\to AX}$. Then let $|\tilde{\phi}_{AXE}\rangle =$

$V|\phi_A\rangle$ and $|\tilde{\psi}_{AXE}\rangle = V|\psi_A\rangle$. By assumption, the reduced states on A are pure so these must be product states

$$|\tilde{\phi}_{AXE}\rangle = |\phi_A\rangle|u_{XE}\rangle \text{ and } |\tilde{\psi}_{AXE}\rangle = |\psi_A\rangle|v_{XE}\rangle.$$

Also, since V is an isometry

$$\langle\phi_A|\psi_A\rangle = \langle\tilde{\phi}_{AXE}|\tilde{\psi}_{AXE}\rangle = \langle\phi_A|\psi_A\rangle\langle u_{XE}|v_{XE}\rangle$$

which implies $\langle u_{XE}|v_{XE}\rangle = 1$ so $|u_{XE}\rangle = |v_{XE}\rangle$. □

This shows that whenever Eve tries to distinguish non-orthogonal states she will disturb the system. The above lemma can be made more quantitative, see for instance Exercise 14.4. In general, the more Eve learns, the more she disturbs the state and Alice and Bob will be able to detect this disturbance if Eve learns too much. For actually *proving* that the BB84 scheme is secure this does not yet suffice, as Eve is not restricted to attempts to distinguish the individual qubits: she can perform *any* quantum channel on the communicated qubits!

The BB84 protocol, and other prepare-and-measure approaches have some practical advantages over entanglement-based protocols: it is often easier to send over qubits one-by-one rather than establishing a joint entangled state. However, for proving security it is useful to reduce a prepare-and-measure protocol to an entanglement-based protocol, which can always be done. The BB84 protocol for example can be turned into an entanglement-based protocol in the following way:

(a) Alice prepares n maximally entangled qubit states $|\Phi^+_{AB}\rangle^{\otimes n}$ locally.

(b) She chooses a random bitstring $a^n = (a_1, \ldots, a_n)$ and applies $\bigotimes_{i=1}^n H^{a_i}$ on the A^n system.

(c) Alice sends over the B-systems to Bob, who signals he has received the quantum state.

(d) Alice divides into two sets I_1 and I_2, and she publicly communicates a_i to Bob for the check qubits $i \in I_2$.

(e) Bob measures the qubits $i \in I_2$. He measures in the $|0\rangle, |1\rangle$-basis if $a_i = 0$ and in the $|+\rangle, |-\rangle$ basis if $a_i = 1$.

(f) Alice and Bob check how much error there is on the I_2 measurement outcomes, if it is too high they abort.

(g) Alice and Bob perform entanglement distillation, so they obtain a (near) maximally entangled state.

(h) Alice and Bob measure this state in the standard basis.

For the entanglement distillation step, they could in theory use the decoupling approach for entanglement distillation. In practice there are more convenient *quantum error correcting codes* which can be used for this step.

14.4 Exercises

14.1 **Key rate:** The goal of this exercise is that for information-theoretic security, the one-time pad is optimal, and the length of the key is at least the amount the information of the message. Suppose that we have a source M, a key K and an encoding C which depends on K and M.

(a) Argue that in order for the message to be perfectly securely encoded we want $I(M : C) = 0$.

(b) Argue that in order to be able to exactly recover the message M from the encoding C if you know the key, we need $H(M|CK) = 0$.

(c) Show that if we want C to be securely encoded, and we want to be able to recover C from K and M, we need $H(K) \geq H(M)$.

(d) Argue that this is even the case if the code word C is a quantum state (but the key K and the message M are still classical).

14.2 **Security definition:** Suppose that Alice and Bob perform a quantum key distribution protocol, with Eve eavesdropping. Let ρ_{ABE}^{pass} be the final state, assuming the protocol has not been aborted. That is,

$$\rho_{ABE}^{\text{pass}} = \sum_{k_A,k_B \in \mathcal{K}} p_{AB}(k_A, k_B)|k_A k_B\rangle\langle k_A k_B| \otimes \rho_E^{k_A,k_B} \ ,$$

where $p_{AB}(k_A, k_B)$ is the probability that Alice and Bob generate keys k_A and k_B respectively, and $\rho_E^{k_A,k_B}$ is Eve's final state in this case. We say that the key distribution protocol is *ϵ-secure* if

$$(1 - p^{\perp})T\Big(\rho_{ABE}^{\text{pass}}, \omega_{AB} \otimes \rho_E^{\text{pass}}\Big) \leq \epsilon \ ,$$

where $\omega_{AB} = \frac{1}{|\mathcal{K}|}\sum_{K \in \mathcal{K}}|KK\rangle\langle KK|$.

(a) Define the state σ_{ABE} similarly to ρ_{ABE}^{pass}, except afterwards Bob has thrown out his state and copied Alice's key. That is,

$$\sigma_{ABE} = \sum_{k_A,k_B \in \mathcal{K}} p_{AB}(k_A, k_B)|k_A k_A\rangle\langle k_A k_A| \otimes \rho_E^{k_A,k_B} \ .$$

Show that

$$T(\rho_{ABE}^{\text{pass}}, \sigma_{ABE}) \leq \frac{1}{1 - p^{\perp}} \Pr(k_A \neq k_B) \ .$$

(b) Next, show that

$$T(\sigma_{ABE}, \omega_{AB} \otimes \rho_E) = T(\rho_{AE}^{\text{pass}}, \omega_A \otimes \sigma_E) \ .$$

(c) Deduce that if the protocol is ϵ_{cor}-correct and ϵ_{sec}-secret then it is $(\epsilon_{\text{cor}} + \epsilon_{\text{sec}})$-secure.

14.3 **Quantum one-time pad:** Let $(k_1, k_2) \in \{0, 1\}^2$ be a key. The *quantum one-time pad* encodes a single qubit A as

$$\rho_A \mapsto X^{k_1} Z^{k_2} \rho_A Z^{k_2} X^{k_1}.$$

(a) Explain why this allows perfect decoding if one knows the key.

(b) Show that if one does not know the key and the key is uniformly random, the encoded state is the maximally mixed state.

(c) Show that one needs at least two bits of key to encode a qubit. *Hint: use Exercise 14.1 and superdense coding.*

14.4 **Quantum money:** What does a bank do? They issue money, say in the form a piece of paper, a banknote, and when someone comes to the bank with a legitimate banknote, the bank confirms that this represents a certain value. The bank needs to make sure that Eve does not *forge* any money! The bank could give serial numbers to all the notes they print. However, Eve can just copy a real serial number and use this to print fake money... Fortunately quantum mechanics offers a solution!

(a) Consider the following protocol: the bank takes two random bitstrings $a^n = (a_1, \dots, a_n)$ and $x^n = (x_1, \dots, x_n)$ and prepares n qubits, where the i-th qubit is in state $H^{a_i}|x_i\rangle$. The bank hands these qubits, together with a unique serial number, to the customer. This is a 'quantum banknote'. If the customer returns to the bank, how can the bank verify that the customer has a valid banknote?

(b) Explain (on an intuitive level) how no-cloning prevents Eve from forging a banknote.

We will show that a natural type of attack by Alice has small chance of succeeding. The goal of the attack will be for Eve to create from one banknote a second banknote such that both notes are accepted by the bank. The forging attempt will be the following. She tries to perform a quantum channel *on each qubit separately.* Ideally, this channel would clone the state, but we know this is impossible. In other words, the ideal channel Eve wants to construct is a cloning channel $\Phi_{A\to AE}$ where A, E are qubit systems and

$$\Phi_{A\to AE}(|\phi_A\rangle\langle\phi_A|) = |\phi_A\rangle\langle\phi_A| \otimes |\phi_E\rangle\langle\phi_E|$$

for $|\phi\rangle = |0\rangle, |1\rangle, |+\rangle, |-\rangle$.

(c) Consider the following test for whether Eve has succesfully fooled the bank: the bank measures whether the final state is $|\psi_A\rangle|\psi_E\rangle$ or not. Argue that the probability p_{pass} of passing thist test, given that the qubit is $|x\rangle$ for a random $x \in \{0, 1, +, 1\}$ is given by

$$p_{\text{pass}} = \sum_{x\in\{0,1,+,1\}} \frac{1}{4} \langle x_A|\langle x_A|\Phi(|x_A\rangle\langle x_A|)|x_A\rangle|x_A\rangle.$$

Show that if $J(\Phi)$ is the Choi matrix of $\Phi_{A\to AE}$ then

$$p_{\text{pass}} = \frac{1}{4} \sum_{x\in\{0,1,+,1\}} \operatorname{tr}\left[|x\rangle\langle x|^{\otimes 3} J(\Phi)\right]$$

(d) Let $Q = \frac{1}{4}\sum_{x\in\{0,1,+,1\}}|x\rangle\langle x|^{\otimes 3}$. Show that $\|Q\|_\infty = \frac{3}{8}$.

(e) Show that the probability p_{pass} of Eve passing the test given a state randomly chosen from $|0\rangle, |1\rangle, |+\rangle, |-\rangle$ is at most $\frac{3}{4}$. *Hint: use Eq.* (6.3). *What is* $\|J(\Phi)\|_1$*?*

(f) Conclude that the probability that Eve (using a strategy of this type on every qubit) manages to produce a quantum banknote of length n which passes the test at the bank is at most $(\frac{3}{4})^n$.

Remark: This scheme historically predates quantum key distribution and was proposed by Wiesner. It already contains the main conceptual idea of quantum key distribution and was one of the sources of inspiration for the BB84 protocol. What makes it (currently) practically infeasible is that it requires the certificate to remain stably in the correct quantum state over a long time. Quantum key distribution has the practical advantage that Bob can immediately measure the quantum systems and never has to store quantum information for a long time.

14.5 **Security from entanglement:** Show that if Alice and Bob share a state ρ_{AB} which is close to the maximally entangled state, so

$$P\left(\rho_{AB}, |\Phi^+_{AB}\rangle\langle\Phi^+_{AB}|\right) \leq \varepsilon$$

then measuring in the standard basis gives a key which is ε-secure and ε-correct.

Bibliography

[1] Scott Aaronson, *The zen anti-interpretation of quantum mechanics*, 2022. https://scottaaronson.blog/?p=5359.

[2] Alain Aspect, Jean Dalibard, and Gérard Roger, *Experimental test of Bell's inequalities using time-varying analyzers*, Physical review letters **49** (1982), no. 25, 1804.

[3] Ingemar Bengtsson and Karol Życzkowski, *Geometry of quantum states: an introduction to quantum entanglement*, Cambridge university press, 2017.

[4] David Bohm, *A suggested interpretation of the quantum theory in terms of "hidden" variables. I*, Physical review **85** (1952), no. 2, 166.

[5] Heinz-Peter Breuer and Francesco Petruccione, *The theory of open quantum systems*, Oxford University Press, USA, 2002.

[6] Eric Chitambar, Debbie Leung, Laura Mančinska, Maris Ozols, and Andreas Winter, *Everything you always wanted to know about locc (but were afraid to ask)*, Communications in Mathematical Physics **328** (2014), 303–326.

[7] Matthias Christandl, Vladimir Lysikov, Vincent Steffan, Albert H Werner, and Freek Witteveen, *The resource theory of tensor networks*, arXiv preprint arXiv:2307.07394 (2023).

[8] Matthias Christandl, Alexander Müller-Hermes, and Michael M Wolf, *When do composed maps become entanglement breaking?*, Annales Henri Poincaré **20** (2019), 2295–2322.

[9] J Ignacio Cirac, David Perez-Garcia, Norbert Schuch, and Frank Verstraete, *Matrix product states and projected entangled pair states: Concepts, symmetries, theorems*, Reviews of Modern Physics **93** (2021), no. 4, 045003.

[10] John F Clauser, Michael A Horne, Abner Shimony, and Richard A Holt, *Proposed experiment to test local hidden-variable theories*, Physical review letters **23** (1969), no. 15, 880.

[11] Charles A Coulson, *Present state of molecular structure calculations*, Reviews of Modern Physics **32** (1960), no. 2, 170.

[12] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.

[13] Bryce Seligman Dewitt and Neill Graham, *The many-worlds interpretation of quantum mechanics*, Vol. 63, Princeton University Press, 2015.

[14] Andrew C Doherty, Pablo A Parrilo, and Federico M Spedalieri, *Complete family of separability criteria*, Physical Review A **69** (2004), no. 2, 022308.

[15] Wolfgang Dür, Guifre Vidal, and J Ignacio Cirac, *Three qubits can be entangled in two inequivalent ways*, Physical Review A **62** (2000), no. 6, 062314.

[16] Christopher A Fuchs, N David Mermin, and Rüdiger Schack, *An introduction to QBism with an application to the locality of quantum mechanics*, American Journal of Physics **82** (2014), no. 8, 749–754.

[17] James Gleick, *The information: A history, a theory, a flood*, Vintage, 2011.

[18] Leonid Gurvits, *Classical deterministic complexity of Edmonds' problem and quantum entanglement*, Proceedings of the thirty-fifth annual acm symposium on theory of computing, 2003, pp. 10–19.

[19] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu, *Sample-optimal tomography of quantum states*, Proceedings of the forty-eighth annual acm symposium on theory of computing, 2016, pp. 913–925.

[20] Patrick Hayden, Richard Jozsa, Denes Petz, and Andreas Winter, *Structure of states which satisfy strong subadditivity of quantum entropy with equality*, Communications in mathematical physics **246** (2004), 359–374.

[21] Bas Hensen, Hannes Bernien, Anaïs E Dréau, Andreas Reiserer, Norbert Kalb, Machiel S Blok, Just Ruitenberg, Raymond FL Vermeulen, Raymond N Schouten, and Carlos Abellán, *Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres*, Nature **526** (2015), no. 7575, 682–686.

[22] Alexander S Holevo, *Quantum systems, channels, information: a mathematical introduction*, Walter de Gruyter GmbH & Co KG, 2019.

[23] Michael Horodecki, Peter W Shor, and Mary Beth Ruskai, *Entanglement breaking channels*, Reviews in Mathematical Physics **15** (2003), no. 06, 629–641.

[24] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki, *Quantum entanglement*, Reviews of modern physics **81** (2009), no. 2, 865.

[25] Elliott H Lieb and Mary Beth Ruskai, *Proof of the strong subadditivity of quantum-mechanical entropy*, Les rencontres physiciens-mathématiciens de Strasbourg-RCP25 **19** (1973), 36–55.

[26] Ting-Chun Lin, Isaac H Kim, and Min-Hsiu Hsieh, *A new operator extension of strong subadditivity of quantum entropy*, Letters in Mathematical Physics **113** (2023), no. 3, 68.

[27] Yi-Kai Liu, *Consistency of local density matrices is QMA-complete*, Random 2006, proceedings, 2006, pp. 438–449.

[28] Yi-Kai Liu, Matthias Christandl, and Frank Verstraete, *Quantum computational complexity of the N-representability problem: QMA complete*, Physical review letters **98** (2007), no. 11, 110503.

[29] David JC MacKay, *Information theory, inference and learning algorithms*, Cambridge university press, 2003.

[30] David Mermin, *What's wrong with this pillow?*, Physics Today **42** (1989), no. 4, 9–11.

[31] Michael A Nielsen and Isaac L Chuang, *Quantum computation and quantum information*, Cambridge university press, 2010.

[32] Masanori Ohya and Dénes Petz, *Quantum entropy and its use*, Springer Science & Business Media, 2004.

[33] Roland Omnes, *The interpretation of quantum mechanics*, Vol. 102, Princeton University Press, 2018.

[34] Nobel Prize Outreach, *The Nobel Prize in Physics 2022*, 2022. https://www.nobelprize.org/prizes/physics/2022/summary/.

[35] Asher Peres, *Separability criterion for density matrices*, Physical Review Letters **77** (1996), no. 8, 1413.

[36] Nicholas Pippenger, *The inequalities of quantum information theory*, IEEE Transactions on Information Theory **49** (2003), no. 4, 773–789.

[37] John Preskill, *Lecture notes for physics 229: Quantum information and computation*, California Institute of Technology **16** (1998), no. 1, 1–8.

[38] Joseph Renes, *Quantum information theory: Concepts and methods*, Walter de Gruyter GmbH & Co KG, 2022.

[39] Benjamin Schumacher, *Quantum coding*, Physical Review A **51** (1995), no. 4, 2738.

[40] Claude Elwood Shannon, *A mathematical theory of communication*, The Bell system technical journal **27** (1948), no. 3, 379–423.

[41] Ivan Šupić and Joseph Bowles, *Self-testing of quantum systems: a review*, Quantum **4** (2020), 337.

[42] Marco Tomamichel, *Quantum information processing with finite resources: mathematical foundations*, Vol. 5, Springer, 2015.

[43] John Von Neumann, *Mathematical foundations of quantum mechanics*, Vol. 53, Princeton University Press, 1955.

[44] Michael Walter, Brent Doran, David Gross, and Matthias Christandl, *Entanglement polytopes: multiparticle entanglement from single-particle information*, Science **340** (2013), no. 6137, 1205–1208.

[45] John Watrous, *The theory of quantum information*, Cambridge university press, 2018.

[46] Gregor Weihs, Thomas Jennewein, Christoph Simon, Harald Weinfurter, and Anton Zeilinger, *Violation of Bell's inequality under strict Einstein locality conditions*, Physical Review Letters **81** (1998), no. 23, 5039.

[47] Mark M Wilde, *Quantum information theory*, Cambridge university press, 2013.

[48] Michael M Wolf and J Ignacio Cirac, *Dividing quantum channels*, Communications in Mathematical Physics **279** (2008), 147–168.

[49] Bei Zeng, Xie Chen, Duan-Lu Zhou, Xiao-Gang Wen, et al., *Quantum information meets quantum matter*, Springer, 2019.

[50] Zhen Zhang and Raymond W Yeung, *On characterization of entropy function via information inequalities*, IEEE Transactions on Information Theory **44** (1998), no. 4, 1440–1452.

Appendix A

Linear algebra

Most of the mathematics involved in quantum information theory and quantum mechanics more generally is linear algebra. We will now set up notation and review some of the crucial basic notions in linear algebra. Unless mentioned otherwise, all vector spaces are complex vector spaces. We will first introduce bra-ket notation. Then we discuss important classes of linear operators: Hermitian, positive, unitary and isometric operators. Next, we define the tensor product.

We denote by $\mathbb{C}^d$ the Hilbert space of column vectors over the complex numbers

$$v = \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{d-1} \end{pmatrix}$$

where $v_i \in \mathbb{C}$ and with the standard inner product

$$\langle v|w\rangle = \sum_{i=1}^{d} \overline{v_i} w_i.$$

If $\mathcal{H}$ is an arbitrary Hilbert space with $\dim(\mathcal{H}) = d$ we may choose an orthonormal basis $\{|e_i\rangle\}_{i=0}^{d-1}$ for $\mathcal{H}$, that is

$$\langle e_i|e_j\rangle = \delta_{ij}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.[1] We may then use this basis to identify $\mathcal{H}$ with $\mathbb{C}^d$ with the standard inner product. Throughout these lectures, a *basis* will always be understood to be an orthonormal basis.

In these lectures we will use *bra-ket* notation. We write vectors in $\mathcal{H}$ as $|\psi\rangle$ (a 'ket') and we write a 'bra' for the dual vector $\langle\psi| \in cH^*$ which is the functional on $\mathcal{H}$ mapping

$$|\phi\rangle \mapsto \langle\psi|\phi\rangle.$$

The logic of this notation is such that composing a 'bra' $\langle\psi|$ with a 'ket' $|\phi\rangle$ gives the 'bracket' inner product $\langle\psi|\phi\rangle$, so

$$\langle\psi||\phi\rangle = \langle\psi|\phi\rangle.$$

[1] We let indices run from 0 to $d-1$ here, because in the special case $d = 2$ we would like the labels 0 and 1 to correspond to a *bit*.

This is perhaps a little abstract, but as we will explain below you can always think of $|\psi\rangle$ as a *column* vector and $\langle\psi|$ as a *row* vector.

In the special case where $\mathcal{H} = \mathbb{C}^d$ we introduce the following notation for the standard basis:

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad |1\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \cdots \qquad |d-1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

If $|\psi\rangle$ is the column vector

$$|\psi\rangle = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{d-1} \end{pmatrix}$$

then we may also write this as

$$|\psi\rangle = \sum_{i=0}^{d-1} \psi_i |i\rangle.$$

The dual vector $\langle\psi|$ is a row vector and is given by

$$\langle\psi| = \begin{pmatrix} \overline{\psi_0} & \overline{\psi_1} & \cdots & \overline{\psi_{d-1}} \end{pmatrix}.$$

Note that a column vector can also be read as a $d \times 1$ matrix, and a row vector as a $1 \times d$ matrix, and the inner product is obtained by multiplying these matrices.

In general, we can always choose a basis to identify $\mathcal{H}$ with $\mathbb{C}^d$ for $d = \dim(\mathcal{H})$. If $\Sigma = \{|e_0\rangle, \ldots, |e_{d-1}\rangle\}$ is a (fixed choice of) orthonormal basis we usually write $|i\rangle := |e_i\rangle$, and we can expand any $|\psi\rangle \in \mathcal{H}$ as

$$|\psi\rangle = \sum_{i=0}^{d-1} \psi_i |i\rangle,$$

where $\psi_i = \langle i|\psi\rangle$. The dual (row) vector is

$$\langle\psi| = \sum_{i=0}^{d-1} \overline{\psi_i} \langle i|.$$

An inner product satisfies the *Cauchy-Schwarz* inequality: for any two $|\phi\rangle, |\psi\rangle \in \mathcal{H}$

$$|\langle\phi|\psi\rangle|^2 \leq \langle\phi|\phi\rangle\langle\psi|\psi\rangle \tag{A.1}$$

with equality only if $|\phi\rangle$ is proportional to $|\psi\rangle$. The norm of a vector is defined by $\||\psi\rangle\| = \sqrt{\langle\psi|\psi\rangle}$, so this can also be written as

$$|\langle\phi|\psi\rangle| \leq \||\phi\rangle\| \; \||\psi\rangle\|.$$

A.1 Linear maps and matrices

Given a Hilbert space $\mathcal{H} = \mathbb{C}^d$ and $\mathcal{K} = \mathbb{C}^e$ we denote by $\mathrm{Lin}(\mathcal{H}, \mathcal{K})$ the set of linear maps from $\mathcal{H}$ to $\mathcal{K}$, which after a choice of basis may be identified with the set of complex $e \times d$ matrices. We abbreviate $\mathrm{Lin}(\mathcal{H}, \mathcal{H}) = \mathrm{Lin}(\mathcal{H})$. We write $\mathrm{im}(M)$ for the image of $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ and $\mathrm{rank}(M)$ for its rank (which is the dimension of $\mathrm{im}(M)$). We let $\mathbb{1}$ denote the identity operator.

The *trace* of $M \in \mathrm{Lin}(\mathcal{H})$ is computed by choosing a basis Σ of $\mathcal{H}$ and summing the diagonal entries of the matrix representation of M. In bra-ket notation we write

$$\mathrm{tr}[M] = \sum_{i \in \Sigma} \langle i|M|i\rangle.$$

This does not depend on the choice of basis and it has the important *cyclicity property*, meaning that for any operators $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ and $N \in \mathrm{Lin}(\mathcal{K}, \mathcal{H})$ we have $\mathrm{tr}[MN] = \mathrm{tr}[NM]$. An important special case is where $N = |\psi\rangle\langle\psi|$ for $|\psi\rangle \in \mathcal{H}$ and $M \in \mathrm{Lin}(\mathcal{H})$

$$\mathrm{tr}[MN] = \langle\psi|M|\psi\rangle. \tag{A.2}$$

If $M \in \mathrm{Lin}(\mathcal{H})$, and if we are given a nonzero $|\psi\rangle \in \mathcal{H}$ and $\lambda \in \mathbb{C}$ such that

$$M|\psi\rangle = \lambda|\psi\rangle$$

we say that $|\psi\rangle$ is an eigenvector of M with eigenvalue λ.

Given $M \in \mathrm{Lin}(\mathcal{H}_1, \mathcal{H}_2)$ we may choose bases Σ_1 and Σ_2 of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively and expand M as

$$M = \sum_{i \in \Sigma_2} \sum_{j \in \Sigma_1} M_{ij} |i\rangle\langle j|$$

where the coefficients M_{ij} are the matrix coefficients and are given by $M_{ij} = \langle i|M|j\rangle$. The adjoint of a linear map $M \in \mathrm{Lin}(\mathcal{H}_1, \mathcal{H}_2)$ is the operator $M^\dagger \in \mathrm{Lin}(\mathcal{H}_2, \mathcal{H}_1)$ which is such that

$$\langle\psi|M|\phi\rangle = \overline{\langle\phi|M^\dagger|\psi\rangle}$$

for all $|\psi\rangle \in \mathcal{H}_2$ and $|\phi\rangle \in \mathcal{H}_1$. In bra-ket notation,

$$M^\dagger = \sum_{i,j} \overline{M_{ij}} |j\rangle\langle i|.$$

The transpose of M is defined by

$$M^\mathsf{T} = \sum_{i,j} M_{ij} |j\rangle\langle i|.$$

In particular, the adjoint is the conjugate transpose $M^\dagger = \overline{M^\mathsf{T}}$. Note that the transpose depends on the choice of basis, whereas the adjoint does not. If $M \in \mathrm{Lin}(\mathcal{H}_2, \mathcal{H}_3)$ and $N \in \mathrm{Lin}(\mathcal{H}_1, \mathcal{H}_2)$ we have $(MN)^\dagger = N^\dagger M^\dagger$.

Hermitian operators

An operator $M \in \mathrm{Lin}(\mathcal{H})$ is called *Hermitian* (or self-adjoint) if $M = M^\dagger$. After a choice of basis, this means that the associated matrix has diagonal entries M_{ii} which must be real, and $M_{ij} = \overline{M_{ji}}$ for $i \neq j$. Hermitian matrices have the property that they have real eigenvalues and one can find a basis of eigenvectors. This result is very important in quantum mechanics!

Theorem A.1 (Spectral theorem for Hermitian operators)**.** *Suppose $M \in \operatorname{Lin}(\mathcal{H})$ is a Hermitian operator and $\dim(\mathcal{H}) = d$, then there exist real numbers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and a basis $\{\psi_i\}_{i=1}^d$ such that*

$$M = \sum_{i=1}^{d} \lambda_i |\psi_i\rangle\langle\psi_i|. \tag{A.3}$$

The λ_i are the eigenvalues of M, and the $|\psi_i\rangle$ are eigenvectors. We will also call the set $\{\lambda_i\}_{i=1}^d$ the *spectrum* of M. Theorem A.1 is equivalent to the fact that a Hermitian matrix can be diagonalized using unitary matrices, and the resulting diagonal matrix has real values.

The spectrum is uniquely determined by the matrix, but the eigenvectors $|\psi_i\rangle$ need not be unique if the spectrum is degenerate. For instance for the identity matrix we have

$$\mathbb{1} = \sum_{i=1}^{d} |\psi_i\rangle\langle\psi_i|$$

for any choice of basis.

The spectral theorem is a useful way to define functions of operators. Note that if M is an operator, there is a natural way to define the operator M^k for integer k (just compose the matrix k times). What about $\sqrt{M}$?

For a Hermitian operator M and some single-variable function f we may take a spectral decomposition as in Eq. (A.3)[2] and we may define $f(M)$ by applying f to the spectrum

$$f(M) = \sum_{i=1}^{d} f(\lambda_i) |\psi_i\rangle\langle\psi_i|. \tag{A.4}$$

This only requires that f is well-defined on the λ_i (for instance, for the square root function $\sqrt{M}$ one requires that $\lambda_i \geq 0$).

Positive operators

An operator $P \in \operatorname{Lin}(\mathcal{H})$ is *positive semidefinite* (abbreviated as PSD) or simply *positive* if for all $|\psi\rangle \in \mathcal{H}$ we have $\langle\psi|P|\psi\rangle \geq 0$. We denote the set of all positive operators on $\mathcal{H}$ by $\operatorname{PSD}(\mathcal{H})$ and we will also write $P \geq 0$ for $P \in \operatorname{PSD}(\mathcal{H})$. Similarly, we can also define a *positive definite* (PD) operator $P \in \operatorname{Lin}(\mathcal{H})$ by demanding that for all nonzero $|\psi\rangle \in \mathcal{H}$ we have the strict inequality $\langle\psi|P|\psi\rangle > 0$. We denote the set of PD operators on $\mathcal{H}$ by $\operatorname{PD}(\mathcal{H})$ and write $P > 0$ for $P \in \operatorname{PD}(\mathcal{H})$. The following result provides a number of different characterizations of positive matrices.

Lemma A.2. *Let $P \in \operatorname{Lin}(\mathcal{H})$. The following are equivalent:*

(a) P is positive, i.e. $\langle\psi|P|\psi\rangle \geq 0$ for all $|\psi\rangle \in \mathcal{H}$.

(b) P is Hermitian and all eigenvalues are non-negative.

(c) There exists $M \in \operatorname{Lin}(\mathcal{H}, \mathcal{K})$ such that $P = M^\dagger M$ for some Hilbert space $\mathcal{K}$.

(d) For every $Q \in \operatorname{PSD}(\mathcal{H})$ we have $\operatorname{tr}[PQ] \geq 0$.

[2] This method of applying functions to operators can be extended to the broader class of *normal operators* for which $A^\dagger A = AA^\dagger$, by generalizing the spectral theorem to this class of operators. We will not need this. [MW: Might be nice to do, so that it also applies to unitaries.]

The proof is Exercise 1.16. Note that in particular positive matrices are Hermitian which is not obvious from the definition. The following is an easy consequence of Lemma A.2.

Corollary A.3. *If* $P \in \mathrm{PSD}(\mathcal{H})$ *and* $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$*, then* $MPM^\dagger \in \mathrm{PSD}(\mathcal{K})$.

Proof. By Lemma A.2 $P = N^\dagger N$ for some $N \in \mathrm{Lin}(\mathcal{H}, \mathcal{H}')$. Then

$$MPM^\dagger = MN^\dagger NM^\dagger = (NM^\dagger)^\dagger (NM^\dagger)$$

and again by Lemma A.2 $MPM^\dagger \geq 0$. □

The notion of positivity give rise to an order relation on matrices where we say that $P \leq Q$ for $P, Q \in \mathrm{Lin}(\mathcal{H})$ if $Q - P \geq 0$. For example, $P \leq \alpha \mathbb{1}$ for $\alpha \in \mathbb{R}$is equivalent to P being Hermitian and having eigenvalues all smaller than α.

Unitaries and isometries

$V \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ is an *isometry* if $V^\dagger V = \mathbb{1}$. This is only possible if $\dim(\mathcal{H}) \leq \dim(\mathcal{K})$. $U \in \mathrm{Lin}(\mathcal{H})$ is a *unitary* if $U^\dagger U = UU^\dagger = \mathbb{1}$ (in which case we must have $\dim(\mathcal{H}) = \dim(\mathcal{K})$). Isometries are such that they preserve inner products (this follows directly from the definition) and hence norms of vectors. This implies that a unitary maps an orthonormal basis to an orthonormal basis.

We denote the set of unitaries on $\mathcal{H}$ by

$$\mathrm{U}(\mathcal{H}) = \{U \in \mathrm{Lin}(\mathcal{H}) : U^\dagger U = UU^\dagger = \mathbb{1}\}$$

and isometries between spaces $\mathcal{H}$ and $\mathcal{K}$ by

$$\mathrm{Isom}(\mathcal{H}, \mathcal{K}) = \{V \in \mathrm{Lin}(\mathcal{H}, \mathcal{K}) : V^\dagger V = \mathbb{1}\}.$$

[MW: It would be nice to say that isometries are unitaries if $\dim(\mathcal{H}) = \dim(\mathcal{K})$, but we did not define $\mathrm{U}(\mathcal{H}, \mathcal{K})$. Should we?] If $V \in \mathrm{Isom}(\mathcal{H}, \mathcal{K})$ then we must have $d \leq e$, where $d = \dim(\mathcal{H})$ and $e = \dim(\mathcal{K})$, and the isometry identifies $\mathcal{H}$ with the subspace $V(\mathcal{H}) \subseteq \mathcal{K}$. If $d = e$, then the two spaces are equal. If $\mathcal{H} \subset \mathcal{K}$ is a subspace then any isometry $V \in \mathrm{Isom}(\mathcal{H}, \mathcal{K})$ can be extended to a unitary $U \in \mathrm{U}(\mathcal{K})$. That is, there exists $U \in \mathrm{U}(\mathcal{K})$ such that U restricted to $\mathcal{H}$ equals V. You may show this in Exercise 1.13.

Projections

An operator $P \in \mathrm{Lin}(\mathcal{H})$ is called a *projection* if $P^2 = P$. We will moreover always assume that P is Hermitian (in other contexts these are called *orthogonal projections* to make the distinction). Suppose that $\{|e_i\rangle\}_{i=1}^r$ is a basis for the image of a projection P. Then

$$P = \sum_{i=1}^r |e_i\rangle\langle e_i|,$$

as you may show in Exercise 1.12. If $M \in \mathrm{Lin}(\mathcal{H})$ is Hermitian, then if

$$M = \sum_{i=1}^d \lambda_i |\psi_i\rangle\langle\psi_i|$$

is a spectral decomposition, we may define the projectors

$$P_\lambda = \sum_{i:\lambda_i=\lambda} |\psi_i\rangle\langle\psi_i|$$

so if we let Λ denote the set of *distinct* eigenvalues

$$M = \sum_{\lambda\in\Lambda} \lambda P_\lambda.$$

This decomposition of M is unique.

A.1.1 Singular value decomposition

For certain classes of operators we know that we can diagonalise the operator (i.e. choose a basis in which the operator is diagonal), in which case the values in the diagonal matrix are the eigenvalues of the matrix. Indeed, the spectral theorem in Theorem A.1 shows that this is the case for Hermitian operators. For arbitrary $M \in \mathrm{Lin}(\mathcal{H},\mathcal{K})$ (so after choosing a basis the matrix need not even be square) there is a useful decomposition known as the *singular value decomposition.*

Theorem A.4 (Singular value decomposition). *Suppose $M \in \mathrm{Lin}(\mathcal{H},\mathcal{K})$. Then there exist bases $\{e_i\}$ and $\{f_i\}$ of $\mathcal{K}$ and $\mathcal{H}$ and a collection of positive numbers $s_1 \geq \cdots \geq s_r > 0$ for $r = \mathrm{rank}(M)$ such that*

$$M = \sum_{i=1}^{r} s_i |e_i\rangle\langle f_i|$$

Note that $M^\dagger M$ and $MM^\dagger$ are Hermitian (and positive) matrices. It is easy to see that the nonzero part of the spectrum of $M^\dagger M$ and $MM^\dagger$ is given by the numbers $\{s_i^2\}$. This observation is key to proving Theorem A.4 and also suggests a way to compute the singular values of a matrix.

An alternative formulation of the singular value decomposition is that for $M \in \mathrm{Lin}(\mathcal{H},\mathcal{K})$, there exist isometries $U \in \mathrm{Isom}(\mathbb{C}^r,\mathcal{H})$ and $V \in \mathrm{U}(\mathbb{C}^r,\mathcal{K})$ such that

$$M = VSU^\dagger \qquad S = \sum_{i=1}^{r} s_i |i\rangle\langle i|$$

where we let $|1\rangle, \dots, |r\rangle$ denote the standard basis for $\mathbb{C}^r$. Note that S is a diagonal matrix with the positive numbers s_i on the diagonal. The isometries U and V can be defined by

$$U = \sum_{i=1}^{r} |e_i\rangle\langle i| \qquad V = \sum_{i=1}^{r} |f_i\rangle\langle i|.$$

In the special case where M is Hermitian and has eigenvalues λ_i and eigenvectors $|\psi_i\rangle$ one finds a singular value decomposition with $s_i = |\lambda_i|$ and $|e_i\rangle = |\psi_i\rangle$, $|f_i\rangle = \mathrm{sign}(\lambda_i)|\psi_i\rangle$ for the nonzero eigenvalues λ_i.

A.2 Tensor products

We now introduce an additional ingredient from linear algebra required to describe combinations of multiple quantum systems.

Definition A.5. If $\mathcal{H}$ and $\mathcal{K}$ are Hilbert spaces, their *tensor product* $\mathcal{H} \otimes \mathcal{K}$ is the Hilbert space which consists of the linear span of elements of the form $v \otimes w$ for $v \in \mathcal{H}$ and $w \in \mathcal{K}$, subject to the relations

$$\begin{aligned}(\alpha v) \otimes (\beta w) &= \alpha\beta v \otimes w \text{ for } \alpha, \beta \in \mathbb{C}\\ (v_1 + v_2) \otimes w &= v_1 \otimes w + v_2 \otimes w\\ v \otimes (w_1 + w_2) &= v \otimes w_1 + v \otimes w_2\end{aligned}$$

for any $v, v_1, v_2 \in \mathcal{H}$ and $w, w_1, w_2 \in \mathcal{K}$. The inner product is defined by linear extension of the relation

$$\langle v_1 \otimes w_1 | v_2 \otimes w_2 \rangle = \langle v_1 | v_2 \rangle \langle w_1 | w_2 \rangle$$

We again use bra-ket notation and write $|\phi\rangle \otimes |\psi\rangle$. If we choose bases $\Sigma_{\mathcal{H}}$ and $\Sigma_{\mathcal{K}}$ for $\mathcal{H}$ and $\mathcal{K}$ and if $|\psi\rangle \in \mathcal{H}$, $|\phi\rangle \in \mathcal{K}$ with $|\psi\rangle = \sum_{i \in \Sigma_{\mathcal{H}}} \psi_i |i\rangle$ and $|\phi\rangle = \sum_{j \in \Sigma_{\mathcal{K}}} \phi_j |j\rangle \in$ then we may expand

$$|\psi\rangle \otimes |\phi\rangle = \sum_{i,j} \psi_i \phi_j |i\rangle \otimes |j\rangle.$$

This implies that the set $\{|i\rangle \otimes |j\rangle\}_{i,j=1}^{d,e}$ is a basis for $\mathcal{H} \otimes \mathcal{K}$ (it is easy to verify that these elements are pairwise orthogonal). In particular, the dimension of the tensor product is multiplicative:

$$\dim(\mathcal{H} \otimes \mathcal{K}) = \dim(\mathcal{H}) \dim(\mathcal{K}).$$

This product basis is in most cases a more useful way to reason about the tensor product than the abstract definition Definition A.5. If Ω_1 and Ω_2 are two finite sets, then by the above construction of a product basis

$$\begin{aligned}\mathbb{C}^{\Omega_1} \otimes \mathbb{C}^{\Omega_2} &\cong \Big\{ \sum_{x_1 \in \Omega_1} \sum_{x_2 \in \Omega_2} v_{x_1,x_2} |x_1\rangle \otimes |x_2\rangle \Big\}\\ &\cong \Big\{ \sum_{x=(x_1,x_2) \in \Omega_1 \times \Omega_2} v_x |x\rangle \Big\} \cong \mathbb{C}^{\Omega_1 \times \Omega_2}.\end{aligned}$$

The tensor product is associative in the sense that if we have three Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$, then

$$\mathcal{H}_1 \otimes (\mathcal{H}_2 \otimes \mathcal{H}_3) \cong (\mathcal{H}_1 \otimes \mathcal{H}_2) \otimes \mathcal{H}_3.$$

We will therefore identify this with a Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_3$, and we may similarly define tensor products with more factors. In particular, we may identify

$$\mathbb{C}^{\Omega_1} \otimes \mathbb{C}^{\Omega_2} \otimes \ldots \otimes \mathbb{C}^{\Omega_n} \cong \mathbb{C}^{\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n}.$$

in which case we have a product basis

$$\{|x_1\rangle \otimes \ldots \otimes |x_n\rangle : (x_1, \ldots, x_n) \in \Omega_1 \times \cdots \times \Omega_n\}.$$

In particular, if $\Omega_k = \{0, \ldots, d_k - 1\}$, where $\mathbb{C}^{\Omega_k} \cong \mathbb{C}^{d_k}$ for $k = 1, \ldots, n$ we get the standard product basis labeled by strings $(i_1, \ldots, i_n)$ for $i_k \in \{0, \ldots, d_k - 1\}$. In bra-ket notation we will often abbreviate product bases as

$$|i_1\rangle \otimes |i_2\rangle \otimes \ldots \otimes |i_n\rangle = |i_1 i_2 \ldots i_n\rangle.$$

We will also use the notation

$$|\phi\rangle|\psi\rangle = |\phi\rangle \otimes |\psi\rangle \text{ for } |\phi\rangle \in \mathcal{H}, |\psi\rangle \in \mathcal{K}$$

so in many cases we do not write the $\otimes$-symbol explicitly. A final piece of notation is that if we have $n \geq 1$ tensor products of the same element we use the shorthand

$$\mathcal{H}^{\otimes n} := \underbrace{\mathcal{H} \otimes \ldots \otimes \mathcal{H}}_{n \text{ times}} \text{ and } |\phi\rangle^{\otimes n} = \underbrace{|\phi\rangle \otimes \ldots \otimes |\phi\rangle}_{n \text{ times}}.$$

Finally, we can also define tensor products of operators. If $M \in \operatorname{Lin}(\mathcal{H}_1, \mathcal{H}_2)$, $N \in \operatorname{Lin}(\mathcal{K}_1, \mathcal{K}_2)$ then we can define a linear $M \otimes N$ in $\operatorname{Lin}(\mathcal{H}_1 \otimes \mathcal{K}_1, \mathcal{H}_2 \otimes \mathcal{K}_2)$ by linear extension of

$$(M \otimes N)|\phi\rangle \otimes |\psi\rangle = M|\phi\rangle \otimes N|\psi\rangle \text{ for all } |\phi\rangle \in \mathcal{H}_1, |\psi\rangle \in \mathcal{K}_1. \tag{A.5}$$

In this way, one can show that $\operatorname{Lin}(\mathcal{H} \otimes \mathcal{K}) \cong \operatorname{Lin}(\mathcal{H}) \otimes \operatorname{Lin}(\mathcal{K})$ as vector spaces. We defined the tensor product of operators by Eq. (A.5) by its action on tensor product states and linear extension. Concretely, in terms of a choice of basis for $\mathcal{H}$ and $\mathcal{K}$ we we may expand $M \in \operatorname{Lin}(\mathcal{H})$ and $N \in \operatorname{Lin}(\mathcal{K})$ as

$$M = \sum_{i,j} M_{ij} |i\rangle\langle j| \text{ and } N = \sum_{k,l} N_{kl} |k\rangle\langle l|.$$

Then, the tensor product operator is given by

$$M \otimes N = \sum_{i,j,k,l} M_{ij} N_{kl} |i\rangle\langle j| \otimes |k\rangle\langle l| = \sum_{i,j,k,l} M_{ij} N_{kl} |ik\rangle\langle jl|.$$

The identification $|i\rangle\langle j| \otimes |k\rangle\langle l| = |ik\rangle\langle jl|$ corresponds to the isomorphism $\operatorname{Lin}(\mathcal{H} \otimes \mathcal{K}) \cong \operatorname{Lin}(\mathcal{H}) \otimes \operatorname{Lin}(\mathcal{K})$.

Useful facts about tensor product operators are

Lemma A.6. *(a) If $P \in \operatorname{PSD}(\mathcal{H})$, $Q \in \operatorname{PSD}(\mathcal{K})$, then $P \otimes Q \in \operatorname{PSD}(\mathcal{H} \otimes \mathcal{K})$.*

(b) For any $M \in \operatorname{Lin}(\mathcal{H})$, $N \in \operatorname{Lin}(\mathcal{K})$, we have

$$\operatorname{tr}[M \otimes N] = \operatorname{tr}[M] \operatorname{tr}[N].$$

(c) For any $M \in \operatorname{Lin}(\mathcal{H})$, $N \in \operatorname{Lin}(\mathcal{K})$, we have

$$\operatorname{rank}(M \otimes N) = \operatorname{rank}(M) \operatorname{rank}(N).$$

The proof is Exercise 2.5.

Remark A.7. For any Hilbert space $\mathcal{H}$ we have $\mathcal{H} \otimes \mathbb{C} \cong \mathcal{H}$ simply by identifying $|\phi\rangle \otimes z \cong z|\phi\rangle$ for all $|\phi\rangle \in \mathcal{H}$ and $z \in \mathbb{C}$. We will often use this identification without comment. We will also consider operators of the form $M \otimes |\psi\rangle$ and $M \otimes \langle\psi|$ for $M \in \operatorname{Lin}(\mathcal{H})$ and $|\psi\rangle \in \mathcal{K}$. The operator $M \otimes |\psi\rangle : \mathcal{H} \to \mathcal{H} \otimes \mathcal{K}$ is defined by

$$(M \otimes |\psi\rangle)|\phi\rangle = M|\phi\rangle \otimes |\psi\rangle \text{ for all } |\phi\rangle \in \mathcal{H}.$$

The operator $M \otimes \langle\psi| : \mathcal{H} \otimes \mathcal{K} \to \mathcal{H} \otimes \mathbb{C} \cong \mathcal{H}$ is similarly given by

$$(M \otimes \langle\psi|)|\phi\rangle \otimes |\chi\rangle = \langle\psi|\chi\rangle M|\phi\rangle \text{ for all } |\phi\rangle \in \mathcal{H}, |\chi\rangle \in \mathcal{K}.$$

A.3 Norms of vectors and linear operators

Here we recall some background on norms. We don't use bra-ket notation because we discuss general vector spaces that do not necessary have an inner product. If V is a complex vector space, a function $\|\cdot\| : V \to \mathbb{R}$ is a *norm* if

(a) $\|v\| \geq 0$ for all $v \in V$, with equality if and only if $v = 0$,

(b) $\|\lambda v\| = |\lambda| \, \|v\|$ for all $v \in V$ and $\lambda \in \mathbb{C}$, and

(c) the *triangle inequality* $\|v + w\| \leq \|v\| + \|w\|$ holds for all $v, w \in V$.

If $\|\cdot\|$ is a norm, then $d(v, w) := \|v - w\|$ defines a metric on V, meaning it satisfies

(a) $d(v, w) \geq 0$ for all $v, w \in V$, with equality if and only if $v = w$,

(b) $d(v, w) = d(w, v)$ for all all $v, w \in V$, and

(c) the *triangle inequality* $d(v, w) \leq d(v, u) + d(u, w)$ holds for all $u, v, w \in V$.

As an important example, if we have a vector

$$v = \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{d-1} \end{pmatrix} \in \mathbb{C}^d,$$

then we may define its *p-norm* (or ℓ^p-norm) for $p \in [1, \infty)$ as

$$\|v\|_p = \left(\sum_{i=0}^{d-1} |v_i|^p \right)^{\frac{1}{p}} .$$

For $p = 2$ this gives the standard Euclidean norm, which comes from the standard inner product. We denote it simply by $\|v\| = \|v\|_2$. More generally, for any Hilbert spaces $\mathcal{H}$ one can define a Euclidean norm by the formula $\|v\| := \sqrt{\langle v|v\rangle}$.

Schatten norms

Since the space $\mathrm{Lin}(\mathcal{H}, \mathcal{K})$ is a vector space, we can also define norms of operators. We may define operator norms by taking the p-norm of the singular values of the operator. We will discuss in Section 6.1 some important special cases, here we give the general definition.

Definition A.8 (Schatten p-norm)**.** If $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ has singular values $(s_i)_{i=1}^r$ we define the *Schatten p-norm* for $p \in [1, \infty)$ by

$$\|M\|_p = \left(\sum_{i=1}^{r} s_i^p \right)^{\frac{1}{p}} .$$

From the definition of the singular values it is easy to see that alternatively

$$\|M\|_p = \left(\mathrm{tr}[(M^\dagger M)^{\frac{p}{2}}] \right)^{\frac{1}{p}} .$$

Except when $p = 2$, these are *not* the same as the p-norm of the vector formed from the entries of the matrix M with respect to a basis.

If we take the limit of p to ∞ we get the following norm:

Definition A.9. The Schatten ∞-norm, or *operator norm* $\|\cdot\|_\infty$ of $M \in \mathrm{Lin}(\mathcal{H}, \mathcal{K})$ is defined as

$$\|M\|_\infty = s_1$$

where s_1 is the largest singular value of M.

The operator norm can also be defined without reference to the singular values:

$$\|M\|_\infty = \max_{\|v\|=1} \|Mv\|,$$

where the norms on the right-hand side are the norms on $\mathcal{H}$ and $\mathcal{K}$, respectively.

Here are some basic properties of the Schatten norms:

Lemma A.10. *(a) The Schatten p-norm defines a norm on $\mathrm{Lin}(\mathcal{H}, \mathcal{K})$ for all $p \in [1, \infty]$.*

(b) The Schatten p-norm is invariant under isometries: if V and W are isometries $\|VMW^\dagger\|_p = \|M\|_p$.

(c) We have

$$\|M\|_p = \|M^\dagger\|_p = \|M^\mathsf{T}\|_p = \|\overline{M}\|_p$$

where the latter two are with respect to any choice of basis.

Appendix B

Probability theory

Classical information theory models *sources* as probability distributions. In this appendix we provide some basic background on probability theory, reminding the reader of a few fundamental facts.

In most of these notes we are only concerned with probability distributions on finite sets. If we have a set of *outcomes* Ω (which we think of as an alphabet of symbols in the information-theoretic setting), then probability distributions on Ω are simply given by

$$\Pr(\Omega) = \{p : \Omega \to \mathbb{R}_{\geq 0} \text{ and } \sum_{x \in \Omega} p(x) = 1\}.$$

To deal with the most general situation for infinite sets one can use *measure theory*, but since we work almost exclusively with finite outcome sets we will not introduce this formalism.

Given a probability distribution $p \in \Omega$, a *random variable* is a function $\mathbf{X} : \Omega \to E$ where E is some space. For our purposes, E will either be a finite set, the real numbers $\mathbb{R}$ or a vector space. Given a random variable, it will take outcome $e \in E$ with probability

$$\Pr(\mathbf{X} = e) = \sum_{\substack{x \in \Omega \\ f(x) = e}} p(x).$$

If E is a set with addition (such as $\mathbb{R}$ or a vector space) the *expectation value* of a random variable is defined to be

$$\mathbb{E}\mathbf{X} = \sum_x x \Pr(\mathbf{X} = x)$$

The expectation value is a linear operation: if $\mathbf{X}$ and $\mathbf{Y}$ are expectation values on the same vector space,

$$\mathbb{E}(\alpha \mathbf{X} + \beta \mathbf{Y}) = \alpha \mathbb{E}\mathbf{X} + \beta \mathbb{E}\mathbf{Y}$$

for scalars α and β. The *variance* of a random variable $\mathbf{X}$ is given by

$$\mathrm{Var}(\mathbf{X}) = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})^2 = \sum_x \Pr(\mathbf{X} = x) \left(x - \mathbb{E}X\right)^2.$$

It is easy to verify that

$$\mathrm{Var}(\mathbf{X}) = \mathbb{E}\left(\mathbf{X}^2\right) - \left(\mathbb{E}\mathbf{X}\right)^2.$$

The *standard deviation* of $\mathbf{X}$ is the square root of $\mathrm{Var}(\mathbf{X})$.

Convex functions and Jensen's inequality

If $\mathbf{X}$ is a random variable taking values in $\mathbb{R}$, we may apply a real-valued function f to the random variable to get a random variable $f(\mathbf{X})$, simply by composition. We may compute the expectation value of $\mathbf{X}$, and then apply f, to get $f(\mathbb{E}\mathbf{X})$, or we may first apply f and then compute the expectation value to find $\mathbb{E}f(\mathbf{X})$. These are in general different. In the special case where f is convex or concave, we can relate these two values by Jensen's inequality. Recall that a subset I of a vector space V is *convex* if for any $x, y \in I$, $t \in [0,1]$ the element $tx + (1-t)y$ is also in I (meaning that if $x, y \in I$, the line segment between x, y is also in I).

Definition B.1. Let I be a convex set. A function $f : I \to \mathbb{R}$ is *convex* if for any $x, y \in I$ and $t \in [0,1]$

$$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$$

and *concave* if for any $x, y \in I$ and $t \in [0,1]$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

A function is *strictly* convex (or concave) if we have strict inequality for $x \neq y$ and $t \in (0,1)$.

Note that f is concave if and only of $-f$ is convex. A smooth real-valued function on an interval is convex if and only if its second derivative is positive everywhere on the interval. Paradigmatic examples are the function $x \mapsto x^2$ on $\mathbb{R}$ which is convex, and the function $\log : \mathbb{R}_{>0} \to \mathbb{R}$ which is concave.

Lemma B.2 (Jensen's inequality). *Let I be a convex set and $\mathbf{X}$ a random variable taking values in I. If $f : I \to \mathbb{R}$ is a convex function*

$$\mathbb{E}f(\mathbf{X}) \geq f(\mathbb{E}\mathbf{X}).$$

If $f : I \to \mathbb{R}$ is a concave function

$$\mathbb{E}f(\mathbf{X}) \leq f(\mathbb{E}\mathbf{X}).$$

If f is strictly convex (or concave) we have equality if and only if $\mathbf{X}$ is constant.

For example, the function $x \mapsto x^2$ is convex, so

$$\mathbb{E}\mathbf{X}^2 \geq (\mathbb{E}X)^2$$

for any real-valued random variable $\mathbf{X}$. This matches with the fact that $\mathrm{Var}(\mathbf{X}) = \mathbb{E}\left(\mathbf{X}^2\right) - (\mathbb{E}\mathbf{X})^2$ is non-negative.

Concentration bounds and limit theorems

Lemma B.3. *Let* $\mathbf{X}$ *be a random variable and let* $x > 0$*.*

(a) Markov's inequality*: If* $\mathbf{X}$ *takes values in* $\mathbb{R}_{\geq 0}$

$$\Pr(\mathbf{X} \geq x) \leq \frac{\mathbb{E}\mathbf{X}}{x}.$$

(b) Chebyshev's inequality*: If* $\mathbf{X}$ *is any real-valued random variable,*

$$\Pr\big(|\mathbf{X} - \mathbb{E}\mathbf{X}| \geq x\big) \leq \frac{\operatorname{Var}(\mathbf{X})}{x^2}.$$

The proof is Exercise B.1. This can be used to prove the *weak law of large numbers*. The law of large numbers captures the intuitive fact that if you take an average of the outcomes of many independent realizations of a distribution, with high probability the average will be close to the expectation value. For example, if we flip 1000 fair coins, then with high probability the number of heads will not be too far from 500.

Theorem B.4 (Weak law of large numbers)**.** *Let* $(\mathbf{X}_i)_{i\in\mathbb{N}}$ *be an independent and identically distributed sequence of real-valued random variables with mean* μ *and finite variance* σ^2*. Then*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty\ .$$

This can be used to estimate unknown parameters of a distribution. For example, given independent random variables $\mathbf{X}_i$ each of which takes value 1 with probability p and outcome 0 with probability $1 - p$, we have expectation value $\mu = p$ and variance $\sigma^2 = p(1 - p)$. We may now give an estimate $\hat{\mu}$ of μ by taking the average

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i.$$

If we take $n \gg \frac{p(1-p)}{\varepsilon^2}$ we see from Theorem B.4 that the probability that the estimate $\hat{\mu}$ is ε-close to the real value μ is close to 1.

There are many refined versions of the law of large numbers. Especially fundamental is the *central limit theorem*, which concerns the behavior of the deviations of the average and states that the deviations can be expected to be of the order $\frac{1}{\sqrt{n}}$ and behave like a *normally distributed* random variable. While not crucial for these lectures on information theory, we nevertheless state this result. It can for example be used, as in Exercise 8.7, to understand the leading corrections to the compression rate at a finite number of copies.

First, we recall that a real-valued random variable $\mathbf{X}$ has a normal (or Gaussian) distribution with mean μ and standard deviation σ if it has probability density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \mathbb{R}.$$

It has cumulative distribution function

$$F(x) = \int_{\infty}^{x} p(x)\mathrm{d}x.$$

The central limit theorem states that if we have a sequence $\mathbf{X}_i$ of IID random variables with mean μ and standard deviation σ, the random variable

$$\mathbf{X}^{(n)} = \sum_{i=1}^{n} \frac{\mathbf{X}_i - \mu}{\sqrt{n}}$$

converges in distribution to a normal distribution with mean σ. Note that $\mathbf{X}^{(n)}$ is the deviation of the average from μ, multiplied by $\sqrt{n}$. This makes precise the fact that the deviations of the average from μ are of the order $\frac{1}{\sqrt{n}}$ (and are normally distributed). Here we give a version that is not maximally general, but gives a nice bound on the speed of convergence (known as the Berry-Esseen bound).

Theorem B.5. *Let $(\mathbf{X}_i)_{i\in\mathbb{N}}$ be an independent and identically distributed sequence of real-valued random variables with mean μ finite variance σ^2 and finite $\rho = \mathbb{E}|X_i|^3$. Let*

$$\mathbf{X}^{(n)} = \sum_{i=1}^{n} \frac{\mathbf{X}_i - \mu}{\sqrt{n}}$$

then

$$\lim_{n\to\infty} \left|\Pr(\mathbf{X}^{(n)} \leq x) - F(x)\right| \leq \frac{C\rho}{\sigma^2\sqrt{n}}$$

where $F(x)$ is the cumulative distribution for a normal distribution with mean zero and standard deviation σ and C is some constant.

Exercises

B.1 **Markov and Chebyshev inequalities:** The goal of this exercise is to prove Lemma B.3.

(a) Let $\mathbf{X}$ be a random variable taking values in $\mathbb{R}_{\geq 0}$ and $x \geq 0$. Show that $\mathbf{X} \geq x\mathbb{1}_{\{\mathbf{X}\geq x\}}$, where $\mathbb{1}_{\{\mathbf{X}\geq x\}}$ is the random variable defined by

$$\mathbb{1}_{\{\mathbf{X}\geq x\}} = \begin{cases} 1 & \text{if } \mathbf{X} \geq x, \\ 0 & \text{if } \mathbf{X} < x, \end{cases}$$

and use this to deduce Markov's inequality.

(b) Apply Markov's inequality to the random variable $\mathbf{Y} = (\mathbf{X} - \mathbb{E}\mathbf{X})^2$ to prove Chebyshev's inequality.

B.2 **Law of large numbers:** Use Chebyshev's inequality to prove Theorem B.4.