Lecture Notes for Physics 229: Quantum Information and Computation

John Preskill
California Institute of Technology

September, 1998

Contents

Chapter 1

Introduction and Overview

The course has a website at

`http://www.theory.caltech.edu/~preskill/ph229`

General information can be found there, including a course outline and links to relevant references.

Our topic can be approached from a variety of points of view, but these lectures will adopt the perspective of a theoretical physicist (that is, it's my perspective and I'm a theoretical physicist). Because of the interdisciplinary character of the subject, I realize that the students will have a broad spectrum of backgrounds, and I will try to allow for that in the lectures. Please give me feedback if I am assuming things that you don't know.

1.1 Physics of information

Why is a physicist teaching a course about information? In fact, the *physics of information and computation* has been a recognized discipline for at least several decades. This is natural. Information, after all, is something that is encoded in the state of a physical system; a computation is something that can be carried out on an actual physically realizable device. So the study of information and computation should be linked to the study of the underlying physical processes. Certainly, from an engineering perspective, mastery of principles of physics and materials science is needed to develop state-of-the-art computing hardware. (Carver Mead calls his Caltech research group, dedicated to advancing the art of chip design, the "Physics of Computation" (Physcmp) group).

From a more abstract theoretical perspective, there have been noteworthy milestones in our understanding of how physics constrains our ability to use and manipulate information. For example:

• **Landauer's principle**. Rolf Landauer pointed out in 1961 that erasure of information is necessarily a *dissipative* process. His insight is that erasure always involves the compression of phase space, and so is irreversible.

For example, I can store one bit of information by placing a single molecule in a box, either on the left side or the right side of a partition that divides the box. Erasure means that we move the molecule to the left side (say) irrespective of whether it started out on the left or right. I can suddenly remove the partition, and then slowly compress the one-molecule "gas" with a piston until the molecule is definitely on the left side. This procedure reduces the entropy of the gas by $\Delta S = k \ln 2$ and there is an associated flow of heat from the box to the environment. If the process is isothermal at temperature T, then work $W = kT \ln 2$ is performed on the box, work that I have to provide. If I am to erase information, someone will have to pay the power bill.

• **Reversible computation**. The logic gates used to perform computation are typically *irreversible*, e.g., the NAND gate

$$(a, b) \to \neg(a \wedge b) \tag{1.1}$$

has two input bits and one output bit, and we can't recover a unique input from the output bit. According to Landauer's principle, since about one bit is erased by the gate (averaged over its possible inputs), at least work $W = kT \ln 2$ is needed to operate the gate. If we have a finite supply of batteries, there appears to be a theoretical limit to how long a computation we can perform.

But Charles Bennett found in 1973 that any computation can be performed using only reversible steps, and so in principle requires no dissipation and no power expenditure. We can actually construct a reversible version of the NAND gate that preserves all the information about the input: For example, the (Toffoli) gate

$$(a, b, c) \to (a, b, c \oplus a \wedge b) \tag{1.2}$$

is a reversible 3-bit gate that flips the third bit if the first two both take the value 1 and does nothing otherwise. The third output bit becomes the NAND of a and b if $c = 1$. We can transform an irreversible computation to a reversible one by replacing the NAND gates by Toffoli gates. This computation could in principle be done with negligible dissipation.

However, in the process we generate a lot of extra junk, and one wonders whether we have only postponed the energy cost; we'll have to pay when we need to erase all the junk. Bennett addressed this issue by pointing out that a reversible computer can run forward to the end of a computation, print out a copy of the answer (a logically reversible operation) and then *reverse* all of its steps to return to its initial configuration. This procedure removes the junk without any energy cost.

In principle, then, we need not pay any power bill to compute. In practice, the (irreversible) computers in use today dissipate orders of magnitude more than $kT\ln 2$ per gate, anyway, so Landauer's limit is not an important engineering consideration. But as computing hardware continues to shrink in size, it may become important to beat Landauer's limit to prevent the components from melting, and then reversible computation may be the only option.

• **Maxwell's demon**. The insights of Landauer and Bennett led Bennett in 1982 to the reconciliation of Maxwell's demon with the second law of thermodynamics. Maxwell had envisioned a gas in a box, divided by a partition into two parts A and B. The partition contains a shutter operated by the demon. The demon observes the molecules in the box as they approach the shutter, allowing fast ones to pass from A to B, and slow ones from B to A. Hence, A cools and B heats up, with a negligible expenditure of work. Heat flows from a cold place to a hot place at no cost, in apparent violation of the second law.

The resolution is that the demon must collect and store information about the molecules. If the demon has a finite memory capacity, he cannot continue to cool the gas indefinitely; eventually, information must be erased. At that point, we finally pay the power bill for the cooling we achieved. (If the demon does not erase his record, or if we want to do the thermodynamic accounting before the erasure, then we should associate some entropy with the recorded information.)

These insights were largely anticipated by Leo Szilard in 1929; he was truly a pioneer of the physics of information. Szilard, in *his* analysis of the Maxwell demon, invented the concept of a *bit* of information, (the *name* "bit" was introduced later, by Tukey) and associated the entropy $\Delta S = k\ln 2$ with the acquisition of one bit (though Szilard does not seem to have fully grasped Landauer's principle, that it is the *erasure* of the bit that carries an inevitable

cost).

These examples illustrate that work at the interface of physics and information has generated noteworthy results of interest to both physicists and computer scientists.

1.2 Quantum information

The moral we draw is that "information is physical." and it is instructive to consider what physics has to tell us about information. But fundamentally, the universe is quantum mechanical. How does quantum theory shed light on the nature of information?

It must have been clear already in the early days of quantum theory that classical ideas about information would need revision under the new physics. For example, the clicks registered in a detector that monitors a radioactive source are described by a *truly random* Poisson process. In contrast, there is no place for true randomness in deterministic classical dynamics (although of course a complex (chaotic) classical system can exhibit behavior that is in practice indistinguishable from random).

Furthermore, in quantum theory, noncommuting observables cannot simultaneously have precisely defined values (the uncertainty principle), and in fact performing a measurement of one observable A will necessarily influence the outcome of a subsequent measurement of an observable B, if A and B do not commute. Hence, the act of acquiring information about a physical system inevitably disturbs the state of the system. There is no counterpart of this limitation in classical physics.

The tradeoff between acquiring information and creating a disturbance is related to quantum randomness. It is because the outcome of a measurement has a random element that we are unable to infer the initial state of the system from the measurement outcome.

That acquiring information causes a disturbance is also connected with another essential distinction between quantum and classical information: quantum information cannot be copied with perfect fidelity (the no-cloning principle annunciated by Wootters and Zurek and by Dieks in 1982). If we *could* make a perfect copy of a quantum state, we could measure an observable of the copy without disturbing the original and we could defeat the principle of disturbance. On the other hand, nothing prevents us from copying classical information perfectly (a welcome feature when you need to back up your hard disk).

These properties of quantum information are important, but the really deep way in which quantum information differs from classical information emerged from the work of John Bell (1964), who showed that the predictions of quantum mechanics cannot be reproduced by any local hidden variable theory. Bell showed that quantum information can be (in fact, typically is) encoded in nonlocal correlations between the different parts of a physical system, correlations with no classical counterpart. We will discuss Bell's theorem in detail later on, and I will also return to it later in this lecture.

The study of quantum information as a coherent discipline began to emerge in the 1980's, and it has blossomed in the 1990's. Many of the central results of classical information theory have quantum analogs that have been discovered and developed recently, and we will discuss some of these developments later in the course, including: compression of quantum information, bounds on classical information encoded in quantum systems, bounds on quantum information sent reliably over a noisy quantum channel.

1.3 Efficient quantum algorithms

Given that quantum information has many unusual properties, it might have been expected that quantum theory would have a profound impact on our understanding of computation. That this is spectacularly true came to many of us as a bolt from the blue unleashed by Peter Shor (an AT&T computer scientist and a former Caltech undergraduate) in April, 1994. Shor demonstrated that, at least in principle, a quantum computer can *factor* a large number efficiently.

Factoring (finding the prime factors of a composite number) is an example of an *intractable* problem with the property:

— The solution can be *easily verified*, once found.

— But the solution is *hard* to find.

That is, if p and q are large prime numbers, the product $n = pq$ can be computed quickly (the number of elementary bit operations required is about $\log_2 p \cdot \log_2 q$). But given n, it is *hard* to find p and q.

The time required to find the factors is strongly believed (though this has never been proved) to be *superpolynomial* in $\log(n)$. That is, as n increases, the time needed in the worst case grows faster than any power of $\log(n)$. The best known factoring algorithm (the "number field sieve") requires

$$\text{time} \simeq \exp[c(\ln n)^{1/3}(\ln \ln n)^{2/3}] \tag{1.3}$$

where $c = (64/9)^{1/3} \sim 1.9$. The current state of the art is that the 65 digit factors of a 130 digit number can be found in the order of one month by a network of hundreds of work stations. Using this to estimate the prefactor in Eq. 1.3, we can estimate that factoring a 400 digit number would take about 10^{10} years, the age of the universe. So even with vast improvements in technology, factoring a 400 digit number will be out of reach for a while.

The factoring problem is interesting from the perspective of complexity theory, as an example of a problem presumed to be intractable; that is, a problem that can't be solved in a time bounded by a polynomial in the size of the input, in this case $\log n$. But it is also of practical importance, because the difficulty of factoring is the basis of schemes for public key cryptography, such as the widely used RSA scheme.

The exciting new result that Shor found is that a quantum computer can factor in polynomial time, *e.g.*, in time $O[(\ln n)^3]$. So if we had a quantum computer that could factor a 130 digit number in one month (of course we don't, at least not yet!), running Shor's algorithm it could factor that 400 digit number in less than 3 years. The harder the problem, the greater the advantage enjoyed by the quantum computer.

Shor's result spurred my own interest in quantum information (were it not for Shor, I don't suppose I would be teaching this course). It's fascinating to contemplate the implications for complexity theory, for quantum theory, for technology.

1.4 Quantum complexity

Of course, Shor's work had important antecedents. That a quantum system can perform a computation was first explicitly pointed out by Paul Benioff and Richard Feynman (independently) in 1982. In a way, this was a natural issue to wonder about in view of the relentless trend toward miniaturization in microcircuitry. If the trend continues, we will eventually approach the regime where quantum theory is highly relevant to how computing devices function. Perhaps this consideration provided some of the motivation behind Benioff's work. But Feynman's primary motivation was quite different and very interesting. To understand Feynman's viewpoint, we'll need to be more explicit about the mathematical description of quantum information and computation.

The indivisible unit of classical information is the bit: an object that can take either one of two values: 0 or 1. The corresponding unit of quantum information is the quantum bit or *qubit*. The qubit is a vector in a two-dimensional complex vector space with inner product; in deference to the classical bit we can call the elements of an orthonormal basis in this space $|0\rangle$ and $|1\rangle$. Then a normalized vector can be represented

$$|\psi\rangle = a|0\rangle + b|1\rangle, \quad |a|^2 + |b|^2 = 1. \tag{1.4}$$

where $a, b \in \mathbf{C}$. We can perform a measurement that projects $|\psi\rangle$ onto the basis $|0\rangle, |1\rangle$. The outcome of the measurement is not deterministic — the probability that we obtain the result $|0\rangle$ is $|a|^2$ and the probability that we obtain the result $|1\rangle$ is $|b|^2$.

The quantum state of N qubits can be expressed as a vector in a space of dimension 2^N. We can choose as an orthonormal basis for this space the states in which each qubit has a definite value, either $|0\rangle$ or $|1\rangle$. These can be labeled by binary strings such as

$$|01110010\cdots 1001\rangle \tag{1.5}$$

A general normalized vector can be expanded in this basis as

$$\sum_{x=0}^{2^N-1} a_x|x\rangle \ , \tag{1.6}$$

where we have associated with each string the number that it represents in binary notation, ranging in value from 0 to 2^N-1. Here the a_x's are complex numbers satisfying $\sum_x |a_x|^2 = 1$. If we measure all N qubits by projecting each onto the $\{|0\rangle, |1\rangle\}$ basis, the probability of obtaining the outcome $|x\rangle$ is $|a_x|^2$.

Now, a quantum computation can be described this way. We assemble N qubits, and prepare them in a standard initial state such as $|0\rangle|0\rangle\cdots|0\rangle$, or $|x=0\rangle$. We then apply a unitary transformation U to the N qubits. (The transformation U is constructed as a product of standard *quantum gates*, unitary transformations that act on just a few qubits at a time). After U is applied, we measure all of the qubits by projecting onto the $\{|0\rangle, |1\rangle\}$ basis. The measurement outcome is the output of the computation. So the final

output is classical information that can be printed out on a piece of paper, and published in Physical Review.

Notice that the algorithm performed by the quantum computer is a *probabilistic* algorithm. That is, we could run exactly the same program twice and obtain different results, because of the randomness of the quantum measurement process. The quantum algorithm actually generates a probability distribution of possible outputs. (In fact, Shor's factoring algorithm is not guaranteed to succeed in finding the prime factors; it just succeeds with a reasonable probability. That's okay, though, because it is easy to verify whether the factors are correct.)

It should be clear from this description that a quantum computer, though it may operate according to different physical principles than a classical computer, cannot do anything that a classical computer can't do. Classical computers can store vectors, rotate vectors, and can model the quantum measurement process by projecting a vector onto mutually orthogonal axes. So a classical computer can surely *simulate* a quantum computer to arbitrarily good accuracy. Our notion of what is *computable* will be the same, whether we use a classical computer or a quantum computer.

But we should also consider how long the simulation will take. Suppose we have a computer that operates on a modest number of qubits, like $N = 100$. Then to represent the typical quantum state of the computer, we would need to write down $2^N = 2^{100} \sim 10^{30}$ complex numbers! No existing or foreseeable digital computer will be able to do that. And performing a general rotation of a vector in a space of dimension 10^{30} is far beyond the computational capacity of any foreseeable classical computer.

(Of course, N classical bits can take 2^N possible values. But for each one of these, it is very easy to write down a complete description of the configuration — a binary string of length N. Quantum information is very different in that writing down a complete description of just one typical configuration of N qubits is enormously complex.)

So it is true that a classical computer can simulate a quantum computer, but the simulation becomes extremely inefficient as the number of qubits N increases. Quantum mechanics is *hard* (computationally) because we must deal with huge matrices – there is too much room in Hilbert space. This observation led Feynman to speculate that a quantum computer would be able to perform certain tasks that are beyond the reach of any conceivable classical computer. (The quantum computer has no trouble simulating *itself*!) Shor's result seems to bolster this view.

Is this conclusion unavoidable? In the end, our simulation should provide a means of assigning probabilities to all the possible outcomes of the final measurement. It is not really necessary, then, for the classical simulation to track the complete description of the N-qubit quantum state. We would settle for a *probabilistic classical algorithm*, in which the outcome is not uniquely determined by the input, but in which various outcomes arise with a probability distribution that coincides with that generated by the quantum computation. We might hope to perform a *local* simulation, in which each qubit has a definite value at each time step, and each quantum gate can act on the qubits in various possible ways, one of which is selected as determined by a (pseudo)-random number generator. This simulation would be much easier than following the evolution of a vector in an exponentially large space.

But the conclusion of John Bell's powerful theorem is *precisely* that this simulation could never work: there is no *local probabilistic algorithm* that can reproduce the conclusions of quantum mechanics. Thus, while there is no known proof, it seems highly likely that simulating a quantum computer is a very hard problem for any classical computer.

To understand better why the mathematical description of quantum information is necessarily so complex, imagine we have a $3N$-qubit quantum system ($N \gg 1$) divided into three subsystems of N qubits each (called subsystems (1),(2), and (3)). We randomly choose a quantum state of the $3N$ qubits, and then we separate the 3 subsystems, sending (1) to Santa Barbara and (3) to San Diego, while (2) remains in Pasadena. Now we would like to make some measurements to find out as much as we can about the quantum state. To make it easy on ourselves, let's imagine that we have a zillion copies of the state of the system so that we can measure any and all the observables we want.[1] Except for one proviso: we are restricted to carrying out each measurement within one of the subsystems — no collective measurements spanning the boundaries between the subsystems are allowed. Then for a *typical* state of the $3N$-qubit system, our measurements will reveal almost *nothing* about what the state is. Nearly all the information that distinguishes one state from another is in the *nonlocal correlations* between measurement outcomes in subsystem (1) (2), and (3). These are the nonlocal correlations that Bell found to be an essential part of the physical description.

[1]We cannot make copies of an unknown quantum state ourselves, but we can ask a friend to prepare many identical copies of the state (he can do it because he knows what the state is), and not tell us what he did.

We'll see that information content can be quantified by entropy (large entropy means little information.) If we choose a state for the $3N$ qubits randomly, we almost always find that the entropy of each subsystem is very close to

$$S \cong N - 2^{-(N+1)}, \tag{1.7}$$

a result found by Don Page. Here N is the maximum possible value of the entropy, corresponding to the case in which the subsystem carries no accessible information at all. Thus, for large N we can access only an exponentially small amount of information by looking at each subsystem separately.

That is, the measurements reveal very little information if we don't consider how measurement results obtained in San Diego, Pasadena, and Santa Barbara are correlated with one another — in the language I am using, a measurement of a correlation is considered to be a "collective" measurement (even though it could actually be performed by experimenters who observe the separate parts of the same copy of the state, and then exchange phone calls to compare their results). By measuring the correlations we can learn much more; in principle, we can completely reconstruct the state.

Any satisfactory description of the state of the $3N$ qubits must characterize these nonlocal correlations, which are exceedingly complex. This is why a classical simulation of a large quantum system requires vast resources. (When such nonlocal correlations exist among the parts of a system, we say that the parts are "entangled," meaning that we can't fully decipher the state of the system by dividing the system up and studying the separate parts.)

1.5 Quantum parallelism

Feynman's idea was put in a more concrete form by David Deutsch in 1985. Deutsch emphasized that a quantum computer can best realize its computational potential by invoking what he called "quantum parallelism." To understand what this means, it is best to consider an example.

Following Deutsch, imagine we have a black box that computes a function that takes a single bit x to a single bit $f(x)$. We don't know what is happening inside the box, but it must be something complicated, because the computation takes 24 hours. There are four possible functions $f(x)$ (because each of $f(0)$ and $f(1)$ can take either one of two possible values) and we'd like to know what the box is computing. It would take 48 hours to find out both $f(0)$ and $f(1)$.

But we don't have that much time; we need the answer in 24 hours, not 48. And it turns out that we would be satisfied to know whether $f(x)$ is *constant* ($f(0) = f(1)$) or balanced ($f(0) \neq f(1)$). Even so, it takes 48 hours to get the answer.

Now suppose we have a quantum black box that computes $f(x)$. Of course $f(x)$ might not be invertible, while the action of our quantum computer is unitary and must be invertible, so we'll need a transformation U_f that takes two qubits to two:

$$U_f : |x\rangle|y\rangle \to |x\rangle|y \oplus f(x)\rangle \ . \tag{1.8}$$

(This machine flips the second qubit if f acting on the first qubit is 1, and doesn't do anything if f acting on the first qubit is 0.) We can determine if $f(x)$ is constant or balanced by using the quantum black box twice. But it still takes a day for it to produce one output, so that won't do. Can we get the answer (in 24 hours) by running the quantum black box *just once*. (This is "Deutsch's problem.")

Because the black box is a quantum computer, we can choose the input state to be a *superposition* of $|0\rangle$ and $|1\rangle$. If the second qubit is initially prepared in the state $\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$, then

$$\begin{aligned} U_f : |x\rangle\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \ &\to \ |x\rangle\frac{1}{\sqrt{2}}(|f(x)\rangle - |1 \oplus f(x)\rangle) \\ &= \ |x\rangle(-1)^{f(x)}\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle), \end{aligned} \tag{1.9}$$

so we have isolated the function f in an x-dependent phase. Now suppose we prepare the first qubit as $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Then the black box acts as

$$\begin{aligned} U_f : \quad & \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \to \\ & \frac{1}{\sqrt{2}}\left[(-1)^{f(0)}|0\rangle + (-1)^{f(1)}|1\rangle\right]\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \ . \end{aligned} \tag{1.10}$$

Finally, we can perform a measurement that projects the first qubit onto the basis

$$|\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle). \tag{1.11}$$

Evidently, we will always obtain $|+\rangle$ if the function is balanced, and $|-\rangle$ if the function is constant.[2]

So we have solved Deutsch's problem, and we have found a separation between what a classical computer and a quantum computer can achieve. The classical computer has to run the black box twice to distinguish a balanced function from a constant function, but a quantum computer does the job in one go!

This is possible because the quantum computer is not limited to computing either $f(0)$ or $f(1)$. It can act on a superposition of $|0\rangle$ and $|1\rangle$, and thereby extract "global" information about the function, information that depends on both $f(0)$ and $f(1)$. This is quantum parallelism.

Now suppose we are interested in global properties of a function that acts on N bits, a function with 2^N possible arguments. To compute a complete table of values of $f(x)$, we would have to calculate f 2^N times, completely infeasible for $N \gg 1$ (*e.g.*, 10^{30} times for $N = 100$). But with a quantum computer that acts according to

$$U_f : |x\rangle|0\rangle \to |x\rangle|f(x)\rangle \ , \tag{1.12}$$

we could choose the input register to be in a state

$$\left[\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\right]^N = \frac{1}{2^{N/2}}\sum_{x=0}^{2^N-1}|x\rangle \ , \tag{1.13}$$

and by computing $f(x)$ only once, we can generate a state

$$\frac{1}{2^{N/2}}\sum_{x=0}^{2^N-1}|x\rangle|f(x)\rangle \ . \tag{1.14}$$

Global properties of f are encoded in this state, and we might be able to extract some of those properties if we can only think of an efficient way to do it.

This quantum computation exhibits "*massive* quantum parallelism;" a simulation of the preparation of this state on a classical computer would

[2] In our earlier description of a quantum computation, we stated that the final measurement would project each qubit onto the $\{|0\rangle, |1\rangle\}$ basis, but here we are allowing measurement in a different basis. To describe the procedure in the earlier framework, we would apply an appropriate unitary change of basis to each qubit before performing the final measurement.

require us to compute f an unimaginably large number of times (for $N \gg 1$). Yet we have done it with the quantum computer in only one go. It is just this kind of massive parallelism that Shor invokes in his factoring algorithm.

As noted earlier, a characteristic feature of quantum information is that it can be encoded in nonlocal correlations among different parts of a physical system. Indeed, this is the case in Eq. (1.14); the properties of the function f are stored as correlations between the "input register" and "output register" of our quantum computer. This nonlocal information, however, is not so easy to decipher.

If, for example, I were to measure the input register, I would obtain a result $|x_0\rangle$, where x_0 is chosen completely at random from the 2^N possible values. This procedure would prepare a state

$$|x_0\rangle|f(x_0)\rangle. \tag{1.15}$$

We could proceed to measure the output register to find the value of $f(x_0)$. But because Eq. (1.14) has been destroyed by the measurement, the intricate correlations among the registers have been lost, and we get no opportunity to determine $f(y_0)$ for any $y_0 \neq x_0$ by making further measurements. In this case, then, the quantum computation provided no advantage over a classical one.

The lesson of the solution to Deutsch's problem is that we can sometimes be more clever in exploiting the correlations encoded in Eq. (1.14). Much of the art of designing quantum algorithms involves finding ways to make efficient use of the nonlocal correlations.

1.6 A new classification of complexity

The computer on your desktop is not a quantum computer, but still it is a remarkable device: in principle, it is capable of performing any conceivable computation. In practice there are computations that you can't do — you either run out of time or you run out of memory. But if you provide an unlimited amount of memory, and you are willing to wait as long as it takes, then anything that deserves to be called a computation can be done by your little PC. We say, therefore, that it is a "universal computer."

Classical complexity theory is the study of which problems are hard and which ones are easy. Usually, "hard" and "easy" are defined in terms of how much time and/or memory are needed. But how can we make meaningful

distinctions between hard and easy without specifying the hardware we will be using? A problem might be hard on the PC, but perhaps I could design a special purpose machine that could solve that problem much faster. Or maybe in the future a much better general purpose computer will be available that solves the problem far more efficiently. Truly meaningful distinctions between hard and easy should be *universal* — they ought not to depend on which machine we are using.

Much of complexity theory focuses on the distinction between "polynomial time" and "exponential time" algorithms. For any algorithm A, which can act on an input of variable length, we may associate a *complexity function* $T_A(N)$, where N is the length of the input in bits. $T_A(N)$ is the longest "time" (that is, number of elementary steps) it takes for the algorithm to run to completion, for any N-bit input. (For example, if A is a factoring algorithm, $T_A(N)$ is the time needed to factor an N-bit number in the worst possible case.) We say that A is polynomial time if

$$T_A(N) \leq \text{ Poly } (N), \tag{1.16}$$

where Poly (N) denotes a polynomial of N. Hence, polynomial time means that the time needed to solve the problem does not grow faster than a power of the number of input bits.

If the problem is not polynomial time, we say it is exponential time (though this is really a misnomer, because of course that are superpolynomial functions like $N^{\log N}$ that actually increase much more slowly than an exponential). This is a reasonable way to draw the line between easy and hard. But the truly compelling reason to make the distinction this way is that it is machine-independent: it does not matter what computer we are using. The universality of the distinction between polynomial and exponential follows from one of the central results of computer science: one universal (classical) computer can simulate another with at worst "polynomial overhead." This means that if an algorithm runs on your computer in polynomial time, then I can always run it on my computer in polynomial time. If I can't think of a better way to do it, I can always have my computer emulate how yours operates; the cost of running the emulation is only polynomial time. Similarly, your computer can emulate mine, so we will always agree on which algorithms are polynomial time.[3]

[3]To make this statement precise, we need to be a little careful. For example, we should exclude certain kinds of "unreasonable" machines, like a parallel computer with an unlimited number of nodes.

Now it is true that information and computation in the physical world are fundamentally quantum mechanical, but this insight, however dear to physicists, would not be of great interest (at least from the viewpoint of complexity theory) were it possible to simulate a quantum computer on a classical computer with polynomial overhead. Quantum algorithms might prove to be of technological interest, but perhaps no more so than future advances in classical algorithms that might speed up the solution of certain problems.

But if, as is indicated (but not proved!) by Shor's algorithm, no polynomial-time simulation of a quantum computer is possible, that changes everything. Thirty years of work on complexity theory will still stand as mathematical truth, as theorems characterizing the capabilities of classical universal computers. But it may fall as physical truth, because a classical Turing machine is not an appropriate model of the computations that can really be performed in the physical world.

If the quantum classification of complexity is indeed different than the classical classification (as is suspected but not proved), then this result will shake the foundations of computer science. In the long term, it may also strongly impact technology. But what is its significance for physics?

I'm not sure. But perhaps it is telling that no conceivable classical computation can accurately predict the behavior of even a modest number of qubits (of order 100). This may suggest that relatively small quantum systems have greater potential than we suspected to surprise, baffle, and delight us.

1.7 What about errors?

As significant as Shor's factoring algorithm may prove to be, there is another recently discovered feature of quantum information that may be just as important: the discovery of quantum error correction. Indeed, were it not for this development, the prospects for quantum computing technology would not seem bright.

As we have noted, the essential property of quantum information that a quantum computer exploits is the existence of nonlocal correlations among the different parts of a physical system. If I look at only part of the system at a time, I can decipher only very little of the information encoded in the system.

Unfortunately, these nonlocal correlations are extremely fragile and tend to decay very rapidly in practice. The problem is that our quantum system is inevitably in contact with a much larger system, its environment. It is virtually impossible to perfectly isolate a big quantum system from its environment, even if we make a heroic effort to do so. Interactions between a quantum device and its environment establish nonlocal correlations between the two. Eventually the quantum information that we initially encoded in the device becomes encoded, instead, in correlations between the device and the environment. At that stage, we can no longer access the information by observing only the device. In practice, the information is irrevocably lost. Even if the coupling between device and environment is quite weak, this happens to a macroscopic device remarkably quickly.

Erwin Schrödinger chided the proponents of the mainstream interpretation of quantum mechanics by observing that the theory will allow a quantum state of a cat of the form

$$|\mathrm{cat}\rangle = \frac{1}{\sqrt{2}}\left(|\mathrm{dead}\rangle + |\mathrm{alive}\rangle\right). \tag{1.17}$$

To Schrödinger, the possibility of such states was a blemish on the theory, because every cat he had seen was either dead or alive, not half dead and half alive.

One of the most important advances in quantum theory over the past 15 years is that we have learned how to answer Schrödinger with growing confidence. The state $|\mathrm{cat}\rangle$ is possible in principle, but is rarely seen because it is *extremely* unstable. The cats Schrödinger observed were never well isolated from the environment. If someone were to prepare the state $|\mathrm{cat}\rangle$, the quantum information encoded in the superposition of $|\mathrm{dead}\rangle$ and $|\mathrm{alive}\rangle$ would *immediately* be transferred to correlations between the cat and the environment, and become completely inaccessible. In effect, the environment continually measures the cat, projecting it onto either the state $|\mathrm{alive}\rangle$ or $|\mathrm{dead}\rangle$. This process is called *decoherence*. We will return to the study of decoherence later in the course.

Now, to perform a complex quantum computation, we need to prepare a delicate superposition of states of a relatively large quantum system (though perhaps not as large as a cat). Unfortunately, this system cannot be perfectly isolated from the environment, so this superposition, like the state $|\mathrm{cat}\rangle$, decays very rapidly. The encoded quantum information is quickly lost, and our quantum computer crashes.

To put it another way, contact between the computer and the environment (decoherence) causes *errors* that degrade the quantum information. To operate a quantum computer reliably, we must find some way to prevent or correct these errors.

Actually, decoherence is not our only problem. Even if we could achieve perfect isolation from the environment, we could not expect to operate a quantum computer with perfect accuracy. The quantum gates that the machine executes are unitary transformations that operate on a few qubits at a time, let's say 4×4 unitary matrices acting on two qubits. Of course, these unitary matrices form a continuum. We may have a protocol for applying U_0 to 2 qubits, but our execution of the protocol will not be flawless, so the actual transformation

$$U = U_0 \left(1 + O(\varepsilon)\right) \tag{1.18}$$

will differ from the intended U_0 by some amount of order ε. After about $1/\varepsilon$ gates are applied, these errors will accumulate and induce a serious failure. Classical analog devices suffer from a similar problem, but small errors are much less of a problem for devices that perform discrete logic.

In fact, modern digital circuits are remarkably reliable. They achieve such high accuracy with help from *dissipation*. We can envision a classical gate that acts on a bit, encoded as a ball residing at one of the two minima of a double-lobed potential. The gate may push the ball over the intervening barrier to the other side of the potential. Of course, the gate won't be implemented perfectly; it may push the ball a little too hard. Over time, these imperfections might accumulate, causing an error.

To improve the performance, we *cool* the bit (in effect) after each gate. This is a dissipative process that releases heat to the environment and compresses the phase space of the ball, bringing it close to the local minimum of the potential. So the small errors that we may make wind up heating the environment rather than compromising the performance of the device.

But we can't cool a quantum computer this way. Contact with the environment may enhance the reliability of classical information, but it would destroy encoded quantum information. More generally, accumulation of error will be a problem for classical reversible computation as well. To prevent errors from building up we need to discard the information about the errors, and throwing away information is always a dissipative process.

Still, let's not give up too easily. A sophisticated machinery has been developed to contend with errors in classical information, the theory of er-

ror correcting codes. To what extent can we coopt this wisdom to protect quantum information as well?

How does classical error correction work? The simplest example of a classical error-correcting code is a repetition code: we replace the bit we wish to protect by 3 copies of the bit,

$$\begin{aligned} 0 &\to (000), \\ 1 &\to (111). \end{aligned} \tag{1.19}$$

Now an error may occur that causes one of the three bits to flip; if it's the first bit, say,

$$\begin{aligned} (000) &\to (100), \\ (111) &\to (011). \end{aligned} \tag{1.20}$$

Now in spite of the error, we can still decode the bit correctly, by majority voting.

Of course, if the probability of error in each bit were p, it would be possible for two of the three bits to flip, or even for all three to flip. A double flip can happen in three different ways, so the probability of a double flip is $3p^2(1-p)$, while the probability of a triple flip is p^3. Altogether, then, the probability that majority voting fails is $3p^2(1-p) + p^3 = 3p^2 - 2p^3$. But for

$$3p^2 - 2p^3 < p \quad \text{or} \quad p < \frac{1}{2}, \tag{1.21}$$

the code improves the reliability of the information.

We can improve the reliability further by using a longer code. One such code (though far from the most efficient) is an N-bit repetition code. The probability distribution for the average value of the bit, by the central limit theorem, approaches a Gaussian with width $1/\sqrt{N}$ as $N \to \infty$. If $P = \frac{1}{2} + \varepsilon$ is the probability that each bit has the correct value, then the probability that the majority vote fails (for large N) is

$$P_{error} \sim e^{-N\varepsilon^2}, \tag{1.22}$$

arising from the tail of the Gaussian. Thus, for any $\varepsilon > 0$, by introducing enough redundancy we can achieve arbitrarily good reliability. Even for $\varepsilon < 0$, we'll be okay if we always assume that majority voting gives the *wrong* result. Only for $P = \frac{1}{2}$ is the cause lost, for then our block of N bits will be *random*, and encode no information.

In the 50's, John Von Neumann showed that a classical computer with noisy components can work reliably, by employing sufficient redundancy. He pointed out that, if necessary, we can compute each logic gate many times, and accept the majority result. (Von Neumann was especially interested in how his brain was able to function so well, in spite of the unreliability of neurons. He was pleased to explain why he was so smart.)

But now we want to use error correction to keep a *quantum computer* on track, and we can immediately see that there are difficulties:

1. **Phase errors**. With quantum information, more things can go wrong. In addition to bit-flip errors
$$\begin{aligned} |0\rangle &\rightarrow |1\rangle, \\ |1\rangle &\rightarrow |0\rangle. \end{aligned} \tag{1.23}$$
there can also be *phase* errors
$$\begin{aligned} |0\rangle &\rightarrow |0\rangle, \\ |1\rangle &\rightarrow -|1\rangle. \end{aligned} \tag{1.24}$$
A phase error is serious, because it makes the state $\frac{1}{\sqrt{2}}[|0\rangle + |1\rangle]$ flip to the orthogonal state $\frac{1}{\sqrt{2}}[|0\rangle - |1\rangle]$. But the classical coding provided no protection against phase errors.

2. **Small errors**. As already noted, quantum information is continuous. If a qubit is intended to be in the state
$$a|0\rangle + b|1\rangle, \tag{1.25}$$
an error might change a and b by an amount of order ε, and these small errors can accumulate over time. The classical method is designed to correct large (bit flip) errors.

3. **Measurement causes disturbance**. In the majority voting scheme, it seemed that we needed to *measure* the bits in the code to detect and correct the errors. But we can't measure qubits without *disturbing* the quantum information that they encode.

4. **No cloning**. With classical coding, we protected information by making extra copies of it. But we know that quantum information cannot be copied with perfect fidelity.

1.8 Quantum error-correcting codes

Despite these obstacles, it turns out that quantum error correction really is possible. The first example of a quantum error-correcting code was constructed about two years ago by (guess who!) Peter Shor. This discovery ushered in a new discipline that has matured remarkably quickly – the theory of quantum error-correcting codes. We will study this theory later in the course.

Probably the best way to understand how quantum error correction works is to examine Shor's original code. It is the most straightforward quantum generalization of the classical 3-bit repetition code.

Let's look at that 3-bit code one more time, but this time mindful of the requirement that, with a quantum code, we will need to be able to correct the errors without measuring any of the encoded information.

Suppose we encode a single qubit with 3 qubits:
$$\begin{aligned} |0\rangle &\rightarrow |\bar{0}\rangle \equiv |000\rangle, \\ |1\rangle &\rightarrow |\bar{1}\rangle \equiv |111\rangle, \end{aligned} \tag{1.26}$$
or, in other words, we encode a superposition
$$a|0\rangle + b|1\rangle \rightarrow a|\bar{0}\rangle + b|\bar{1}\rangle = a|000\rangle + b|111\rangle \ . \tag{1.27}$$
We would like to be able to correct a bit flip error without destroying this superposition.

Of course, it won't do to measure a single qubit. If I measure the first qubit and get the result $|0\rangle$, then I have prepared the state $|\bar{0}\rangle$ of all three qubits, and we have lost the quantum information encoded in the coefficients a and b.

But there is no need to restrict our attention to single-qubit measurements. I could also perform collective measurements on two-qubits at once, and collective measurements suffice to diagnose a bit-flip error. For a 3-qubit state $|x, y, z\rangle$ I could measure, say, the two-qubit observables $y \oplus z$, or $x \oplus z$ (where $\oplus$ denotes addition modulo 2). For both $|x, y, z\rangle = |000\rangle$ and $|111\rangle$ these would be 0, but if any one bit flips, then at least one of these quantities will be 1. In fact, if there is a single bit flip, the two bits
$$(y \oplus z, x \oplus z), \tag{1.28}$$

just designate in binary notation the position (1,2 or 3) of the bit that flipped. These two bits constitute a *syndrome* that diagnoses the error that occurred.

For example, if the first bit flips,

$$a|000\rangle + b|111\rangle \to a|100\rangle + b|011\rangle, \tag{1.29}$$

then the measurement of $(y \oplus z, x \oplus z)$ yields the result $(0,1)$, which instructs us to flip the first bit; this indeed repairs the error.

Of course, instead of a (large) bit flip there could be a small error:

$$\begin{aligned} |000\rangle &\to |000\rangle + \varepsilon|100\rangle \\ |111\rangle &\to |111\rangle - \varepsilon|011\rangle. \end{aligned} \tag{1.30}$$

But even in this case the above procedure would work fine. In measuring $(y \oplus z, x \oplus z)$, we would project out an eigenstate of this observable. Most of the time (probability $1 - |\varepsilon|^2$) we obtain the result $(0,0)$ and project the damaged state back to the original state, and so correct the error. Occasionally (probability $|\varepsilon|^2$) we obtain the result $(0,1)$ and project the state onto Eq. 1.29. But then the syndrome instructs us to flip the first bit, which restores the original state. Similarly, if there is an amplitude of order ε for each of the three qubits to flip, then with a probability of order $|\varepsilon|^2$ the syndrome measurement will project the state to one in which one of the three bits is flipped, and the syndrome will tell us which one.

So we have already overcome 3 of the 4 obstacles cited earlier. We see that it is possible to make a measurement that diagnoses the error without damaging the information (answering (3)), and that a quantum measurement can project a state with a small error to either a state with no error or a state with a large discrete error that we know how to correct (answering (2)). As for (4), the issue didn't come up, because the state $a|\bar{0}\rangle + b|\bar{1}\rangle$ is not obtained by cloning – it is not the same as $(a|0\rangle + b|1\rangle)^3$; that is, it differs from three copies of the unencoded state.

Only one challenge remains: (1) phase errors. Our code does not yet provide any protection against phase errors, for if any one of the three qubits undergoes a phase error then our encoded state $a|\bar{0}\rangle + b|\bar{1}\rangle$ is transformed to $a|\bar{0}\rangle - b|\bar{1}\rangle$, and the encoded quantum information is damaged. In fact, phase errors have become three times more likely than if we hadn't used the code. But with the methods in hand that conquered problems (2)-(4), we can approach problem (1) with new confidence. Having protected against bit-flip

errors by encoding bits redundantly, we are led to protect against phase-flip errors by encoding phases redundantly.

Following Shor, we encode a single qubit using nine qubits, according to

$$\begin{aligned} |0\rangle \to |\bar{0}\rangle &\equiv \frac{1}{2^{3/2}}\left(|000\rangle + |111\rangle\right)\left(|000\rangle + |111\rangle\right)\left(|000\rangle + |111\rangle\right), \\ |1\rangle \to |\bar{1}\rangle &\equiv \frac{1}{2^{3/2}}\left(|000\rangle - |111\rangle\right)\left(|000\rangle - |111\rangle\right)\left(|000\rangle - |111\rangle\right). \end{aligned} \tag{1.31}$$

Both $|\bar{0}\rangle$ and $|\bar{1}\rangle$ consist of three clusters of three qubits each, with each cluster prepared in the same quantum state. Each of the clusters has triple bit redundancy, so we can correct a single bit flip in any cluster by the method discussed above.

Now suppose that a phase flip occurs in one of the clusters. The error changes the relative sign of $|000\rangle$ and $|111\rangle$ in that cluster so that

$$\begin{aligned} |000\rangle + |111\rangle &\to |000\rangle - |111\rangle, \\ |000\rangle - |111\rangle &\to |000\rangle + |111\rangle. \end{aligned} \tag{1.32}$$

This means that the relative phase of the damaged cluster differs from the phases of the other two clusters. Thus, as in our discussion of bit-flip correction, we can identify the damaged cluster, not by *measuring* the relative phase in each cluster (which would disturb the encoded information) but by *comparing* the phases of pairs of clusters. In this case, we need to measure a six-qubit observable to do the comparison, e.g., the observable that flips qubits 1 through 6. Since flipping twice is the identity, this observable squares to 1, and has eigenvalues ± 1. A pair of clusters with the same sign is an eigenstate with eigenvalue $+1$, and a pair of clusters with opposite sign is an eigenstate with eigenvalue -1. By measuring the six-qubit observable for a second pair of clusters, we can determine which cluster has a different sign than the others. Then, we apply a unitary phase transformation to one of the qubits in that cluster to reverse the sign and correct the error.

Now suppose that a unitary error $U = 1 + 0(\varepsilon)$ occurs for each of the 9 qubits. The most general single-qubit unitary transformation (aside from a physically irrelevant overall phase) can be expanded to order ε as

$$U = 1 + i\varepsilon_x \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + i\varepsilon_y \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} + i\varepsilon_z \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{1.33}$$

the three terms of order ε in the expansion can be interpreted as a bit flip operator, a phase flip operator, and an operator in which both a bit flip and a phase flip occur. If we prepare an encoded state $a|\bar{0}\rangle + b|\bar{1}\rangle$, allow the unitary errors to occur on each qubit, and then measure the bit-flip and phase-flip syndromes, then most of the time we will project the state back to its original form, but with a probability of order $|\varepsilon|^2$, one qubit will have a large error: a bit flip, a phase flip, or both. From the syndrome, we learn which bit flipped, and which cluster had a phase error, so we can apply the suitable one-qubit unitary operator to fix the error.

Error recovery will fail if, after the syndrome measurement, there are two bit flip errors in each of two clusters (which induces a phase error in the encoded data) or if phase errors occur in two different clusters (which induces a bit-flip error in the encoded data). But the probability of such a double phase error is of order $|\varepsilon|^4$. So for $|\varepsilon|$ small enough, coding improves the reliability of the quantum information.

The code also protects against decoherence. By restoring the quantum state irrespective of the nature of the error, our procedure removes any entanglement between the quantum state and the environment.

Here as always, error correction is a dissipative process, since information about the nature of the errors is flushed out of the quantum system. In this case, that information resides in our recorded measurement results, and heat will be dissipated when that record is erased.

Further developments in quantum error correction will be discussed later in the course, including:

• As with classical coding it turns out that there are "good" quantum codes that allow us to achieve arbitrarily high reliability as long as the error rate per qubit is small enough.

• We've assumed that the error recovery procedure is itself executed flawlessly. But the syndrome measurement was complicated – we needed to measure two-qubit and six-qubit collective observables to diagnose the errors – so we actually might further damage the data when we try to correct it. We'll show, though, that error correction can be carried out so that it still works effectively even if we make occasional errors during the recovery process.

• To operate a quantum computer we'll want not only to store quantum information reliably, but also to process it. We'll show that it is possible to apply quantum gates to encoded information.

Let's summarize the essential ideas that underlie our quantum error correction scheme:

1. We *digitized* the errors. Although the errors in the quantum information were small, we performed measurements that projected our state onto either a state with no error, or a state with one of a discrete set of errors that we knew how to convert.

2. We measured the errors without measuring the data. Our measurements revealed the nature of the errors without revealing (and hence disturbing) the encoded information.

3. The errors are local, and the encoded information is nonlocal. It is important to emphasize the central assumption underlying the construction of the code – that errors affecting different qubits are, to a good approximation, *uncorrelated*. We have tacitly assumed that an event that causes errors in two qubits is much less likely than an event causing an error in a single qubit. It is of course a physics question whether this assumption is justified or not – we can easily envision processes that will cause errors in two qubits at once. If such correlated errors are common, coding will fail to improve reliability.

The code takes advantage of the presumed local nature of the errors by encoding the information in a nonlocal way - that is the information is stored in *correlations* involving several qubits. There is no way to distinguish $|\bar{0}\rangle$ and $|\bar{1}\rangle$ by measuring a single qubit of the nine. If we measure one qubit we will find $|0\rangle$ with probability $\frac{1}{2}$ and $|1\rangle$ with probability $\frac{1}{2}$ irrespective of the value of the encoded qubit. To access the encoded information we need to measure a 3-qubit observable (the operator that flips all three qubits in a cluster can distinguish $|000\rangle + |111\rangle$ from $|000\rangle - |111\rangle$).

The environment might occasionally kick one of the qubits, in effect "measuring" it. But the encoded information cannot be damaged by disturbing that one qubit, because a single qubit, by itself, actually carries no information at all. Nonlocally encoded information is invulnerable to local influences – this is the central principle on which quantum error-correcting codes are founded.

1.9 Quantum hardware

The theoretical developments concerning quantum complexity and quantum error correction have been accompanied by a burgeoning experimental effort to process coherent quantum information. I'll briefly describe some of this activity here.

To build hardware for a quantum computer, we'll need technology that enables us to manipulate qubits. The hardware will need to meet some stringent specifications:

1. **Storage**: We'll need to store qubits for a long time, long enough to complete an interesting computation.

2. **Isolation**: The qubits must be well isolated from the environment, to minimize decoherence errors.

3. **Readout**: We'll need to measure the qubits efficiently and reliably.

4. **Gates**: We'll need to manipulate the quantum states of individual qubits, and to induce controlled interactions among qubits, so that we can perform quantum gates.

5. **Precision**: The quantum gates should be implemented with high precision if the device is to perform reliably.

1.9.1 Ion Trap

One possible way to achieve these goals was suggested by Ignacio Cirac and Peter Zoller, and has been pursued by Dave Wineland's group at the National Institute for Standards and Technology (NIST), as well as other groups. In this scheme, each qubit is carried by a single ion held in a linear Paul trap. The quantum state of each ion is a linear combination of the ground state $|g\rangle$ (interpreted as $|0\rangle$) and a particular long-lived metastable excited state $|e\rangle$ (interpreted as $|1\rangle$). A coherent linear combination of the two levels,

$$a|g\rangle + be^{i\omega t}|e\rangle, \tag{1.34}$$

can survive for a time comparable to the lifetime of the excited state (though of course the relative phase oscillates as shown because of the energy splitting $\hbar\omega$ between the levels). The ions are so well isolated that spontaneous decay can be the dominant form of decoherence.

It is easy to read out the ions by performing a measurement that projects onto the $\{|g\rangle, |e\rangle\}$ basis. A laser is tuned to a transition from the state $|g\rangle$ to a short-lived excited state $|e'\rangle$. When the laser illuminates the ions, each

qubit with the value $|0\rangle$ repeatedly absorbs and reemits the laser light, so that it flows visibly (fluoresces). Qubits with the value $|1\rangle$ remain dark.

Because of their mutual Coulomb repulsion, the ions are sufficiently well separated that they can be individually addressed by pulsed lasers. If a laser is tuned to the frequency ω of the transition and is focused on the nth ion, then Rabi oscillations are induced between $|0\rangle$ and $|1\rangle$. By timing the laser pulse properly and choosing the phase of the laser appropriately, we can apply any one-qubit unitary transformation. In particular, acting on $|0\rangle$, the laser pulse can prepare any desired linear combination of $|0\rangle$ and $|1\rangle$.

But the most difficult part of designing and building quantum computing hardware is getting two qubits to interact with one another. In the ion trap, interactions arise because of the Coulomb repulsion between the ions. Because of the mutual Couloumb repulsion, there is a spectrum of coupled normal modes of vibration for the trapped ions. When the ion absorbs or emits a laser photon, the center of mass of the ion recoils. But if the laser is properly tuned, then when a single ion absorbs or emits, a normal mode involving many ions will recoil coherently (the Mössbauer effect).

The vibrational mode of lowest frequency (frequency ν) is the center-of-mass (cm) mode, in which the ions oscillate in lockstep in the harmonic well of the trap. The ions can be laser cooled to a temperature much less than ν, so that each vibrational mode is very likely to occupy its quantum-mechanical ground state. Now imagine that a laser tuned to the frequency $\omega - \nu$ shines on the nth ion. For a properly time pulse the state $|e\rangle_n$ will rotate to $|g\rangle_n$, while the cm oscillator makes a transition from its ground state $|0\rangle_{cm}$ to its first excited state $|1\rangle_{cm}$ (a cm "phonon" is produced). However, the state $|g\rangle_n|0\rangle_{cm}$ is not on resonance for any transition and so is unaffected by the pulse. Thus the laser pulse induces a unitary transformation acting as

$$\begin{aligned} |g\rangle_n|0\rangle_{cm} &\to |g\rangle_n|0\rangle_{cm}, \\ |e\rangle_n|0\rangle_{cm} &\to -i|g\rangle_n|1\rangle_{cm}. \end{aligned} \tag{1.35}$$

This operation removes a bit of information that is initially stored in the internal state of the nth ion, and deposits that bit in the *collective* state of motion of *all* the ions.

This means that the state of motion of the mth ion ($m \neq n$) has been influenced by the internal state of the nth ion. In this sense, we have succeeded in inducing an interaction between the ions. To complete the quantum gate, we should transfer the quantum information from the cm phonon back to the internal state of one of the ions. The procedure should be designed so that the cm mode always returns to its ground state $|0\rangle_{cm}$ at the conclusion of the gate implementation. For example, Cirac and Zoller showed that the quantum XOR (or controlled not) gate

$$|x, y\rangle \to |x, y \oplus x\rangle, \tag{1.36}$$

can be implemented in an ion trap with altogether 5 laser pulses. The conditional excitation of a phonon, Eq. (1.35) has been demonstrated experimentally, for a single trapped ion, by the NIST group.

One big drawback of the ion trap computer is that it is an intrinsically slow device. Its speed is ultimately limited by the energy-time uncertainty relation. Since the uncertainty in the energy of the laser photons should be small compared to the characteristic vibrational splitting ν, each laser pulse should last a time long compared to ν^{-1}. In practice, ν is likely to be of order 100 kHz.

1.9.2 Cavity QED

An alternative hardware design (suggested by Pellizzari, Gardiner, Cirac, and Zoller) is being pursued by Jeff Kimble's group here at Caltech. The idea is to trap several neutral atoms inside a small high finesse optical cavity. Quantum information can again be stored in the internal states of the atoms. But here the atoms interact because they all couple to the normal modes of the electromagnetic field in the cavity (instead of the vibrational modes as in the ion trap). Again, by driving transitions with pulsed lasers, we can induce a transition in one atom that is conditioned on the internal state of another atom.

Another possibility is to store a qubit, not in the internal state of an ion, but in the polarization of a photon. Then a trapped atom can be used as the intermediary that causes one photon to interact with another (instead of a photon being used to couple one atom to another). In their "flying qubit" experiment two years ago. The Kimble group demonstrated the operation of a two-photon quantum gate, in which the circular polarization of one photon

influences the *phase* of another photon:

$$\begin{aligned} |L\rangle_1|L\rangle_2 &\to |L\rangle_1|L\rangle_2 \\ |L\rangle_1|R\rangle_2 &\to |L\rangle_1|R\rangle_2 \\ |R\rangle_1|L\rangle_2 &\to |R\rangle_1|L\rangle_2 \\ |R\rangle_1|R\rangle_2 &\to e^{i\Delta}|R\rangle_1|R\rangle_2 \end{aligned} \tag{1.37}$$

where $|L\rangle, |R\rangle$ denote photon states with left and right circular polarization. To achieve this interaction, one photon is stored in the cavity, where the $|L\rangle$ polarization does not couple to the atom, but the $|R\rangle$ polarization couples strongly. A second photon transverses the cavity, and for the second photon as well, one polarization interacts with the atom preferentially. The second photon wave pocket acquires a particular phase shift $e^{i\Delta}$ only if both photons have $|R\rangle$ polarization. Because the phase shift is conditioned on the polarization of *both* photons, this is a nontrivial two-qubit quantum gate.

1.9.3 NMR

A third (dark horse) hardware scheme has sprung up in the past year, and has leap frogged over the ion trap and cavity QED to take the current lead in coherent quantum processing. The new scheme uses nuclear magnetic resonance (NMR) technology. Now qubits are carried by certain nuclear spins in a particular molecule. Each spin can either be aligned ($|\uparrow\rangle = |0\rangle$) or antialigned ($|\downarrow\rangle = |1\rangle$) with an applied constant magnetic field. The spins take a long time to relax or decohere, so the qubits can be stored for a reasonable time.

We can also turn on a pulsed rotating magnetic field with frequency ω (where the ω is the energy splitting between the spin-up and spin-down states), and induce Rabi oscillations of the spin. By timing the pulse suitably, we can perform a desired unitary transformation on a single spin (just as in our discussion of the ion trap). All the spins in the molecule are exposed to the rotating magnetic field but only those on resonance respond.

Furthermore, the spins have dipole-dipole interactions, and this coupling can be exploited to perform a gate. The splitting between $|\uparrow\rangle$ and $|\downarrow\rangle$ for one spin actually depends on the state of neighboring spins. So whether a driving pulse is on resonance to tip the spin over is conditioned on the state of another spin.

All this has been known to chemists for decades. Yet it was only in the past year that Gershenfeld and Chuang, and independently Cory, Fahmy, and Havel, pointed out that NMR provides a useful implementation of quantum computation. This was not obvious for several reasons. Most importantly, NMR systems are very *hot*. The typical temperature of the spins (room temperature, say) might be of order a million times larger than the energy splitting between $|0\rangle$ and $|1\rangle$. This means that the quantum state of our computer (the spins in a single molecule) is very noisy – it is subject to strong random thermal fluctuations. This noise will disguise the quantum information. Furthermore, we actually perform our processing not on a single molecule, but on a macroscopic sample containing of order 10^{23} "computers," and the signal we read out of this device is actually averaged over this ensemble. But quantum algorithms are *probabilistic*, because of the randomness of quantum measurement. Hence averaging over the ensemble is not equivalent to running the computation on a single device; averaging may obscure the results.

Gershenfeld and Chuang and Cory, Fahmy, and Havel, explained how to overcome these difficulties. They described how "effective pure states" can be prepared, manipulated, and monitored by performing suitable operations on the thermal ensemble. The idea is to arrange for the fluctuating properties of the molecule to average out when the signal is detected, so that only the underlying coherent properties are measured. They also pointed out that some quantum algorithms (including Shor's factoring algorithm) can be cast in a deterministic form (so that at least a large fraction of the computers give the same answer); then averaging over many computations will not spoil the result.

Quite recently, NMR methods have been used to prepare a maximally entangled state of three qubits, which had never been achieved before.

Clearly, quantum computing hardware is in its infancy. Existing hardware will need to be scaled up by many orders of magnitude (both in the number of stored qubits, and the number of gates that can be applied) before ambitious computations can be attempted. In the case of the NMR method, there is a particularly serious limitation that arises as a matter of principle, because the ratio of the coherent signal to the background declines exponentially with the number of spins per molecule. In practice, it will be very challenging to perform an NMR quantum computation with more than of order 10 qubits.

Probably, if quantum computers are eventually to become practical devices, new ideas about how to construct quantum hardware will be needed.

1.10 Summary

This concludes our introductory overview to quantum computation. We have seen that three converging factors have combined to make this subject exciting.

1. **Quantum computers can solve hard problems**. It seems that a new classification of complexity has been erected, a classification better founded on the fundamental laws of physics than traditional complexity theory. (But it remains to characterize more precisely the class of problems for which quantum computers have a big advantage over classical computers.)

2. **Quantum errors can be corrected**. With suitable coding methods, we can protect a complicated quantum system from the debilitating effects of decoherence. We may never see an actual cat that is half dead and half alive, but perhaps we can prepare and preserve an *encoded cat* that is half dead and half alive.

3. **Quantum hardware can be constructed**. We are privileged to be witnessing the dawn of the age of coherent manipulation of quantum information in the laboratory.

Our aim, in this course, will be to deepen our understanding of points (1), (2), and (3).

Chapter 2

Foundations I: States and Ensembles

2.1 Axioms of quantum mechanics

For a few lectures I have been talking about quantum this and that, but I have never defined what quantum theory is. It is time to correct that omission.

Quantum theory is a mathematical model of the physical world. To characterize the model, we need to specify how it will represent: states, observables, measurements, dynamics.

1. **States**. A state is a complete description of a physical system. In quantum mechanics, a state is a *ray* in a *Hilbert space.*

   What is a Hilbert space?

   a) It is a vector space over the complex numbers **C**. Vectors will be denoted $|\psi\rangle$ (Dirac's ket notation).

   b) It has an inner product $\langle\psi|\varphi\rangle$ that maps an ordered pair of vectors to **C**, defined by the properties

   (i) Positivity: $\langle\psi|\psi\rangle > 0$ for $|\psi\rangle = 0$

   (ii) Linearity: $\langle\varphi|(a|\psi_1\rangle + b|\psi_2\rangle) = a\langle\varphi|\psi_1\rangle + b\langle\varphi|\psi_2\rangle$

   (iii) Skew symmetry: $\langle\varphi|\psi\rangle = \langle\psi|\varphi\rangle^*$

   c) It is *complete* in the norm $||\psi|| = \langle\psi|\psi\rangle^{1/2}$

   (Completeness is an important proviso in infinite-dimensional function spaces, since it will ensure the convergence of certain eigenfunction expansions – e.g., Fourier analysis. But mostly we'll be content to work with finite-dimensional inner product spaces.)

   What is a ray? It is an equivalence class of vectors that differ by multiplication by a nonzero complex scalar. We can choose a representative of this class (for any nonvanishing vector) to have unit norm

$$\langle\psi|\psi\rangle = 1. \tag{2.1}$$

   We will also say that $|\psi\rangle$ and $e^{i\alpha}|\psi\rangle$ describe the same physical state, where $|e^{i\alpha}| = 1$.

   (Note that every ray corresponds to a possible state, so that given two states $|\varphi\rangle, |\psi\rangle$, we can form another as $a|\varphi\rangle + b|\psi\rangle$ (the "superposition principle"). The *relative* phase in this superposition *is* physically significant; we identify $a|\varphi\rangle + b|\varphi\rangle$ with $e^{i\alpha}(a|\varphi\rangle + b|\psi\rangle)$ but *not* with $a|\varphi\rangle + e^{i\alpha}b|\psi\rangle$.)

2. **Observables**. An observable is a property of a physical system that in principle can be measured. In quantum mechanics, an observable is a *self-adjoint operator.* An operator is a linear map taking vectors to vectors

$$\mathbf{A} : |\psi\rangle \to \mathbf{A}|\psi\rangle, \mathbf{A}\left(a|\psi\rangle + b|\psi\rangle\right) = a\mathbf{A}|\psi\rangle + b\mathbf{B}|\psi\rangle. \tag{2.2}$$

   The adjoint of the operator **A** is defined by

$$\langle\varphi|\mathbf{A}\psi\rangle = \langle\mathbf{A}^\dagger\varphi|\psi\rangle, \tag{2.3}$$

   for all vectors $|\varphi\rangle, |\psi\rangle$ (where here I have denoted $\mathbf{A}|\psi\rangle$ as $|\mathbf{A}\psi\rangle$). **A** is self-adjoint if $\mathbf{A} = \mathbf{A}^\dagger$.

   If **A** and **B** are self adjoint, then so is $\mathbf{A} + \mathbf{B}$ (because $(\mathbf{A} + \mathbf{B})^\dagger = \mathbf{A}^\dagger + \mathbf{B}^\dagger$) but $(\mathbf{AB})^\dagger = \mathbf{B}^\dagger\mathbf{A}^\dagger$, so **AB** is self adjoint only if **A** and **B** commute. Note that $\mathbf{AB}+\mathbf{BA}$ and $i(\mathbf{AB}-\mathbf{BA})$ are always self-adjoint if **A** and **B** are.

   A self-adjoint operator in a Hilbert space $\mathcal{H}$ has a spectral representation – it's eigenstates form a complete orthonormal basis in $\mathcal{H}$. We can express a self-adjoint operator **A** as

$$\mathbf{A} = \sum_n a_n \mathbf{P}_n. \tag{2.4}$$

Here each a_n is an eigenvalue of $\mathbf{A}$, and $\mathbf{P}_n$ is the corresponding orthogonal projection onto the space of eigenvectors with eigenvalue a_n. (If a_n is nondegenerate, then $\mathbf{P}_n = |n\rangle\langle n|$; it is the projection onto the corresponding eigenvector.) The $\mathbf{P}_n$'s satisfy

$$\begin{aligned} \mathbf{P}_n \mathbf{P}_m &= \delta_{n,m} \mathbf{P}_n \\ \mathbf{P}_n^\dagger &= \mathbf{P}_n . \end{aligned} \tag{2.5}$$

(For unbounded operators in an infinite-dimensional space, the definition of self-adjoint and the statement of the spectral theorem are more subtle, but this need not concern us.)

3. **Measurement**. In quantum mechanics, the numerical outcome of a measurement of the observable $\mathbf{A}$ is an eigenvalue of $\mathbf{A}$; right after the measurement, the quantum state is an eigenstate of $\mathbf{A}$ with the measured eigenvalue. If the quantum state just prior to the measurement is $|\psi\rangle$, then the outcome a_n is obtained with *probability*

$$\textbf{Prob}\ (a_n) = \parallel \mathbf{P}_n|\psi\rangle \parallel^2 = \langle\psi|\mathbf{P}_n|\psi\rangle ; \tag{2.6}$$

If the outcome is a_n is attained, then the (normalized) quantum state becomes

$$\frac{\mathbf{P}_n|\psi\rangle}{\left(\langle\psi|\mathbf{P}_n|\psi\rangle\right)^{1/2}} . \tag{2.7}$$

(Note that if the measurement is immediately repeated, then according to this rule the same outcome is attained again, with probability one.)

4. **Dynamics**. Time evolution of a quantum state is unitary; it is generated by a self-adjoint operator, called the *Hamiltonian* of the system. In the *Schrödinger picture* of dynamics, the vector describing the system moves in time as governed by the *Schrödinger equation*

$$\frac{d}{dt}|\psi(t)\rangle = -i\mathbf{H}|\psi(t)\rangle , \tag{2.8}$$

where $\mathbf{H}$ is the Hamiltonian. We may reexpress this equation, to first order in the infinitesimal quantity dt, as

$$|\psi(t+dt)\rangle = (\mathbf{1} - i\mathbf{H}dt)|\psi(t)\rangle . \tag{2.9}$$

The operator $\mathbf{U}(dt) \equiv \mathbf{1} - i\mathbf{H}dt$ is unitary; because $\mathbf{H}$ is self-adjoint it satisfies $\mathbf{U}^\dagger\mathbf{U} = 1$ to linear order in dt. Since a product of unitary operators is finite, time evolution over a finite interval is also unitary

$$|\psi(t)\rangle = \mathbf{U}(t)|\psi(0)\rangle . \tag{2.10}$$

In the case where $\mathbf{H}$ is t-independent; we may write $\mathbf{U} = e^{-it\mathbf{H}}$.

This completes the mathematical formulation of quantum mechanics. We immediately notice some curious features. One oddity is that the Schrödinger equation is linear, while we are accustomed to nonlinear dynamical equations in classical physics. This property seems to beg for an explanation. But far more curious is the mysterious dualism; there are two quite distinct ways for a quantum state to change. On the one hand there is unitary evolution, which is deterministic. If we specify $|\psi(0)\rangle$, the theory predicts the state $|\psi(t)\rangle$ at a later time.

But on the other hand there is measurement, which is probabilistic. The theory does not make definite predictions about the measurement outcomes; it only assigns probabilities to the various alternatives. This is troubling, because it is unclear why the measurement process should be governed by different physical laws than other processes.

Beginning students of quantum mechanics, when first exposed to these rules, are often told not to ask "why?" There is much wisdom in this advice. But I believe that it can be useful to ask way. In future lectures. we will return to this disconcerting dualism between unitary evolution and measurement, and will seek a resolution.

2.2 The Qubit

The indivisible unit of classical information is the *bit*, which takes one of the two possible values $\{0, 1\}$. The corresponding unit of quantum information is called the "quantum bit" or *qubit*. It describes a state in the simplest possible quantum system.

The smallest nontrivial Hilbert space is two-dimensional. We may denote an orthonormal basis for a two-dimensional vector space as $\{|0\rangle, |1\rangle\}$. Then the most general normalized state can be expressed as

$$a|0\rangle + b|1\rangle , \tag{2.11}$$

where a, b are complex numbers that satisfy $|a|^2 + |b|^2 = 1$, and the overall phase is physically irrelevant. A *qubit* is a state in a two-dimensional Hilbert space that can take any value of the form eq. (2.11).

We can perform a measurement that projects the qubit onto the basis $\{|0\rangle, |1\rangle\}$. Then we will obtain the outcome $|0\rangle$ with probability $|a|^2$, and the outcome $|1\rangle$ with probability $|b|^2$. Furthermore, except in the cases $a = 0$ and $b = 0$, the measurement irrevocably disturbs the state. If the value of the qubit is initially unknown, then there is no way to determine a and b with that single measurement, or any other conceivable measurement. However, *after* the measurement, the qubit has been prepared in a *known* state – either $|0\rangle$ or $|1\rangle$ – that differs (in general) from its previous state.

In this respect, a qubit differs from a classical bit; we can measure a classical bit without disturbing it, and we can decipher all of the information that it encodes. But suppose we have a classical bit that really does have a definite value (either 0 or 1), but that value is initially unknown to us. Based on the information available to us we can only say that there is a *probability* p_0 that the bit has the value 0, and a probability p_1 that the bit has the value 1, where $p_0 + p_1 = 1$. When we measure the bit, we acquire additional information; afterwards we know the value with 100% confidence.

An important question is: what is the essential difference between a qubit and a *probabilistic* classical bit? In fact they are *not* the same, for several reasons that we will explore.

2.2.1 Spin-$\frac{1}{2}$

First of all, the coefficients a and b in eq. (2.11) encode more than just the probabilities of the outcomes of a measurement in the $\{|0\rangle, |1\rangle\}$ basis. In particular, the *relative phase* of a and b also has physical significance.

For a physicist, it is natural to interpret eq. (2.11) as the spin state of an object with spin-$\frac{1}{2}$ (like an electron). Then $|0\rangle$ and $|1\rangle$ are the spin up ($|\uparrow\rangle$) and spin down ($|\downarrow\rangle$) states along a particular axis such as the z-axis. The two real numbers characterizing the qubit (the complex numbers a and b, modulo the normalization and overall phase) describe the *orientation* of the spin in three-dimensional space (the polar angle θ and the azimuthal angle φ).

We cannot go deeply here into the theory of symmetry in quantum mechanics, but we will briefly recall some elements of the theory that will prove useful to us. A symmetry is a transformation that acts on a state of a system,

yet leaves all observable properties of the system unchanged. In quantum mechanics, observations are measurements of self-adjoint operators. If $\mathbf{A}$ is measured in the state $|\psi\rangle$, then the outcome $|a\rangle$ (an eigenvector of $\mathbf{A}$) occurs with probability $|\langle a|\psi\rangle|^2$. A symmetry should leave these probabilities unchanged (when we "rotate" both the system *and* the apparatus).

A symmetry, then, is a mapping of vectors in Hilbert space

$$|\psi\rangle \to |\psi'\rangle, \tag{2.12}$$

that preserves the absolute values of inner products

$$|\langle\varphi|\psi\rangle| = |\langle\varphi'|\psi'\rangle|, \tag{2.13}$$

for all $|\varphi\rangle$ and $|\psi\rangle$. According to a famous theorem due to Wigner, a mapping with this property can always be chosen (by adopting suitable phase conventions) to be either unitary or antiunitary. The antiunitary alternative, while important for discrete symmetries, can be excluded for continuous symmetries. Then the symmetry acts as

$$|\psi\rangle \to |\psi'\rangle = \mathbf{U}|\psi\rangle, \tag{2.14}$$

where $\mathbf{U}$ is unitary (and in particular, *linear*).

Symmetries form a group: a symmetry transformation can be inverted, and the product of two symmetries is a symmetry. For each symmetry operation R acting on our physical system, there is a corresponding unitary transformation $\mathbf{U}(R)$. Multiplication of these unitary operators must respect the group multiplication law of the symmetries – applying $R_1 \circ R_2$ should be equivalent to first applying R_2 and subsequently R_1. Thus we demand

$$\mathbf{U}(R_1)\mathbf{U}(R_2) = \text{ Phase } (R_1, R_2)\mathbf{U}(R_1 \circ R_2) \tag{2.15}$$

The phase is permitted in eq. (2.15) because quantum states are *rays*; we need only demand that $\mathbf{U}(R_1 \circ R_2)$ act the same way as $\mathbf{U}(R_1)\mathbf{U}(R_2)$ on rays, not on vectors. $\mathbf{U}(R)$ provides a unitary representation (up to a phase) of the symmetry group.

So far, our concept of symmetry has no connection with dynamics. Usually, we demand of a symmetry that it respect the dynamical evolution of the system. This means that it should not matter whether we first transform the system and then evolve it, or first evolve it and then transform it. In other words, the diagram

$$
\begin{tikzcd}
\boxed{\text{Initial}} \arrow[r, "dynamics"] \arrow[d, "rotation"] & \boxed{\text{Final}} \arrow[d, "rotation"] \\
\boxed{\text{New Initial}} \arrow[r, "dynamics"] & \boxed{\text{New Final}}
\end{tikzcd}
$$

is commutative. This means that the time evolution operator $e^{it\mathbf{H}}$ should commute with the symmetry transformation $\mathbf{U}(R)$:

$$\mathbf{U}(R)e^{-it\mathbf{H}} = e^{-it\mathbf{H}}\mathbf{U}(R), \tag{2.16}$$

and expanding to linear order in t we obtain

$$\mathbf{U}(R)\mathbf{H} = \mathbf{H}\mathbf{U}(R) \tag{2.17}$$

For a continuous symmetry, we can choose R infinitesimally close to the identity, $R = I + \epsilon T$, and then $\mathbf{U}$ is close to $\mathbf{1}$,

$$\mathbf{U} = \mathbf{1} - i\varepsilon\mathbf{Q} + O(\varepsilon^2). \tag{2.18}$$

From the unitarity of $\mathbf{U}$ (to order ε) it follows that $\mathbf{Q}$ is an observable, $\mathbf{Q} = \mathbf{Q}^\dagger$. Expanding eq. (2.17) to linear order in ε we find

$$[\mathbf{Q}, \mathbf{H}] = 0; \tag{2.19}$$

the observable $\mathbf{Q}$ commutes with the Hamiltonian.

Eq. (2.19) is a *conservation law*. It says, for example, that if we prepare an eigenstate of $\mathbf{Q}$, then time evolution governed by the Schrödinger equation will preserve the eigenstate. We have seen that symmetries imply conservation laws. Conversely, given a conserved quantity $\mathbf{Q}$ satisfying eq. (2.19) we can construct the corresponding symmetry transformations. Finite transformations can be built as a product of many infinitesimal ones

$$R = (1 + \frac{\theta}{N}T)^N \Rightarrow \mathbf{U}(R) = (\mathbf{1} + i\frac{\theta}{N}\mathbf{Q})^N \to e^{i\theta\mathbf{Q}}, \tag{2.20}$$

(taking the limit $N \to \infty$). Once we have decided how infinitesimal symmetry transformations are represented by unitary operators, then it is also

determined how finite transformations are represented, for these can be built as a product of infinitesimal transformations. We say that $\mathbf{Q}$ is the *generator* of the symmetry.

Let us briefly recall how this general theory applies to spatial rotations and angular momentum. An infinitesimal rotation by $d\theta$ about the axis specified by the unit vector $\hat{n} = (n_1, n_2, n_3)$ can be expressed as

$$R(\hat{n}, d\theta) = I - id\theta\hat{n}\cdot\vec{J}, \tag{2.21}$$

where (J_1, J_2, J_3) are the components of the angular momentum. A finite rotation is expressed as

$$R(\hat{n}, \theta) = \exp(-i\theta\hat{n}\cdot\vec{J}). \tag{2.22}$$

Rotations about distinct axes don't commute. From elementary properties of rotations, we find the commutation relations

$$[J_k, J_\ell] = i\varepsilon_{k\ell m}J_m, \tag{2.23}$$

where $\varepsilon_{k\ell m}$ is the totally antisymmetric tensor with $\varepsilon_{123} = 1$, and repeated indices are summed. To implement rotations on a quantum system, we find self-adjoint operators $\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3$ in Hilbert space that satisfy these relations.

The "defining" representation of the rotation group is three dimensional, but the simplest nontrivial irreducible representation is two dimensional, given by

$$\mathbf{J}_k = \frac{1}{2}\boldsymbol{\sigma}_k, \tag{2.24}$$

where

$$\boldsymbol{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \boldsymbol{\sigma}_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \boldsymbol{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{2.25}$$

are the Pauli matrices. This is the unique two-dimensional irreducible representation, up to a unitary change of basis. Since the eigenvalues of $\mathbf{J}_k$ are $\pm\frac{1}{2}$, we call this the spin-$\frac{1}{2}$ representation. (By identifying $\mathbf{J}$ as the angular-momentum, we have implicitly chosen units with $\hbar = 1$).

The Pauli matrices also have the properties of being mutually anticommuting and squaring to the identity,

$$\boldsymbol{\sigma}_k\boldsymbol{\sigma}_\ell + \boldsymbol{\sigma}_\ell\boldsymbol{\sigma}_k = 2\delta_{k\ell}\mathbf{1}, \tag{2.26}$$

So we see that $(\hat{n}\cdot\vec{\boldsymbol{\sigma}})^2 = n_k n_\ell \boldsymbol{\sigma}_k \boldsymbol{\sigma}_\ell = n_k n_k \mathbf{1} = \mathbf{1}$. By expanding the exponential series, we see that finite rotations are represented as

$$\mathbf{U}(\hat{n},\theta) = e^{-i\frac{\theta}{2}\hat{n}\cdot\vec{\boldsymbol{\sigma}}} = \mathbf{1}\cos\frac{\theta}{2} - i\hat{n}\cdot\vec{\boldsymbol{\sigma}}\sin\frac{\theta}{2}. \tag{2.27}$$

The most general 2×2 unitary matrix with determinant 1 can be expressed in this form. Thus, we are entitled to think of a qubit as the state of a spin-$\frac{1}{2}$ object, and an arbitrary unitary transformation acting on the state (aside from a possible rotation of the overall phase) is a *rotation* of the spin.

A peculiar property of the representation $\mathbf{U}(\hat{n},\theta)$ is that it is *double-valued*. In particular a rotation by 2π about any axis is represented nontrivially:

$$\mathbf{U}(\hat{n},\theta = 2\pi) = -\mathbf{1}. \tag{2.28}$$

Our representation of the rotation group is really a representation "up to a sign"

$$\mathbf{U}(R_1)\mathbf{U}(R_2) = \pm\mathbf{U}(R_1\circ R_2). \tag{2.29}$$

But as already noted, this is acceptable, because the group multiplication is respected on *rays*, though not on vectors. These double-valued representations of the rotation group are called *spinor* representations. (The existence of spinors follows from a topological property of the group — it is not simply connected.)

While it is true that a rotation by 2π has no detectable effect on a spin-$\frac{1}{2}$ object, it would be wrong to conclude that the spinor property has no observable consequences. Suppose I have a machine that acts on a pair of spins. If the first spin is up, it does nothing, but if the first spin is down, it rotates the second spin by 2π. Now let the machine act when the first spin is in a *superposition* of up and down. Then

$$\frac{1}{\sqrt{2}}(|\uparrow\rangle_1 + |\downarrow\rangle_1)\,|\uparrow\rangle_2 \rightarrow \frac{1}{\sqrt{2}}(|\uparrow\rangle_1 - |\downarrow\rangle_1)\,|\uparrow\rangle_2\ . \tag{2.30}$$

While there is no detectable effect on the second spin, the state of the first has flipped to an orthogonal state, which is very much observable.

In a rotated frame of reference, a rotation $R(\hat{n},\theta)$ becomes a rotation through the same angle but about a rotated axis. It follows that the three components of angular momentum transform under rotations as a vector:

$$\mathbf{U}(R)\mathbf{J}_k\mathbf{U}(R)^\dagger = R_{k\ell}\mathbf{J}_\ell. \tag{2.31}$$

Thus, if a state $|m\rangle$ is an eigenstate of $\mathbf{J}_3$

$$\mathbf{J}_3|m\rangle = m|m\rangle, \tag{2.32}$$

then $\mathbf{U}(R)|m\rangle$ is an eigenstate of $R\mathbf{J}_3$ with the same eigenvalue:

$$\begin{aligned} R\mathbf{J}_3\left(\mathbf{U}(R)|m\rangle\right) &= \mathbf{U}(R)\mathbf{J}_3\mathbf{U}(R)^\dagger\mathbf{U}(R)|m\rangle \\ &= \mathbf{U}(R)\mathbf{J}_3|m\rangle = m\left(\mathbf{U}(R)|m\rangle\right). \end{aligned} \tag{2.33}$$

Therefore, we can construct eigenstates of angular momentum along the axis $\hat{n} = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$ by applying a rotation through θ, about the axis $\hat{n}' = (-\sin\varphi, \cos\varphi, 0)$, to a $\mathbf{J}_3$ eigenstate. For our spin-$\frac{1}{2}$ representation, this rotation is

$$\begin{aligned} \exp\left[-i\frac{\theta}{2}\hat{n}'\cdot\vec{\boldsymbol{\sigma}}\right] &= \exp\left[\frac{\theta}{2}\begin{pmatrix} 0 & -e^{-i\varphi} \\ e^{i\varphi} & 0 \end{pmatrix}\right] \\ &= \begin{pmatrix} \cos\frac{\theta}{2} & -e^{-i\varphi}\sin\frac{\theta}{2} \\ e^{i\varphi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}, \end{aligned} \tag{2.34}$$

and applying it to $\binom{1}{0}$, the $\mathbf{J}_3$ eigenstate with eigenvalue 1, we obtain

$$|\psi(\theta,\varphi)\rangle = \begin{pmatrix} e^{-i\varphi/2}\cos\frac{\theta}{2} \\ e^{i\varphi/2}\sin\frac{\theta}{2} \end{pmatrix}, \tag{2.35}$$

(up to an overall phase). We can check directly that this is an eigenstate of

$$\hat{n}\cdot\vec{\boldsymbol{\sigma}} = \begin{pmatrix} \cos\theta & e^{-i\varphi}\sin\theta \\ e^{i\varphi}\sin\theta & -\cos\theta \end{pmatrix}, \tag{2.36}$$

with eigenvalue one. So we have seen that eq. (2.11) with $a = e^{-i\varphi/2}\cos\frac{\theta}{2}$, $b = e^{i\varphi/2}\sin\frac{\theta}{2}$, can be interpreted as a spin pointing in the (θ,φ) direction.

We noted that we cannot determine a and b with a single measurement. Furthermore, even with many identical copies of the state, we cannot completely determine the state by measuring each copy only along the z-axis. This would enable us to estimate $|a|$ and $|b|$, but we would learn nothing about the relative phase of a and b. Equivalently, we would find the component of the spin along the z-axis

$$\langle\psi(\theta,\varphi)|\sigma_3|\psi(\theta,\varphi)\rangle = \cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2} = \cos\theta, \tag{2.37}$$

but we would not learn about the component in the $x-y$ plane. The problem of determining $|\psi\rangle$ by measuring the spin is equivalent to determining the unit vector $\hat{n}$ by measuring its components along various axes. Altogether, measurements along three different axes are required. *E.g.*, from $\langle\sigma_3\rangle$ and $\langle\sigma_1\rangle$ we can determine n_3 and n_1, but the sign of n_2 remains undetermined. Measuring $\langle\sigma_2\rangle$ would remove this remaining ambiguity.

Of course, if we are permitted to rotate the spin, then only measurements along the z-axis will suffice. That is, measuring a spin along the $\hat{n}$ axis is equivalent to first applying a rotation that rotates the $\hat{n}$ axis to the axis $\hat{z}$, and then measuring along $\hat{z}$.

In the special case $\theta = \frac{\pi}{2}$ and $\varphi = 0$ (the $\hat{x}$-axis) our spin state is

$$|\uparrow_x\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow_z\rangle + |\downarrow_z\rangle\right), \tag{2.38}$$

("spin-up along the x-axis"). The orthogonal state ("spin down along the x-axis") is

$$|\downarrow_x\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow_z\rangle - |\downarrow_z\rangle\right). \tag{2.39}$$

For either of these states, if we measure the spin along the z-axis, we will obtain $|\uparrow_z\rangle$ with probability $\frac{1}{2}$ and $|\downarrow_z\rangle$ with probability $\frac{1}{2}$.

Now consider the combination

$$\frac{1}{\sqrt{2}}\left(|\uparrow_x\rangle + |\downarrow_x\rangle\right). \tag{2.40}$$

This state has the property that, if we measure the spin along the x-axis, we obtain $|\uparrow_x\rangle$ or $|\downarrow_x\rangle$, each with probability $\frac{1}{2}$. Now we may ask, what if we measure the state in eq. (2.40) along the z-axis?

If these were probabilistic classical bits, the answer would be obvious. The state in eq. (2.40) is in one of two states, and for *each* of the two, the probability is $\frac{1}{2}$ for pointing up or down along the z-axis. So of course we should find up with probability $\frac{1}{2}$ when we measure along the z-axis.

But not so for qubits! By adding eq. (2.38) and eq. (2.39), we see that the state in eq. (2.40) is really $|\uparrow_z\rangle$ in disguise. When we measure along the z-axis, we always find $|\uparrow_z\rangle$, never $|\downarrow_z\rangle$.

We see that for qubits, as opposed to probabilistic classical bits, probabilities can add in unexpected ways. This is, in its simplest guise, the phenomenon called "quantum interference," an important feature of quantum information.

It should be emphasized that, while this *formal* equivalence with a spin-$\frac{1}{2}$ object applies to any two-level quantum system, of course not every two-level system transforms as a spinor under rotations!

2.2.2 Photon polarizations

Another important two-state system is provided by a *photon*, which can have two independent polarizations. These photon polarization states also transform under rotations, but photons differ from our spin-$\frac{1}{2}$ objects in two important ways: (1) Photons are massless. (2) Photons have spin-1 (they are not spinors).

Now is not a good time for a detailed discussion of the unitary representations of the Poincare group. Suffice it to say that the *spin* of a particle classifies how it transforms under the *little group*, the subgroup of the Lorentz group that preserves the particle's momentum. For a massive particle, we may always boost to the particle's rest frame, and then the little group is the rotation group.

For massless particles, there is no rest frame. The finite-dimensional unitary representations of the little group turn out to be representations of the rotation group in *two* dimensions, the rotations about the axis determined by the momentum. Of course, for a photon, this corresponds to the familiar property of classical light – the waves are polarized transverse to the direction of propagation.

Under a rotation about the axis of propagation, the two linear polarization states ($|x\rangle$ and $|y\rangle$ for horizontal and vertical polarization) transform as

$$\begin{aligned} |x\rangle &\rightarrow \cos\theta|x\rangle + \sin\theta|y\rangle \\ |y\rangle &\rightarrow -\sin\theta|x\rangle + \cos\theta|y\rangle. \end{aligned} \tag{2.41}$$

This two-dimensional representation is actually reducible. The matrix

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{2.42}$$

has the eigenstates

$$|R\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ i \end{pmatrix} \qquad |L\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} i \\ 1 \end{pmatrix}, \tag{2.43}$$

with eigenvalues $e^{i\theta}$ and $e^{-i\theta}$, the states of right and left circular polarization. That is, these are the eigenstates of the rotation generator

$$J = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \sigma_y, \tag{2.44}$$

with eigenvalues ± 1. Because the eigenvalues are ± 1 (*not* $\pm\frac{1}{2}$) we say that the photon has spin-1.

In this context, the quantum interference phenomenon can be described this way: Suppose that we have a polarization analyzer that allows only one of the two linear photon polarizations to pass through. Then an x or y polarized photon has prob $\frac{1}{2}$ of getting through a 45^o rotated polarizer, and a 45^o polarized photon has prob $\frac{1}{2}$ of getting through an x and y analyzer. But an x photon *never* passes through a y analyzer. If we put a 45^o rotated analyzer in between an x and y analyzer, then $\frac{1}{2}$ the photons make it through each analyzer. But if we remove the analyzer in the middle *no* photons make it through the y analyzer.

A device can be constructed easily that rotates the linear polarization of a photon, and so applies the transformation Eq. (2.41) to our qubit. As noted, this is not the most general possible unitary transformation. But if we also have a device that alters the relative phase of the two orthogonal linear polarization states

$$\begin{aligned} |x\rangle &\rightarrow e^{i\omega/2}|x\rangle \\ |y\rangle &\rightarrow e^{-i\omega/2}|y\rangle \ , \end{aligned} \tag{2.45}$$

the two devices can be employed together to apply an arbitrary 2×2 unitary transformation (of determinant 1) to the photon polarization state.

2.3 The density matrix

2.3.1 The bipartite quantum system

The last lecture was about one qubit. This lecture is about *two* qubits. (Guess what the next lecture will be about!) Stepping up from one qubit to two is a bigger leap than you might expect. Much that is weird and wonderful about quantum mechanics can be appreciated by considering the properties of the quantum states of two qubits.

The axioms of §2.1 provide a perfectly acceptable general formulation of the quantum theory. Yet under many circumstances, we find that the axioms appear to be violated. The trouble is that our axioms are intended to characterize the quantum behavior of the entire universe. Most of the time, we are not so ambitious as to attempt to understand the physics of the whole universe; we are content to observe just our little corner. In practice, then, the observations we make are always limited to a small part of a much larger quantum system.

In the next several lectures, we will see that, when we limit our attention to just part of a larger system, then (contrary to the axioms):

1. States are *not* rays.
2. Measurements are *not* orthogonal projections.
3. Evolution is *not* unitary.

We can best understand these points by considering the simplest possible example: a two-qubit world in which we observe only one of the qubits.

So consider a system of two qubits. Qubit A is here in the room with us, and we are free to observe or manipulate it any way we please. But qubit B is locked in a vault where we can't get access to it. Given some quantum state of the two qubits, we would like to find a compact way to characterize the observations that can be made on qubit A alone.

We'll use $\{|0\rangle_A, |1\rangle_A\}$ and $\{|0\rangle_B, |1\rangle_B\}$ to denote orthonormal bases for qubits A and B respectively. Consider a quantum state of the two-qubit world of the form

$$|\psi\rangle_{AB} = a|0\rangle_A \otimes |0\rangle_B + b|1\rangle_A \otimes |1\rangle_B. \tag{2.46}$$

In this state, qubits A and B are *correlated*. Suppose we measure qubit A by projecting onto the $\{|0\rangle_A, |1\rangle_A\}$ basis. Then with probability $|a|^2$ we obtain the result $|0\rangle_A$, and the measurement prepares the state

$$|0\rangle_A \otimes |0\rangle_B. \tag{2.47}$$

with probability $|b|^2$, we obtain the result $|1\rangle_A$ and prepare the state

$$|1\rangle_A \otimes |1\rangle_B. \tag{2.48}$$

In either case, a definite state of qubit B is picked out by the measurement. If we subsequently measure qubit B, then we are guaranteed (with probability one) to find $|0\rangle_B$ if we had found $|0\rangle_A$, and we are guaranteed to find $|1\rangle_B$ if we found $|1\rangle_A$. In this sense, the outcomes of the $\{|0\rangle_A, |1\rangle_A\}$ and $\{|0\rangle_B, |1\rangle_B\}$ measurements are perfectly correlated in the state $|\psi\rangle_{AB}$.

But now I would like to consider more general observables acting on qubit A, and I would like to characterize the measurement outcomes for A alone (irrespective of the outcomes of any measurements of the inaccessible qubit B). An observable acting on qubit A only can be expressed as

$$\mathbf{M}_A \otimes \mathbf{1}_B, \tag{2.49}$$

where $\mathbf{M}_A$ is a self-adjoint operator acting on A, and $\mathbf{1}_B$ is the identity operator acting on B. The expectation value of the observable in the state $|\psi\rangle$ is:

$$\begin{aligned} &\langle\psi|\mathbf{M}_A \otimes \mathbf{1}_B|\psi\rangle \\ &= (a^*{}_A\langle 0| \otimes {}_B\langle 0| + b^*{}_B\langle 1| \otimes {}_B\langle 1|)\,(\mathbf{M}_A \otimes \mathbf{1}_B) \\ &\quad (a|0\rangle_A \otimes |0\rangle_B + b|1\rangle_A \otimes |1\rangle_B) \\ &= |a|^2{}_A\langle 0|\mathbf{M}_A|0\rangle_A + |b|^2{}_A\langle 1|\mathbf{M}_A|1\rangle_A, \end{aligned} \tag{2.50}$$

(where we have used the orthogonality of $|0\rangle_B$ and $|1\rangle_B$). This expression can be rewritten in the form

$$\langle \mathbf{M}_A\rangle = \mathrm{tr}\,(\mathbf{M}_A \boldsymbol{\rho}_A)\,, \tag{2.51}$$

$$\boldsymbol{\rho}_A = |a|^2|0\rangle_A\,{}_A\langle 0| + |b|^2|1\rangle_A\,{}_A\langle 1|, \tag{2.52}$$

and $\mathrm{tr}(\cdot)$ denotes the *trace*. The operator $\boldsymbol{\rho}_A$ is called the *density operator* (or *density matrix*) for qubit A. It is self-adjoint, positive (its eigenvalues are nonnegative) and it has unit trace (because $|\psi\rangle$ is a normalized state.)

Because $\langle \mathbf{M}_A\rangle$ has the form eq. (2.51) for *any* observable $\mathbf{M}_A$ acting on qubit A, it is consistent to interpret $\boldsymbol{\rho}_A$ as representing an *ensemble* of possible quantum states, each occurring with a specified probability. That is, we would obtain precisely the same result for $\langle \mathbf{M}_A\rangle$ if we stipulated that qubit A is in one of two quantum states. With probability $p_0 = |a|^2$ it is in the quantum state $|0\rangle_A$, and with probability $p_1 = |b|^2$ it is in the state

$|1\rangle_A$. If we are interested in the result of any possible measurement, we can consider $\mathbf{M}_A$ to be the projection $\mathbf{E}_A(a)$ onto the relevant eigenspace of a particular observable. Then

$$\mathrm{Prob}\ (a) = p_0\ {}_A\langle 0|\mathbf{E}_A(a)|0\rangle_A + p_1\ {}_A\langle 1|\mathbf{E}_A(a)|1\rangle_A, \tag{2.53}$$

which is the probability of outcome a summed over the ensemble, and weighted by the probability of each state in the ensemble.

We have emphasized previously that there is an essential difference between a coherent superposition of the states $|0\rangle_A$ and $|1\rangle_A$, and a probabilistic ensemble, in which $|0\rangle_A$ and $|1\rangle_A$ can each occur with specified probabilities. For example, for a spin-$\frac{1}{2}$ object we have seen that if we measure $\boldsymbol{\sigma}_1$ in the state $\frac{1}{\sqrt{2}}(|\uparrow_z\rangle + |\downarrow_z\rangle)$, we will obtain the result $|\uparrow_x\rangle$ with probability one. But the ensemble in which $|\uparrow_z\rangle$ and $|\downarrow_z\rangle$ each occur with probability $\frac{1}{2}$ is represented by the density operator

$$\begin{aligned} \boldsymbol{\rho} &= \frac{1}{2}(|\uparrow_z\rangle\langle\uparrow_z| + |\downarrow_z\rangle\langle\downarrow_z|) \\ &= \frac{1}{2}\mathbf{1}, \end{aligned} \tag{2.54}$$

and the projection onto $|\uparrow_x\rangle$ then has the expectation value

$$\mathrm{tr}\,(|\uparrow_x\rangle\langle\uparrow_x|\boldsymbol{\rho}) = \frac{1}{2}. \tag{2.55}$$

In fact, we have seen that any state of one qubit represented by a ray can be interpreted as a spin pointing in some definite direction. But because the identity is left unchanged by any unitary change of basis, and the state $|\psi(\theta,\varphi)\rangle$ can be obtained by applying a suitable unitary transformation to $|\uparrow_z\rangle$, we see that for $\boldsymbol{\rho}$ given by eq. (2.54), we have

$$\mathrm{tr}\,(|\psi(\theta,\varphi)\rangle\langle\psi(\theta,\varphi)|\boldsymbol{\rho}) = \frac{1}{2}. \tag{2.56}$$

Therefore, if the state $|\psi\rangle_{AB}$ in eq. (2.57) is prepared, with $|a|^2 = |b|^2 = \frac{1}{2}$, and we measure the spin A along *any* axis, we obtain a completely random result; spin up or spin down can occur, each with probability $\frac{1}{2}$.

This discussion of the correlated two-qubit state $|\psi\rangle_{AB}$ is easily generalized to an arbitrary state of any bipartite quantum system (a system divided into two parts). The Hilbert space of a bipartite system is $\mathcal{H}_A \otimes \mathcal{H}_B$ where

$\mathcal{H}_{A,B}$ are the Hilbert spaces of the two parts. This means that if $\{|i\rangle_A\}$ is an orthonormal basis for $\mathcal{H}_A$ and $\{|\mu\rangle_B\}$ is an orthonormal basis for $\mathcal{H}_B$, then $\{|i\rangle_A \otimes |\mu\rangle_B\}$ is an orthonormal basis for $\mathcal{H}_A \otimes \mathcal{H}_B$. Thus an arbitrary pure state of $\mathcal{H}_A \otimes \mathcal{H}_B$ can be expanded as

$$|\psi\rangle_{AB} = \sum_{i,\mu} a_{i\mu} |i\rangle_A \otimes |\mu\rangle_B, \tag{2.57}$$

where $\sum_{i,\mu} |a_{i\mu}|^2 = 1$. The expectation value of an observable $\mathbf{M}_A \otimes \mathbf{1}_B$, that acts only on subsystem A is

$$\begin{aligned}\langle \mathbf{M}_A \rangle &= {}_{AB}\langle \psi | \mathbf{M}_A \otimes \mathbf{1}_B | \psi \rangle_{AB} \\ &= \sum_{j,\nu} a^*_{j\nu} \left({}_A\langle j| \otimes {}_B\langle \nu| \right) (\mathbf{M}_A \otimes \mathbf{1}_B) \sum_{i,\mu} a_{i\mu} \left(|i\rangle_A \otimes |\mu\rangle_B \right) \\ &= \sum_{i,j,\mu} a^*_{j\mu} a_{i\mu} \; {}_A\langle j | \mathbf{M}_A | i \rangle_A \\ &= \operatorname{tr}\left(\mathbf{M}_A \boldsymbol{\rho}_A \right),\end{aligned} \tag{2.58}$$

where

$$\begin{aligned}\boldsymbol{\rho}_A &= \operatorname{tr}_B \left(|\psi\rangle_{AB} \; {}_{AB}\langle \psi| \right) \\ &\equiv \sum_{i,j,\mu} a_{i\mu} a^*_{j\mu} |i\rangle_A \; {}_A\langle j| \; .\end{aligned} \tag{2.59}$$

We say that the density operator $\boldsymbol{\rho}_A$ for subsystem A is obtained by performing a partial *trace* over subsystem B of the density matrix (in this case a pure state) for the combined system AB.

From the definition eq. (2.59), we can immediately infer that $\boldsymbol{\rho}_A$ has the following properties:

1. $\boldsymbol{\rho}_A$ is self-adjoint: $\boldsymbol{\rho}_A = \boldsymbol{\rho}_A^\dagger$.
2. $\boldsymbol{\rho}_A$ is positive: For any $|\psi\rangle_A$ ${}_A\langle \psi | \boldsymbol{\rho}_A | \psi \rangle_A = \sum_\mu |\sum_i a_{i\mu} \; {}_A\langle \psi | i \rangle_A|^2 \geq 0$.
3. $\operatorname{tr}(\boldsymbol{\rho}_A) = 1$: We have tr $\boldsymbol{\rho}_A = \sum_{i,\mu} |a_{i\mu}|^2 = 1$, since $|\psi\rangle_{AB}$ is normalized.

It follows that $\boldsymbol{\rho}_A$ can be diagonalized, that the eigenvalues are all real and nonnegative, and that the eigenvalues sum to one.

If we are looking at a subsystem of a larger quantum system, then, even if the state of the larger system is a ray, the state of the subsystem need

not be; in general, the state is represented by a density operator. In the case where the state of the subsystem *is* a ray, and we say that the state is *pure*. Otherwise the state is *mixed*. If the state is a pure state $|\psi\rangle_A$, then the density matrix $\boldsymbol{\rho}_A = |\psi\rangle_A \; {}_A\langle\psi|$ is the *projection* onto the one-dimensional space spanned by $|\psi\rangle_A$. Hence a pure density matrix has the property $\boldsymbol{\rho}^2 = \boldsymbol{\rho}$. A general density matrix, expressed in the basis in which it is diagonal, has the form

$$\boldsymbol{\rho}_A = \sum_a p_a |\psi_a\rangle\langle\psi_a|, \tag{2.60}$$

where $0 < p_a \leq 1$ and $\sum_a p_a = 1$. If the state is not pure, there are two or more terms in this sum, and $\boldsymbol{\rho}^2 \neq \boldsymbol{\rho}$; in fact, tr $\boldsymbol{\rho}^2 = \sum p_a^2 < \sum p_a = 1$. We say that $\boldsymbol{\rho}$ is an *incoherent* superposition of the states $\{|\psi_a\rangle\}$; incoherent meaning that the relative phases of the $|\psi_a\rangle$ are experimentally inaccessible.

Since the expectation value of *any* observable $\mathbf{M}$ acting on the subsystem can be expressed as

$$\langle \mathbf{M} \rangle = \operatorname{tr}\mathbf{M}\boldsymbol{\rho} = \sum_a p_a \langle \psi_a | \mathbf{M} | \psi_a \rangle, \tag{2.61}$$

we see as before that we may interpret $\boldsymbol{\rho}$ as describing an *ensemble* of pure quantum states, in which the state $|\psi_a\rangle$ occurs with probability p_a. We have, therefore, come a long part of the way to understanding how probabilities arise in quantum mechanics when a quantum system A interacts with another system B. A and B become *entangled*, that is, correlated. The entanglement *destroys the coherence* of a superposition of states of A, so that some of the phases in the superposition become inaccessible if we look at A alone. We may describe this situation by saying that the state of system A *collapses* — it is in one of a set of alternative states, each of which can be assigned a probability.

2.3.2 Bloch sphere

Let's return to the case in which system A is a single qubit, and consider the form of the general density matrix. The most general self-adjoint 2×2 matrix has four real parameters, and can be expanded in the basis $\{\mathbf{1}, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3\}$. Since each $\boldsymbol{\sigma}_i$ is traceless, the coefficient of $\mathbf{1}$ in the expansion of a density matrix $\boldsymbol{\rho}$ must be $\frac{1}{2}$ (so that $\mathrm{tr}(\boldsymbol{\rho}) = 1$), and $\boldsymbol{\rho}$ may be expressed as

$$\begin{aligned}\boldsymbol{\rho}(\vec{P}) &= \frac{1}{2}\left(\mathbf{1} + \vec{P}\cdot\vec{\boldsymbol{\sigma}}\right)\\ &\equiv \frac{1}{2}\left(\mathbf{1} + P_1\boldsymbol{\sigma}_1 + P_2\boldsymbol{\sigma}_2 + P_3\boldsymbol{\sigma}_3\right)\\ &= \frac{1}{2}\begin{pmatrix} 1+P_3 & P_1 - iP_2 \\ P_1 + iP_2 & 1-P_3\end{pmatrix}. \end{aligned} \tag{2.62}$$

We can compute $\det\boldsymbol{\rho} = \frac{1}{4}\left(1-\vec{P}^2\right)$. Therefore, a necessary condition for $\boldsymbol{\rho}$ to have nonnegative eigenvalues is $\det\boldsymbol{\rho} \geq 0$ or $\vec{P}^2 \leq 1$. This condition is also sufficient; since $\mathrm{tr}\boldsymbol{\rho} = 1$, it is not possible for $\boldsymbol{\rho}$ to have two negative eigenvalues. Thus, there is a $1-1$ correspondence between the possible density matrices of a single qubit and the points on the *unit 3-ball* $0 \leq |\vec{P}| \leq 1$. This ball is usually called the *Bloch sphere* (although of course it is really a ball, not a sphere).

The boundary $\left(|\vec{P}| = 1\right)$ of the ball (which really is a sphere) contains the density matrices with vanishing determinant. Since $\mathrm{tr}\boldsymbol{\rho} = 1$, these density matrices must have the eigenvalues 0 and 1. They are one-dimensional projectors, and hence pure states. We have already seen that every pure state of a single qubit is of the form $|\psi(\theta,\varphi)\rangle$ and can be envisioned as a spin pointing in the (θ,φ) direction. Indeed using the property

$$(\hat{n}\cdot\vec{\boldsymbol{\sigma}})^2 = \mathbf{1}, \tag{2.63}$$

where $\hat{n}$ is a unit vector, we can easily verify that the pure-state density matrix

$$\boldsymbol{\rho}(\hat{n}) = \frac{1}{2}\left(\mathbf{1} + \hat{n}\cdot\vec{\boldsymbol{\sigma}}\right) \tag{2.64}$$

satisfies the property

$$(\hat{n}\cdot\vec{\boldsymbol{\sigma}})\,\vec{\boldsymbol{\rho}}(\hat{n}) = \boldsymbol{\rho}(\hat{n})\,(\hat{n}\cdot\vec{\boldsymbol{\sigma}}) = \boldsymbol{\rho}(\hat{n}), \tag{2.65}$$

and, therefore is the projector

$$\boldsymbol{\rho}(\hat{n}) = |\psi(\hat{n})\rangle\langle\psi(\hat{n})|; \tag{2.66}$$

that is, $\hat{n}$ is the direction along which the spin is pointing up. Alternatively, from the expression

$$|\psi(\theta,\phi)\rangle = \begin{pmatrix} e^{-i\varphi/2} & \cos\frac{\theta}{2} \\ e^{i\varphi/2} & \sin\frac{\theta}{2}\end{pmatrix}, \tag{2.67}$$

we may compute directly that

$$\begin{aligned}\rho(\theta,\phi) &= |\psi(\theta,\phi)\rangle\langle\psi(\theta,\phi)|\\ = \begin{pmatrix} \cos^2\frac{\theta}{2} & \cos\frac{\theta}{2}\sin\frac{\theta}{2}e^{-i\varphi} \\ \cos\frac{\theta}{2}\sin\frac{\theta}{2}e^{i\varphi} & \sin^2\frac{\theta}{2}\end{pmatrix} &= \frac{1}{2}\mathbf{1} + \frac{1}{2}\begin{pmatrix}\cos\theta & \sin\theta e^{-i\varphi} \\ \sin\theta e^{i\varphi} & -\cos\theta\end{pmatrix}\\ &= \frac{1}{2}(\mathbf{1} + \hat{n}\cdot\vec{\boldsymbol{\sigma}})\end{aligned} \tag{2.68}$$

where $\hat{n} = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$. One nice property of the Bloch parametrization of the pure states is that while $|\psi(\theta,\varphi)\rangle$ has an arbitrary overall phase that has no physical significance, there is no phase ambiguity in the density matrix $\boldsymbol{\rho}(\theta,\varphi) = |\psi(\theta,\varphi)\rangle\langle\psi(\theta,\varphi)|$; all the parameters in $\boldsymbol{\rho}$ have a physical meaning.

From the property

$$\frac{1}{2}\mathrm{tr}\;\boldsymbol{\sigma}_i\boldsymbol{\sigma}_j = \delta_{ij} \tag{2.69}$$

we see that

$$\langle\hat{n}\cdot\vec{\boldsymbol{\sigma}}\rangle_{\vec{P}} = \mathrm{tr}\left(\hat{n}\cdot\vec{\boldsymbol{\sigma}}\boldsymbol{\rho}(\vec{P})\right) = \hat{n}\cdot\vec{P}\;. \tag{2.70}$$

Thus the vector $\vec{P}$ in Eq. (2.62) parametrizes the *polarization* of the spin. If there are many identically prepared systems at our disposal, we can determine $\vec{P}$ (and hence the complete density matrix $\boldsymbol{\rho}(\vec{P})$) by measuring $\langle\hat{n}\cdot\vec{\boldsymbol{\sigma}}\rangle$ along each of three linearly independent axes.

2.3.3 Gleason's theorem

We arrived at the density matrix $\boldsymbol{\rho}$ and the expression $\mathrm{tr}(\mathbf{M}\boldsymbol{\rho})$ for the expectation value of an observable $\mathbf{M}$ by starting from our axioms of quantum mechanics, and then considering the description of a portion of a larger quantum system. But it is encouraging to know that the density matrix formalism is a very general feature in a much broader framework. This is the content of *Gleason's theorem* (1957).

Gleason's theorem starts from the premise that it is the task of quantum theory to assign consistent probabilities to all possible orthogonal projections in a Hilbert space (in other words, to all possible measurements of observables).

A state of a quantum system, then, is a mapping that take each projection ($\mathbf{E}^2 = \mathbf{E}$ and $\mathbf{E} = \mathbf{E}^\dagger$) to a nonnegative real number less than one:

$$\mathbf{E} \to p(\mathbf{E}); \quad 0 \le p(E) \le 1. \tag{2.71}$$

This mapping must have the properties:

(1) $p(\mathbf{0}) = 0$

(2) $p(\mathbf{1}) = 1$

(3) If $\mathbf{E}_1\mathbf{E}_2 = 0$, then $p(\mathbf{E}_1 + \mathbf{E}_2) = p(\mathbf{E}_1) + p(\mathbf{E}_2)$.

Here (3) is the crucial assumption. It says that (since projections on to mutually orthogonal spaces can be viewed as mutually exclusive alternatives) the probabilities assigned to mutually orthogonal projections must be additive. This assumption is very powerful, because there are so many different ways to choose $\mathbf{E}_1$ and $\mathbf{E}_2$. Roughly speaking, the first two assumptions say that whenever we make a measurement; (1) there is always an outcome, and (2) the probabilities of all possible outcomes sum to 1.

Under these assumptions, Gleason showed that for any such map, there is a hermitian, positive $\boldsymbol{\rho}$ with $\mathrm{tr}\boldsymbol{\rho} = 1$ such that

$$p(\mathbf{E}) = \mathrm{tr}(\boldsymbol{\rho}\mathbf{E}). \tag{2.72}$$

as long as the dimension of the Hilbert space is greater than 2. Thus, the density matrix formalism is really *necessary*, if we are to represent observables as self-adjoint operators in Hilbert space, and we are to consistently assign probabilities to all possible measurement outcomes. Crudely speaking, the requirement of additivity of probabilities for mutually exclusive outcomes is so strong that we are inevitably led to the linear expression eq. (2.72).

The case of a two-dimensional Hilbert space is special because there just are not enough mutually exclusive projections in two dimensions. All nontrivial projections are of the form

$$\mathbf{E}(\hat{n}) = \frac{1}{2}(\mathbf{1} + \hat{n} \cdot \vec{\boldsymbol{\sigma}}), \tag{2.73}$$

and

$$\mathbf{E}(\hat{n})\mathbf{E}(\hat{m}) = 0 \tag{2.74}$$

only for $\hat{m} = -\hat{n}$; therefore, any function $f(\hat{n})$ on the two-sphere such that $f(\hat{n}) + f(-\hat{n}) = 1$ satisfies the premises of Gleason's theorem, and there are many such functions. However, in three-dimensions, there are may more alternative ways to partition unity, so that Gleason's assumptions are far more powerful. The proof of the theorem will not be given here. See Peres, p. 190 ff, for a discussion.

2.3.4 Evolution of the density operator

So far, we have not discussed the time evolution of mixed states. In the case of a bipartite pure state governed by the usual axioms of quantum theory, let us suppose that the Hamiltonian on $\mathcal{H}_A \otimes \mathcal{H}_B$ has the form

$$\mathbf{H}_{AB} = \mathbf{H}_A \otimes \mathbf{1}_B + \mathbf{1}_A \otimes \mathbf{H}_B. \tag{2.75}$$

Under this assumption, there is no coupling between the two subsystems A and B, so that each evolves independently. The time evolution operator for the combined system

$$\mathbf{U}_{AB}(t) = \mathbf{U}_A(t) \otimes \mathbf{U}_B(t), \tag{2.76}$$

decomposes into separate unitary time evolution operators acting on each system.

In the Schrödinger picture of dynamics, then, an initial pure state $|\psi(0)\rangle_{AB}$ of the bipartite system given by eq. (2.57) evolves to

$$|\psi(t)\rangle_{AB} = \sum_{i,\mu} a_{i\mu}|i(t)\rangle_A \otimes |\mu(t)\rangle_B, \tag{2.77}$$

where

$$\begin{aligned} |i(t)\rangle_A &= U_A(t)|i(0)\rangle_A, \\ |\mu(t)\rangle_B &= U_B(t)|\mu(0)\rangle_B, \end{aligned} \tag{2.78}$$

define new orthonormal basis for $\mathcal{H}_A$ and $\mathcal{H}_B$ (since $U_A(t)$ and $U_B(t)$ are unitary). Taking the partial trace as before, we find

$$\begin{aligned} \boldsymbol{\rho}_A(t) &= \sum_{i,j,\mu} a_{i\mu}a^*_{j\nu}|i(t)\rangle_A\, {}_A\langle j(t)| \\ &= \mathbf{U}_A(t)\boldsymbol{\rho}_A(0)\mathbf{U}_A(t)^\dagger. \end{aligned} \tag{2.79}$$

Thus $\mathbf{U}_A(t)$, acting by conjugation, determines the time evolution of the density matrix.

In particular, in the basis in which $\boldsymbol{\rho}_A(0)$ is diagonal, we have

$$\boldsymbol{\rho}_A(t) = \sum_a p_a \mathbf{U}_A(t)|\psi_a(0)\rangle_A\ {}_A\langle\psi_a(0)|\mathbf{U}_A(t). \tag{2.80}$$

Eq. (2.80) tells us that the evolution of $\boldsymbol{\rho}_A$ is perfectly consistent with the ensemble interpretation. Each state in the ensemble evolves forward in time governed by $\mathbf{U}_A(t)$. If the state $|\psi_a(0)\rangle$ occurs with probability p_a at time 0, then $|\psi_a(t)\rangle$ occurs with probability p_a at the subsequent time t.

On the other hand, it should be clear that eq. (2.80) applies only under the assumption that systems A and B are not *coupled* by the Hamiltonian. Later, we will investigate how the density matrix evolves under more general conditions.

2.4 Schmidt decomposition

A bipartite pure state can be expressed in a standard form (*the Schmidt decomposition*) that is often very useful.

To arrive at this form, note that an arbitrary vector in $\mathcal{H}_A \otimes \mathcal{H}_B$ can be expanded as

$$|\psi\rangle_{AB} = \sum_{i,\mu} a_{i\mu}|i\rangle_A|\mu\rangle_B \equiv \sum_i |i\rangle_A|\tilde{i}\rangle_B. \tag{2.81}$$

Here $\{|i\rangle_A\}$ and $\{|\mu\rangle_B\}$ are orthonormal basis for $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively, but to obtain the second equality in eq. (2.81) we have defined

$$|\tilde{i}\rangle_B \equiv \sum_\mu a_{i\mu}|\mu\rangle_B. \tag{2.82}$$

Note that the $|\tilde{i}\rangle_B$'s need *not* be mutually orthogonal or normalized.

Now let's suppose that the $\{|i\rangle_A\}$ basis is chosen to be the basis in which $\boldsymbol{\rho}_A$ is diagonal,

$$\boldsymbol{\rho}_A = \sum_i p_i|i\rangle_A\ {}_A\langle i|. \tag{2.83}$$

We can also compute $\boldsymbol{\rho}_A$ by performing a partial trace,

$$\boldsymbol{\rho}_A = \mathrm{tr}_B(|\psi\rangle_{AB}\ {}_{AB}\langle\psi|)$$

$$= \mathrm{tr}_B(\sum_{ij}|i\rangle_A\ {}_A\langle j| \otimes |\tilde{i}\rangle_B\ {}_B\langle\tilde{j}|) = \sum_{ij} {}_B\langle\tilde{j}|\tilde{i}\rangle_B\ \left(|i\rangle_A\ {}_A\langle j|\right)\ . \tag{2.84}$$

We obtained the last equality in eq. (2.84) by noting that

$$\begin{aligned}\mathrm{tr}_B\left(|\tilde{i}\rangle_B\ {}_B\langle\tilde{j}|\right) &= \sum_k\ {}_B\langle k|\tilde{i}\rangle_B\ {}_B\langle\tilde{j}|k\rangle_B \\ &= \sum_k\ {}_B\langle\tilde{j}|k\rangle_B\ {}_B\langle k|\tilde{i}\rangle_B =\ {}_B\langle\tilde{j}|\tilde{i}\rangle_B,\end{aligned} \tag{2.85}$$

where $\{|k\rangle_B\}$ is an orthonormal basis for $\mathcal{H}_B$. By comparing eq. (2.83) and eq. (2.84), we see that

$${}_B\langle\tilde{j}|\tilde{i}\rangle_B = p_i\delta_{ij}. \tag{2.86}$$

Hence, it turns out that the $\{|\tilde{i}\rangle_B\}$ *are* orthogonal after all. We obtain orthonormal vectors by rescaling,

$$|i'\rangle_B = p_i^{-1/2}|i\rangle_B \tag{2.87}$$

(we may assume $p_i \neq 0$, because we will need eq. (2.87) only for i appearing in the sum eq. (2.83)), and therefore obtain the expansion

$$|\psi\rangle_{AB} = \sum_i \sqrt{p_i}|i\rangle_A|i'\rangle_B, \tag{2.88}$$

in terms of a *particular* orthonormal basis of $\mathcal{H}_A$ and $\mathcal{H}_B$.

Eq. (2.88) is the Schmidt decomposition of the bipartite pure state $|\psi\rangle_{AB}$. Any bipartite pure state can be expressed in this form, but of course the bases used depend on the pure state that is being expanded. In general, we can't simultaneously expand *both* $|\psi\rangle_{AB}$ and $|\varphi\rangle_{AB} \in \mathcal{H}_A \otimes \mathcal{H}_B$ in the form eq. (2.88) using the *same* orthonormal bases for $\mathcal{H}_A$ and $\mathcal{H}_B$.

Using eq. (2.88), we can also evaluate the partial trace over $\mathcal{H}_A$ to obtain

$$\boldsymbol{\rho}_B = \mathrm{tr}_A\left(|\psi\rangle_{AB}\ {}_{AB}\langle\psi|\right) = \sum_i p_i|i'\rangle_B\ {}_B\langle i'|. \tag{2.89}$$

We see that $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ have the *same nonzero eigenvalues*. Of course, there is no need for $\mathcal{H}_A$ and $\mathcal{H}_B$ to have the same dimension, so the number of *zero* eigenvalues of $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ can differ.

If $\boldsymbol{\rho}_A$ (and hence $\boldsymbol{\rho}_B$) have no degenerate eigenvalues other than zero, then the Schmidt decomposition of $|\psi\rangle_{AB}$ is essentially uniquely determined by $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$. We can diagonalize $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ to find the $|i\rangle_A$'s and $|i'\rangle_B$'s, and then we pair up the eigenstates of $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ with the same eigenvalue to obtain eq. (2.88). We have chosen the phases of our basis states so that no phases appear in the coefficients in the sum; the only remaining freedom is to redefine $|i\rangle_A$ and $|i'\rangle_B$ by multiplying by opposite phases (which of course leaves the expression eq. (2.88) unchanged).

But if $\boldsymbol{\rho}_A$ has degenerate nonzero eigenvalues, then we need more information than that provided by $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ to determine the Schmidt decomposition; we need to know which $|i'\rangle_B$ gets paired with each $|i\rangle_A$. For example, if both $\mathcal{H}_A$ and $\mathcal{H}_B$ are N-dimensional and $\mathbf{U}_{ij}$ is any $N \times N$ unitary matrix, then

$$|\psi\rangle_{AB} = \frac{1}{\sqrt{N}} \sum_{i,j=1}^{N} |i\rangle_A \mathbf{U}_{ij} |j'\rangle_B, \tag{2.90}$$

will yield $\boldsymbol{\rho}_A = \boldsymbol{\rho}_B = \frac{1}{N}\mathbf{1}$ when we take partial traces. Furthermore, we are free to apply simultaneous unitary transformations in $\mathcal{H}_A$ and $\mathcal{H}_B$,

$$|\psi\rangle_{AB} = \frac{1}{\sqrt{N}} \sum_{i} |i\rangle_A |i'\rangle_B = \frac{1}{\sqrt{N}} \sum_{ijk} \mathbf{U}^*_{ij} |j\rangle_A \mathbf{U}_{ik} |k'\rangle_B; \tag{2.91}$$

this preserves the state $|\psi\rangle_{AB}$, but illustrates that there is an ambiguity in the basis used when we express $|\psi\rangle_{AB}$ in the Schmidt form.

2.4.1 Entanglement

With any bipartite pure state $|\psi\rangle_{AB}$ we may associate a positive integer, the *Schmidt number*, which is the number of nonzero eigenvalues in $\boldsymbol{\rho}_A$ (or $\boldsymbol{\rho}_B$) and hence the number of terms in the Schmidt decomposition of $|\psi\rangle_{AB}$. In terms of this quantity, we can define what it means for a bipartite pure state to be *entangled*: $|\psi\rangle_{AB}$ is entangled (or nonseparable) if its Schmidt number is greater than one; otherwise, it is *separable* (or unentangled). Thus, a separable bipartite pure state is a direct product of pure states in $\mathcal{H}_A$ and $\mathcal{H}_B$,

$$|\psi\rangle_{AB} = |\varphi\rangle_A \otimes |\chi\rangle_B; \tag{2.92}$$

then the reduced density matrices $\boldsymbol{\rho}_A = |\varphi\rangle_A\,{}_A\langle\varphi|$ and $\boldsymbol{\rho}_B = |\chi\rangle_B\,{}_B\langle\chi|$ are pure. Any state that cannot be expressed as such a direct product is entangled; then $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ are mixed states.

One of our main goals this term will be to understand better the significance of entanglement. It is not strictly correct to say that subsystems A and B are *uncorrelated* if $|\psi\rangle_{AB}$ is separable; after all, the two spins in the separable state

$$|\uparrow\rangle_A |\uparrow\rangle_B, \tag{2.93}$$

are surely correlated – they are both pointing in the same direction. But the correlations between A and B in an entangled state have a different character than those in a separable state. Perhaps the critical difference is that *entanglement cannot be created locally.* The only way to entangle A and B is for the two subsystems to directly interact with one another.

We can prepare the state eq. (2.93) without allowing spins A and B to ever come into contact with one another. We need only send a (classical!) message to two preparers (Alice and Bob) telling both of them to prepare a spin pointing along the z-axis. But the only way to turn the state eq. (2.93) into an entangled state like

$$\frac{1}{\sqrt{2}} \left(|\uparrow\rangle_A |\uparrow\rangle_B + |\downarrow\rangle_A |\downarrow\rangle_B \right), \tag{2.94}$$

is to apply a *collective* unitary transformation to the state. Local unitary transformations of the form $U_A \otimes U_B$, and local measurements performed by Alice or Bob, *cannot increase the Schmidt number* of the two-qubit state, no matter how much Alice and Bob discuss what they do. To entangle two qubits, we *must* bring them together and allow them to interact.

As we will discuss later, it is also possible to make the distinction between entangled and separable bipartite *mixed* states. We will also discuss various ways in which local operations can modify the form of entanglement, and some ways that entanglement can be put to use.

2.5 Ambiguity of the ensemble interpretation

2.5.1 Convexity

Recall that an operator $\boldsymbol{\rho}$ acting on a Hilbert space $\mathcal{H}$ may be interpreted as a density operator if it has the three properties:

(1) $\boldsymbol{\rho}$ is self-adjoint.

(2) $\boldsymbol{\rho}$ is nonnegative.

(3) $\mathrm{tr}(\boldsymbol{\rho}) = 1$.

It follows immediately that, given two density matrices $\boldsymbol{\rho}_1$, and $\boldsymbol{\rho}_2$, we can always construct another density matrix as a convex linear combination of the two:

$$\boldsymbol{\rho}(\lambda) = \lambda\boldsymbol{\rho}_1 + (1-\lambda)\boldsymbol{\rho}_2 \tag{2.95}$$

is a density matrix for any real λ satisfying $0 \leq \lambda \leq 1$. We easily see that $\boldsymbol{\rho}(\lambda)$ satisfies (1) and (3) if $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ do. To check (2), we evaluate

$$\langle\psi|\boldsymbol{\rho}(\lambda)|\psi\rangle = \lambda\langle\psi|\boldsymbol{\rho}_1|\psi\rangle + (1-\lambda)\langle\psi|\boldsymbol{\rho}_2|\psi\rangle \geq 0; \tag{2.96}$$

$\langle\boldsymbol{\rho}(\lambda)\rangle$ is guaranteed to be nonnegative because $\langle\boldsymbol{\rho}_1\rangle$ and $\langle\boldsymbol{\rho}_2\rangle$ are. We have, therefore, shown that in a Hilbert space $\mathcal{H}$ of dimension N, the density operators are a *convex subset* of the real vector space of $N \times N$ hermitian matrices. (A subset of a vector space is said to be convex if the set contains the straight line segment connecting any two points in the set.)

Most density operators can be expressed as a sum of other density operators in many different ways. But the pure states are special in this regard – it is *not* possible to express a pure state as a convex sum of two other states. Consider a pure state $\boldsymbol{\rho} = |\psi\rangle\langle\psi|$, and let $|\psi_\perp\rangle$ denote a vector orthogonal to $|\psi\rangle$, $\langle\psi_\perp|\psi\rangle = 0$. Suppose that $\boldsymbol{\rho}$ can be expanded as in eq. (2.95); then

$$\begin{aligned}\langle\psi_\perp|\boldsymbol{\rho}|\psi_\perp\rangle = 0 = \lambda\langle\psi_\perp|\boldsymbol{\rho}_1|\psi_\perp\rangle& \\ + (1-\lambda)\langle\psi_\perp|\boldsymbol{\rho}_2|\psi_\perp\rangle&.\end{aligned} \tag{2.97}$$

Since the right hand side is a sum of two nonnegative terms, and the sum vanishes, both terms must vanish. If λ is not 0 or 1, we conclude that $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ are orthogonal to $|\psi_\perp\rangle$. But since $|\psi_\perp\rangle$ can be *any* vector orthogonal to $|\psi\rangle$, we conclude that $\boldsymbol{\rho}_1 = \boldsymbol{\rho}_2 = \boldsymbol{\rho}$.

The vectors in a convex set that cannot be expressed as a linear combination of other vectors in the set are called the *extremal points* of the set. We have just shown that the pure states are extremal points of the set of density matrices. Furthermore, *only* the pure states are extremal, because any mixed state can be written $\boldsymbol{\rho} = \sum_i p_i|i\rangle\langle i|$ in the basis in which it is diagonal, and so is a convex sum of pure states.

We have already encountered this structure in our discussion of the special case of the Bloch sphere. We saw that the density operators are a (unit) ball in the three-dimensional set of 2×2 hermitian matrices with unit trace. The ball is convex, and its extremal points are the points on the boundary. Similarly, the $N \times N$ density operators are a convex subset of the (N^2-1)-dimensional set of $N \times N$ hermitian matrices with unit trace, and the extremal points of the set are the pure states.

However, the 2×2 case is atypical in one respect: for $N > 2$, the points on the boundary of the set of density matrices are not necessarily pure states. The boundary of the set consists of all density matrices with at least one vanishing eigenvalue (since there are nearby matrices with negative eigenvalues). Such a density matrix need not be pure, for $N > 2$, since the number of nonvanishing eigenvalues can exceed one.

2.5.2 Ensemble preparation

The convexity of the set of density matrices has a simple and enlightening physical interpretation. Suppose that a preparer agrees to prepare one of two possible states; with probability λ, the state $\boldsymbol{\rho}_1$ is prepared, and with probability $1-\lambda$, the state $\boldsymbol{\rho}_2$ is prepared. (A random number generator might be employed to guide this choice.) To evaluate the expectation value of any observable $\mathbf{M}$, we average over *both* the choices of preparation *and* the outcome of the quantum measurement:

$$\begin{aligned}\langle\mathbf{M}\rangle &= \lambda\langle\mathbf{M}\rangle_1 + (1-\lambda)\langle\mathbf{M}\rangle_2 \\ &= \lambda\mathrm{tr}(\mathbf{M}\boldsymbol{\rho}_1) + (1-\lambda)\mathrm{tr}(\mathbf{M}\boldsymbol{\rho}_2) \\ &= \mathrm{tr}\left(\mathbf{M}\boldsymbol{\rho}(\lambda)\right).\end{aligned} \tag{2.98}$$

All expectation values are thus indistinguishable from what we would obtain if the state $\boldsymbol{\rho}(\lambda)$ had been prepared instead. Thus, we have an operational procedure, given methods for preparing the states $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$, for preparing any convex combination.

Indeed, for any mixed state $\boldsymbol{\rho}$, there are an infinite variety of ways to express $\boldsymbol{\rho}$ as a convex combination of other states, and hence an infinite variety of procedures we could employ to prepare $\boldsymbol{\rho}$, all of which have exactly the same consequences for any conceivable observation of the system. But a pure state is different; it can be prepared in only one way. (This is what is "pure" about a pure state.) Every pure state is an eigenstate of some observable, e.g., for the state $\boldsymbol{\rho} = |\psi\rangle\langle\psi|$, measurement of the projection $\mathbf{E} = |\psi\rangle\langle\psi|$ is guaranteed to have the outcome 1. (For example, recall that every pure state of a single qubit is "spin-up" along some axis.) Since $\boldsymbol{\rho}$ is the only state for which the outcome of measuring $\mathbf{E}$ is 1 with 100% probability, there is no way to reproduce this observable property by choosing one of several possible preparations. Thus, the preparation of a pure state is unambiguous (we can determine a unique preparation if we have many copies of the state to experiment with), but the preparation of a mixed state is always ambiguous.

How ambiguous is it? Since any $\boldsymbol{\rho}$ can be expressed as a sum of pure states, let's confine our attention to the question: in how many ways can a density operator be expressed as a convex sum of pure states? Mathematically, this is the question: in how many ways can $\boldsymbol{\rho}$ be written as a sum of *extremal* states?

As a first example, consider the "maximally mixed" state of a single qubit:

$$\boldsymbol{\rho} = \frac{1}{2}\mathbf{1}. \tag{2.99}$$

This can indeed be prepared as an ensemble of pure states in an infinite variety of ways. For example,

$$\boldsymbol{\rho} = \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\downarrow_z\rangle\langle\downarrow_z|, \tag{2.100}$$

so we obtain $\boldsymbol{\rho}$ if we prepare either $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$, each occurring with probability $\frac{1}{2}$. But we also have

$$\boldsymbol{\rho} = \frac{1}{2}|\uparrow_x\rangle\langle\uparrow_x| + \frac{1}{2}|\downarrow_x\rangle\langle\downarrow_x|, \tag{2.101}$$

so we obtain $\boldsymbol{\rho}$ if we prepare either $|\uparrow_x\rangle$ or $|\downarrow_x\rangle$, each occurring with probability $\frac{1}{2}$. Now the preparation procedures are undeniably *different*. Yet there is no possible way to tell the difference by making observations of the spin.

More generally, the point at the center of the Bloch ball is the sum of any two antipodal points on the sphere – preparing either $|\uparrow_{\hat{n}}\rangle$ or $|\downarrow_{\hat{n}}\rangle$, each occurring with probability $\frac{1}{2}$ will generate $\boldsymbol{\rho} = \frac{1}{2}\mathbf{1}$.

Only in the case where $\boldsymbol{\rho}$ has two (or more) degenerate eigenvalues will there be distinct ways of generating $\boldsymbol{\rho}$ from an ensemble of *mutually orthogonal* pure states, but there is no good reason to confine our attention to

ensembles of mutually orthogonal pure states. We may consider a point in the interior of the Bloch ball

$$\boldsymbol{\rho}(\vec{P}) = \frac{1}{2}(\mathbf{1} + \vec{P}\cdot\vec{\sigma}), \tag{2.102}$$

with $0 < |\vec{P}| < 1$, and it too can be expressed as

$$\boldsymbol{\rho}(\vec{P}) = \lambda\boldsymbol{\rho}(\hat{n}_1) + (1-\lambda)\boldsymbol{\rho}(\hat{n}_2), \tag{2.103}$$

if $\vec{P} = \lambda\hat{n}_1 + (1-\lambda)\hat{n}_2$ (or in other words, if $\vec{P}$ lies somewhere on the line segment connecting the points $\hat{n}_1$ and $\hat{n}_2$ on the sphere). Evidently, for any $\vec{P}$, there is a solution associated with any chord of the sphere that passes through the point $\vec{P}$; all such chords comprise a two-parameter family.

This highly ambiguous nature of the preparation of a mixed quantum state is one of the characteristic features of quantum information that contrasts sharply with classical probability distributions. Consider, for example, the case of a probability distribution for a single classical bit. The two extremal distributions are those in which either 0 or 1 occurs with 100% probability. *Any* probability distribution for the bit is a convex sum of these two extremal points. Similarly, if there are N possible states, there are N extremal distributions, and any probability distribution has a *unique* decomposition into extremal ones (the convex set of probability distributions is a *simplex*). If 0 occurs with 21% probability, 1 with 33% probability, and 2 with 46% probability, there is a unique preparation procedure that yields this probability distribution!

2.5.3 Faster than light?

Let's now return to our earlier viewpoint – that a mixed state of system A arises because A is *entangled* with system B – to further consider the implications of the ambiguous preparation of mixed states. If qubit A has density matrix

$$\boldsymbol{\rho}_A = \frac{1}{2}|\uparrow_z\rangle_A\,{}_A\langle\uparrow_z| + \frac{1}{2}|\downarrow_z\rangle_A\,{}_A\langle\downarrow_z|, \tag{2.104}$$

this density matrix could arise from an entangled bipartite pure state $|\psi\rangle_{AB}$ with the Schmidt decomposition

$$|\psi\rangle_{AB} = \frac{1}{\sqrt{2}}\left(|\uparrow_z\rangle_A|\uparrow_z\rangle_B + |\downarrow_z\rangle_A|\downarrow_z\rangle_B\right). \tag{2.105}$$

Therefore, the ensemble interpretation of $\boldsymbol{\rho}_A$ in which either $|\uparrow_z\rangle_A$ or $|\downarrow_z\rangle_A$ is prepared (each with probability $p=\frac{1}{2}$) can be realized by performing a measurement of qubit B. We measure qubit B in the $\{|\uparrow_z\rangle_B, |\downarrow_z\rangle_B\}$ basis; if the result $|\uparrow_z\rangle_B$ is obtained, we have prepared $|\uparrow_z\rangle_A$, and if the result $|\downarrow_?\rangle_B$ is obtained, we have prepared $|\downarrow_z\rangle_A$.

But as we have already noted, in this case, because $\boldsymbol{\rho}_A$ has degenerate eigenvalues, the Schmidt basis is not unique. We can apply simultaneous unitary transformations to qubits A and B (actually, if we apply U to A we must apply U^* to B) without modifying the bipartite pure state $|\psi\rangle_{AB}$. Therefore, for *any* unit 3-vector $\hat{n}$, $|\psi\rangle_{AB}$ has a Schmidt decomposition of the form

$$|\psi\rangle_{AB} = \frac{1}{\sqrt{2}}\left(|\uparrow_{\hat{n}}\rangle_A |\uparrow_{\hat{n}'}\rangle_B + |\downarrow_{\hat{n}}\rangle_A |\downarrow_{\hat{n}'}\rangle_B\right). \tag{2.106}$$

We see that by measuring qubit B in a suitable basis, we can realize *any* interpretation of $\boldsymbol{\rho}_A$ as an ensemble of two pure states.

Bright students, upon learning of this property, are sometimes inspired to suggest a mechanism for faster-than-light communication. Many copies of $|\psi\rangle_{AB}$ are prepared. Alice takes all of the A qubits to the Andromeda galaxy and Bob keeps all of the B qubits on earth. When Bob wants to send a one-bit message to Alice, he chooses to measure either $\boldsymbol{\sigma}_1$ or $\boldsymbol{\sigma}_3$ for all his spins, thus preparing Alice's spins in either the $\{|\uparrow_z\rangle_A, |\downarrow_z\rangle_A\}$ or $\{|\uparrow_x\rangle_A, |\downarrow_x\rangle_A\}$ ensembles.[1] To read the message, Alice immediately measures her spins to see which ensemble has been prepared.

But *exceptionally* bright students (or students who heard the previous lecture) can see the flaw in this scheme. Though the two preparation methods are surely different, both ensembles are described by precisely the same density matrix $\boldsymbol{\rho}_A$. Thus, there is no conceivable measurement Alice can make that will distinguish the two ensembles, and no way for Alice to tell what action Bob performed. The "message" is unreadable.

Why, then, do we confidently state that "the two preparation methods are surely different?" To qualm any doubts about that, imagine that Bob either (1) measures all of his spins along the $\hat{z}$-axis, or (2) measures all of his spins along the $\hat{x}$-axis, and then calls Alice on the intergalactic telephone. He does *not* tell Alice whether he did (1) or (2), but he does tell her the results of all his measurements: "the first spin was up, the second was down," etc. Now

[1] U is real in this case, so $U = U^*$ and $\hat{n} = \hat{n}'$.

Alice performs either (1) or (2) on *her* spins. If both Alice and Bob measured along the same axis, Alice will find that every single one of her measurement outcomes agrees with what Bob found. But if Alice and Bob measured along different (orthogonal) axes, then Alice will find *no correlation* between her results and Bob's. About half of her measurements agree with Bob's and about half disagree. If Bob promises to do either (1) or (2), and assuming no preparation or measurement errors, then Alice will know that Bob's action was different than hers (even though Bob never told her this information) as soon as one of her measurements disagrees with what Bob found. If all their measurements agree, then if many spins are measured, Alice will have very high statistical confidence that she and Bob measured along the same axis. (Even with occasional measurement errors, the statistical test will still be highly reliable if the error rate is low enough.) So Alice does have a way to distinguish Bob's two preparation methods, but in this case there is certainly no faster-than-light communication, because Alice had to receive Bob's phone call before she could perform her test.

2.5.4 Quantum erasure

We had said that the density matrix $\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1}$ describes a spin in an *incoherent* superposition of the pure states $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$. This was to be distinguished from *coherent* superpositions of these states, such as

$$|\uparrow_x, \downarrow_x\rangle = \frac{1}{2}\left(|\uparrow_z\rangle \pm |\downarrow_z\rangle\right) ; \tag{2.107}$$

in the case of a coherent superposition, the *relative phase* of the two states has observable consequences (distinguishes $|\uparrow_x\rangle$ from $|\downarrow_x\rangle$). In the case of an incoherent superposition, the relative phase is completely unobservable. The superposition becomes incoherent if spin A becomes entangled with another spin B, and spin B is inaccessible.

Heuristically, the states $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$ can *interfere* (the relative phase of these states can be observed) only if we have no information about whether the spin state is $|\uparrow_z\rangle_A$ or $|\downarrow_z\rangle_A$. More than that, interference can occur only if there is *in principle no possible way* to find out whether the spin is up or down along the z-axis. Entangling spin A with spin B destroys interference, (causes spin A to *decohere*) because it is possible in principle for us to determine if spin A is up or down along $\hat{z}$ by performing a suitable measurement of spin B.

But we have now seen that the statement that entanglement causes decoherence requires a qualification. Suppose that Bob measures spin B along the $\hat{x}$-axis, obtaining either the result $|\uparrow_x\rangle_B$ or $|\downarrow_x\rangle_B$, and that he sends his measurement result to Alice. *Now* Alice's spin is a pure state (either $|\uparrow_x\rangle_A$ or $|\downarrow_x\rangle_A$) and in fact a coherent superposition of $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$. We have managed to recover the purity of Alice's spin before the jaws of decoherence could close!

Suppose that Bob allows his spin to pass through a Stern–Gerlach apparatus oriented along the $\hat{z}$-axis. Well, of course, Alice's spin can't behave like a coherent superposition of $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$; all Bob has to do is look to see which way his spin moved, and he will know whether Alice's spin is up or down along $\hat{z}$. But suppose that Bob does not look. Instead, he carefully refocuses the two beams without maintaining any record of whether his spin moved up or down, and *then* allows the spin to pass through a second Stern–Gerlach apparatus oriented along the $\hat{x}$-axis. *This* time he looks, and communicates the result of his $\boldsymbol{\sigma}_1$ measurement to Alice. Now the coherence of Alice's spin has been restored!

This situation has been called a *quantum eraser*. Entangling the two spins creates a "measurement situation" in which the coherence of $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$ is lost because we can find out if spin A is up or down along $\hat{z}$ by observing spin B. But when we measure spin B along $\hat{x}$, this information is "erased." Whether the result is $|\uparrow_x\rangle_B$ or $|\downarrow_x\rangle_B$ does not tell us anything about whether spin A is up or down along $\hat{z}$, because Bob has been careful not to retain the "which way" information that he might have acquired by looking at the first Stern–Gerlach apparatus.[2] Therefore, it is possible again for spin A to behave like a coherent superposition of $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$ (and it does, *after* Alice hears about Bob's result).

We can best understand the quantum eraser from the ensemble viewpoint. Alice has many spins selected from an ensemble described by $\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1}$, and there is no way for her to observe interference between $|\uparrow_z\rangle_A$ and $|\downarrow_z\rangle_A$. When Bob makes his measurement along $\hat{x}$, a particular preparation of the ensemble is realized. However, this has no effect that Alice can perceive – her spin is *still* described by $\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1}$ as before. But, when Alice receives Bob's phone call, she can select a *subensemble* of her spins that are all in the pure state $|\uparrow_x\rangle_A$. The information that Bob sends allows Alice to distill purity from a maximally mixed state.

[2] One often says that the "welcher weg" information has been erased, because it sounds more sophisticated in German.

Another wrinkle on the quantum eraser is sometimes called *delayed choice*. This just means that the situation we have described is really completely symmetric between Alice and Bob, so it can't make any difference who measures first. (Indeed, if Alice's and Bob's measurements are spacelike separated events, there is no invariant meaning to which came first; it depends on the frame of reference of the observer.) Alice could measure all of her spins today (say along $\hat{x}$) before Bob has made his mind up how he will measure his spins. Next week, Bob can decide to "prepare" Alice's spins in the states $|\uparrow_{\hat{n}}\rangle_A$ and $|\downarrow_{\hat{n}}\rangle_A$ (that is the "delayed choice"). He then tells Alice which were the $|\uparrow_{\hat{n}}\rangle_A$ spins, and she can check her measurement record to verify that

$$\langle\sigma_1\rangle_{\hat{n}} = \hat{n}\cdot\hat{x} \ . \tag{2.108}$$

The results are the same, irrespective of whether Bob "prepares" the spins before or after Alice measures them.

We have claimed that the density matrix $\boldsymbol{\rho}_A$ provides a complete physical description of the state of subsystem A, because it characterizes all possible measurements that can be performed on A. One sometimes hears the objection[3] that the quantum eraser phenomenon demonstrates otherwise. Since the information received from Bob enables Alice to recover a pure state from the mixture, how can we hold that everything Alice can know about A is encoded in $\boldsymbol{\rho}_A$?

I don't think this is the right conclusion. Rather, I would say that quantum erasure provides yet another opportunity to recite our mantra: "Information is physical." The state $\boldsymbol{\rho}_A$ of system A is not the same thing as $\boldsymbol{\rho}_A$ accompanied by the information that Alice has received from Bob. This information (which attaches labels to the subensembles) changes the physical description. One way to say this mathematically is that we should include Alice's "state of knowledge" in our description. An ensemble of spins for which Alice has no information about whether each spin is up or down is a *different* physical state than an ensemble in which Alice knows which spins are up and which are down.[4]

[3] For example, from Roger Penrose in *Shadows of the Mind.*

[4] This "state of knowledge" need not really be the state of a human mind; any (inanimate) record that labels the subensemble will suffice.

2.5.5 The GHJW theorem

So far, we have considered the quantum eraser only in the context of a single qubit, described by an ensemble of equally probable mutually orthogonal states, ($i.e.$, $\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1}$). The discussion can be considerably generalized.

We have already seen that a mixed state of any quantum system can be realized as an ensemble of pure states in an infinite number of different ways. For a density matrix $\boldsymbol{\rho}_A$, consider one such realization:

$$\boldsymbol{\rho}_A = \sum_i p_i |\varphi_i\rangle_{A\ A}\langle\varphi_i|, \quad \sum p_i = 1. \tag{2.109}$$

Here the states $\{|\varphi_i\rangle_A\}$ are all normalized vectors, but we do *not* assume that they are mutually orthogonal. Nevertheless, $\boldsymbol{\rho}_A$ can be realized as an ensemble, in which each pure state $|\varphi_i\rangle_{A\ A}\langle\varphi_i|$ occurs with probability p_i.

Of course, for any such $\boldsymbol{\rho}_A$, we can construct a "purification" of $\boldsymbol{\rho}_A$, a bipartite pure state $|\Phi_1\rangle_{AB}$ that yields $\boldsymbol{\rho}_A$ when we perform a partial trace over $\mathcal{H}_B$. One such purification is of the form

$$|\Phi_1\rangle_{AB} = \sum_i \sqrt{p_i}|\varphi_i\rangle_A|\alpha_i\rangle_B, \tag{2.110}$$

where the vectors $|\alpha_i\rangle_B \in \mathcal{H}_B$ *are* mutually orthogonal and normalized,

$$_B\langle\alpha_i|\alpha_j\rangle_B = \delta_{ij}. \tag{2.111}$$

Clearly, then,

$$\mathrm{tr}_B\left(|\Phi_1\rangle_{AB\ AB}\langle\Phi_1|\right) = \boldsymbol{\rho}_A. \tag{2.112}$$

Furthermore, we can imagine performing an orthogonal measurement in system B that projects onto the $|\alpha_i\rangle_B$ basis.[5] The outcome $|\alpha_i\rangle_B$ will occur with probability p_i, and will prepare the pure state $|\varphi_i\rangle_{A\ A}\langle\varphi_i|$ of system A. Thus, given the purification $|\Phi\rangle_{AB}$ of $\boldsymbol{\rho}_A$, there is a measurement we can perform in system B that realizes the $|\varphi_i\rangle_A$ ensemble interpretation of $\boldsymbol{\rho}_A$. When the measurement outcome in B is known, we have successfully extracted one of the pure states $|\varphi_i\rangle_A$ from the mixture $\boldsymbol{\rho}_A$.

What we have just described is a generalization of preparing $|\uparrow_z\rangle_A$ by measuring spin B along $\hat{z}$ (in our discussion of two entangled qubits). But

[5] The $|\alpha_i\rangle_B$'s might not span $\mathcal{H}_B$, but in the state $|\Phi\rangle_{AB}$, measurement outcomes orthogonal to all the $|\alpha_i\rangle_B$'s never occur.

to generalize the notion of a quantum eraser, we wish to see that in the state $|\Phi_1\rangle_{AB}$, we can realize a *different* ensemble interpretation of $\boldsymbol{\rho}_A$ by performing a different measurement of B. So let

$$\boldsymbol{\rho}_A = \sum_\mu q_\mu |\psi_\mu\rangle_{A\ A}\langle\psi_\mu|, \tag{2.113}$$

be another realization of the same density matrix $\boldsymbol{\rho}_A$ as an ensemble of pure states. For this ensemble as well, there is a corresponding purification

$$|\Phi_2\rangle_{AB} = \sum_\mu \sqrt{q_\mu}|\psi_\mu\rangle_A \otimes |\beta_\mu\rangle_B, \tag{2.114}$$

where again the $\{|\beta_\mu\rangle_B$'s$\}$ are orthonormal vectors in $\mathcal{H}_B$. So in the state $|\Phi_2\rangle_{AB}$, we can realize the ensemble by performing a measurement in $\mathcal{H}_B$ that projects onto the $\{|\beta_\mu\rangle_B\}$ basis.

Now, how are $|\Phi_1\rangle_{AB}$ and $|\Phi_2\rangle_{AB}$ related? In fact, we can easily show that

$$|\Phi_1\rangle_{AB} = (\mathbf{1}_A \otimes \mathbf{U}_B)\,|\Phi_2\rangle_{AB}; \tag{2.115}$$

the two states differ by a unitary change of basis acting in $\mathcal{H}_B$ alone, or

$$|\Phi_1\rangle_{AB} = \sum_\mu \sqrt{q_\mu}|\psi_\mu\rangle_A|\gamma_\mu\rangle_B, \tag{2.116}$$

where

$$|\gamma_\mu\rangle_B = \mathbf{U}_B|\beta_\mu\rangle_B, \tag{2.117}$$

is yet another orthonormal basis for $\mathcal{H}_B$. We see, then, that there is a *single* purification $|\Phi_1\rangle_{AB}$ of $\boldsymbol{\rho}_A$, such that we can realize either the $\{|\varphi_i\rangle_A\}$ ensemble or $\{|\psi_\mu\rangle_A\}$ ensemble by choosing to measure the appropriate observable in system B!

Similarly, we may consider many ensembles that all realize $\boldsymbol{\rho}_A$, where the maximum number of pure states appearing in any of the ensembles is n. Then we may choose a Hilbert space $\mathcal{H}_B$ of dimension n, and a pure state $|\Phi\rangle_{AB} \in \mathcal{H}_A \otimes \mathcal{H}_B$, such that any one of the ensembles can be realized by measuring a suitable observable of B. This is the *GHJW* [6] *theorem*. It expresses the quantum eraser phenomenon in its most general form.

[6] For Gisin and Hughston, Jozsa, and Wootters.

In fact, the GHJW theorem is an almost trivial corollary to the Schmidt decomposition. Both $|\Phi_1\rangle_{AB}$ and $|\Phi_2\rangle_{AB}$ have a Schmidt decomposition, and because both yield the same $\boldsymbol{\rho}_A$ when we take the partial trace over B, these decompositions must have the form

$$|\Phi_1\rangle_{AB} = \sum_k \sqrt{\lambda_k}|k\rangle_A|k_1'\rangle_B,$$
$$|\Phi_2\rangle_{AB} = \sum_k \sqrt{\lambda_k}|k\rangle_A|k_2'\rangle_B, \tag{2.118}$$

where the λ_k's are the eigenvalues of $\boldsymbol{\rho}_A$ and the $|k\rangle_A$'s are the corresponding eigenvectors. But since $\{|k_1'\rangle_B\}$ and $\{|k_2'\rangle_B\}$ are both orthonormal bases for $\mathcal{H}_B$, there is a unitary $\mathbf{U}_B$ such that

$$|k_1'\rangle_B = \mathbf{U}_B|k_2'\rangle_B, \tag{2.119}$$

from which eq. (2.115) immediately follows.

In the ensemble of pure states described by Eq. (2.109), we would say that the pure states $|\varphi_i\rangle_A$ are superposed *incoherently* — an observer in system A cannot detect the relative phases of these states. Heuristically, the reason that these states cannot interfere is that it is possible in principle to find out which representative of the ensemble is actually realized by performing a measurement in system B, a projection onto the orthonormal basis $\{|\alpha_i\rangle_B\}$. However, by projecting onto the $\{|\gamma_\mu\rangle_B\}$ basis instead, and relaying the information about the measurement outcome to system A, we can extract one of the pure states $|\psi_\mu\rangle_A$ from the ensemble, even though this state may be a coherent superposition of the $|\varphi_i\rangle_A$'s. In effect, measuring B in the $\{|\gamma_\mu\rangle_B\}$ basis "erases" the "welcher weg" information (whether the state of A is $|\varphi_i\rangle_A$ or $|\varphi_j\rangle_A$). In this sense, the GHJW theorem characterizes the general quantum eraser. The moral, once again, is that *information is physical* — the information acquired by measuring system B, when relayed to A, changes the physical description of a state of A.

2.6 Summary

Axioms. The arena of quantum mechanics is a Hilbert space $\mathcal{H}$. The fundamental assumptions are:

(1) A *state* is a *ray* in $\mathcal{H}$.

(2) An *observable* is a *self-adjoint operator* on $\mathcal{H}$.

(3) A *measurement* is an orthogonal *projection*.

(4) *Time evolution* is *unitary*.

Density operator. But if we confine our attention to only a portion of a larger quantum system, assumptions (1)-(4) need not be satisfied. In particular, a quantum state is described not by a ray, but by a density operator $\boldsymbol{\rho}$, a nonnegative operator with unit trace. The density operator is *pure* (and the state can be described by a ray) if $\boldsymbol{\rho}^2 = \boldsymbol{\rho}$; otherwise, the state is *mixed*. An observable $\mathbf{M}$ has expectation value $\mathrm{tr}(\mathbf{M}\boldsymbol{\rho})$ in this state.

Qubit. A quantum system with a two-dimensional Hilbert space is called a *qubit*. The general density matrix of a qubit is

$$\boldsymbol{\rho}(\vec{P}) = \frac{1}{2}(\mathbf{1} + \vec{P}\cdot\vec{\boldsymbol{\sigma}}) \tag{2.120}$$

where $\vec{P}$ is a three-component vector of length $|\vec{P}| \le 1$. Pure states have $|\vec{P}| = 1$.

Schmidt decomposition. For any quantum system divided into two parts A and B (a *bipartite* system), the Hilbert space is a tensor product $\mathcal{H}_A \oplus \mathcal{H}_B$. For any pure state $|\psi\rangle_{AB}$ of a bipartite system, there are orthonormal bases $\{|i\rangle_A\}$ for $\mathcal{H}_A$ and $\{|i'\rangle_B\}$ for $\mathcal{H}_B$ such that

$$|\psi\rangle_{AB} = \sum_i \sqrt{p_i}|i\rangle_A|i'\rangle_B; \tag{2.121}$$

Eq. (2.121) is called the *Schmidt decomposition* of $|\psi\rangle_{AB}$. In a bipartite pure state, subsystems A and B separately are described by density operators $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$; it follows from eq. (2.121) that $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ have the same nonvanishing eigenvalues (the p_i's). The number of nonvanishing eigenvalues is called the *Schmidt number* of $|\psi\rangle_{AB}$. A bipartite pure state is said to be *entangled* if its Schmidt number is greater than one.

Ensembles. The density operators on a Hilbert space form a convex set, and the pure states are the *extremal points* of the set. A mixed state of a system A can be prepared as an *ensemble* of pure states in many different ways, all of which are experimentally indistinguishable if we observe system A alone. Given any mixed state $\boldsymbol{\rho}_A$ of system A, any preparation of $\boldsymbol{\rho}_A$ as an ensemble of pure states can be realized in principle by performing a measurement in another system B with which A is entangled. In fact given many such preparations of $\boldsymbol{\rho}_A$, there is a single entangled state of A and B such that any one of these preparations can be realized by measuring a suitable observable in B (the *GHJW theorem*). By measuring in system B and reporting the measurement outcome to system A, we can extract from the mixture a pure state chosen from one of the ensembles.

2.7 Exercises

2.1 Fidelity of a random guess

A single qubit (spin-$\frac{1}{2}$ object) is in an unknown *pure* state $|\psi\rangle$, selected at random from an ensemble uniformly distributed over the Bloch sphere. We guess at random that the state is $|\phi\rangle$. On the average, what is the *fidelity* F of our guess, defined by

$$F \equiv |\langle\phi|\psi\rangle|^2 \ . \tag{2.122}$$

2.2 Fidelity after measurement

After randomly selecting a one-qubit pure state as in the previous problem, we perform a measurement of the spin along the $\hat{z}$-axis. This measurement prepares a state described by the density matrix

$$\boldsymbol{\rho} = \mathbf{P}_\uparrow\langle\psi|\mathbf{P}_\uparrow|\psi\rangle + \mathbf{P}_\downarrow\langle\psi|\mathbf{P}_\downarrow|\psi\rangle \tag{2.123}$$

(where $\mathbf{P}_{\uparrow,\downarrow}$ denote the projections onto the spin-up and spin-down states along the $\hat{z}$-axis). On the average, with what fidelity

$$F \equiv \langle\psi|\boldsymbol{\rho}|\psi\rangle \tag{2.124}$$

does this density matrix represent the initial state $|\psi\rangle$? (The improvement in F compared to the answer to the previous problem is a crude measure of how much we learned by making the measurement.)

2.3 Schmidt decomposition

For the two-qubit state

$$\Phi = \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(\frac{1}{2}|\uparrow\rangle_B + \frac{\sqrt{3}}{2}|\downarrow\rangle_B\right) + \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(\frac{\sqrt{3}}{2}|\uparrow\rangle_B + \frac{1}{2}|\downarrow\rangle_B\right), \tag{2.125}$$

a. Compute $\boldsymbol{\rho}_A = \mathrm{tr}_B\left(|\Phi\rangle\langle\Phi|\right)$ and $\boldsymbol{\rho}_B = \mathrm{tr}_A\left(|\Phi\rangle\langle\Phi|\right)$.

b. Find the Schmidt decomposition of $|\Phi\rangle$.

2.4 Tripartite pure state

Is there a Schmidt decomposition for an arbitrary *tripartite* pure state? That is if $|\psi\rangle_{ABC}$ is an arbitrary vector in $\mathcal{H}_A\otimes\mathcal{H}_B\otimes\mathcal{H}_C$, can we find orthonormal bases $\{|i\rangle_A\}$, $\{|i\rangle_B\}$, $\{|i\rangle_C\}$ such that

$$|\psi\rangle_{ABC} = \sum_i \sqrt{p_i}|i\rangle_A\otimes|i\rangle_B\otimes|i\rangle_C \ ? \tag{2.126}$$

Explain your answer.

2.5 Quantum correlations in a mixed state

Consider a density matrix for two qubits

$$\boldsymbol{\rho} = \frac{1}{8}\mathbf{1} + \frac{1}{2}|\psi^-\rangle\langle\psi^-| \ , \tag{2.127}$$

where $\mathbf{1}$ denotes the 4×4 unit matrix, and

$$|\psi^-\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow\rangle|\downarrow\rangle - |\downarrow\rangle|\uparrow\rangle\right) \ . \tag{2.128}$$

Suppose we measure the first spin along the $\hat{n}$ axis and the second spin along the $\hat{m}$ axis, where $\hat{n}\cdot\hat{m} = \cos\theta$. What is the probability that both spins are "spin-up" along their respective axes?

Chapter 3

Foundations II: Measurement and Evolution

3.1 Orthogonal Measurement and Beyond

3.1.1 Orthogonal Measurements

We would like to examine the properties of the *generalized* measurements that can be realized on system A by performing orthogonal measurements on a larger system that contains A. But first we will briefly consider how (orthogonal) measurements of an arbitrary observable can be achieved in principle, following the classic treatment of Von Neumann.

To measure an observable $\mathbf{M}$, we will modify the Hamiltonian of the world by turning on a coupling between that observable and a "pointer" variable that will serve as the apparatus. The coupling establishes entanglement between the eigenstates of the observable and the distinguishable states of the pointer, so that we can prepare an eigenstate of the observable by "observing" the pointer.

Of course, this is not a fully satisfying model of measurement because we have not explained how it is possible to measure the pointer. Von Neumann's attitude was that one can see that it is possible in principle to correlate the state of a microscopic quantum system with the value of a macroscopic classical variable, and we may take it for granted that we can perceive the value of the classical variable. A more complete explanation is desirable and possible; we will return to this issue later.

We may think of the pointer as a particle that propagates freely apart

from its tunable coupling to the quantum system being measured. Since we intend to measure the position of the pointer, it should be prepared initially in a wavepacket state that is narrow in position space — but not too narrow, because a vary narrow wave packet will spread too rapidly. If the initial width of the wave packet is Δx, then the uncertainty in it velocity will be of order $\Delta v = \Delta p/m \sim \hbar/m\Delta x$, so that after a time t, the wavepacket will spread to a width

$$\Delta x(t) \sim \Delta x + \frac{\hbar t}{m\Delta x}, \tag{3.1}$$

which is minimized for $[\Delta x(t)]^2 \sim [\Delta x]^2 \sim \hbar t/m$. Therefore, if the experiment takes a time t, the resolution we can achieve for the final position of the pointer is limited by

$$\Delta x \gtrsim (\Delta x)_{SQL} \sim \sqrt{\frac{\hbar t}{m}}, \tag{3.2}$$

the "standard quantum limit." We will choose our pointer to be sufficiently heavy that this limitation is not serious.

The Hamiltonian describing the coupling of the quantum system to the pointer has the form

$$\mathbf{H} = \mathbf{H}_0 + \frac{1}{2m}\mathbf{P}^2 + \lambda\mathbf{M}\mathbf{P}, \tag{3.3}$$

where $\mathbf{P}^2/2m$ is the Hamiltonian of the free pointer particle (which we will henceforth ignore on the grounds that the pointer is so heavy that spreading of its wavepacket may be neglected), H_0 is the unperturbed Hamiltonian of the system to be measured, and λ is a coupling constant that we are able to turn on and off as desired. The observable to be measured, $\mathbf{M}$, is coupled to the momentum $\mathbf{P}$ of the pointer.

If $\mathbf{M}$ does not commute with $\mathbf{H}_0$, then we have to worry about how the observable evolves during the course of the measurement. To simplify the analysis, let us suppose that either $[\mathbf{M}, \mathbf{H}_0] = 0$, or else the measurement is carried out quickly enough that the free evolution of the system can be neglected during the measurement procedure. Then the Hamiltonian can be approximated as $\mathcal{H} \simeq \lambda\mathbf{M}\mathbf{P}$ (where of course $[\mathbf{M}, \mathbf{P}] = 0$ because $\mathbf{M}$ is an observable of the system and $\mathbf{P}$ is an observable of the pointer), and the time evolution operator is

$$\mathbf{U}(t) \simeq \exp[-i\lambda t\mathbf{M}\mathbf{P}]. \tag{3.4}$$

Expanding in the basis in which $\mathbf{M}$ is diagonal,

$$\mathbf{M} = \sum_a |a\rangle M_a \langle a|, \tag{3.5}$$

we express $\mathbf{U}(t)$ as

$$\mathbf{U}(t) = \sum_a |a\rangle \exp[-i\lambda t M_a \mathbf{P}]\langle a|. \tag{3.6}$$

Now we recall that $\mathbf{P}$ generates a translation of the *position* of the pointer: $\mathbf{P} = -i\frac{d}{dx}$ in the position representation, so that $e^{-ix_o\mathbf{P}} = \exp\left(-x_o\frac{d}{dx}\right)$, and by Taylor expanding,

$$e^{-ix_o\mathbf{P}}\psi(x) = \psi(x - x_o); \tag{3.7}$$

In other words $e^{-ix_o\mathbf{P}}$ acting on a wavepacket translates the wavepacket by x_o. We see that if our quantum system starts in a superposition of $\mathbf{M}$ eigenstates, initially unentangled with the position-space wavepacket $|\psi(x)$ of the pointer, then after time t the quantum state has evolved to

$$\mathbf{U}(t)\left(\sum_a \alpha_a|a\rangle \otimes |\psi(x)\rangle\right)$$

$$= \sum_a \alpha_a|a\rangle \otimes |\psi(x - \lambda t M_a)\rangle; \tag{3.8}$$

the position of the pointer is now correlated with the value of the observable $\mathbf{M}$. If the pointer wavepacket is narrow enough for us to resolve all values of the M_a that occur ($\Delta x \lesssim \lambda t \Delta M_a$), then when we observe the position of the pointer (never mind how!) we will prepare an eigenstate of the observable. With probability $|\alpha_a|^2$, we will detect that the pointer has shifted its position by $\lambda t M_a$, in which case we will have prepared the $\mathbf{M}$ eigenstate $|a\rangle$. In the end, then, we conclude that the initial state $|\varphi\rangle$ or the quantum system is projected to $|a\rangle$ with probability $|\langle a|\varphi\rangle|^2$. This is Von Neumann's model of orthogonal measurement.

The classic example is the Stern–Gerlach apparatus. To measure $\boldsymbol{\sigma}_3$ for a spin-$\frac{1}{2}$ object, we allow the object to pass through a region of inhomogeneous magnetic field

$$B_3 = \lambda z. \tag{3.9}$$

The magnetic moment of the object is $\mu\vec{\boldsymbol{\sigma}}$, and the coupling induced by the magnetic field is

$$H = -\lambda\mu\mathbf{z}\boldsymbol{\sigma}_3. \tag{3.10}$$

In this case $\boldsymbol{\sigma}_3$ is the observable to be measured, coupled to the position $\mathbf{z}$ rather than the momentum of the pointer, but that's all right because $\mathbf{z}$ generates a translation of $\mathbf{P}_z$, and so the coupling imparts an *impulse* to the pointer. We can perceive whether the object is pushed up or down, and so project out the spin state $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$. Of course, by rotating the magnet, we can measure the observable $\hat{n}\cdot\vec{\boldsymbol{\sigma}}$ instead.

Our discussion of the quantum eraser has cautioned us that establishing the entangled state eq. (3.8) is *not* sufficient to explain why the measurement procedure prepares an eigenstate of $\mathbf{M}$. In principle, the measurement of the pointer could project out a peculiar superposition of position eigenstates, and so prepare the quantum system in a superposition of $\mathbf{M}$ eigenstates. To achieve a deeper understanding of the measurement process, we will need to explain why the position eigenstate basis of the pointer enjoys a privileged status over other possible bases.

If indeed we can couple any observable to a pointer as just described, and we can observe the pointer, then we can perform any conceivable orthogonal projection in Hilbert space. Given a set of operators $\{\mathbf{E}_a\}$ such that

$$\mathbf{E}_a = \mathbf{E}_a^\dagger, \quad \mathbf{E}_a\mathbf{E}_b = \delta_{ab}\mathbf{E}_a, \quad \sum_a \mathbf{E}_a = \mathbf{1}, \tag{3.11}$$

we can carry out a measurement procedure that will take a pure state $|\psi\rangle\langle\psi|$ to

$$\frac{\mathbf{E}_a|\psi\rangle\langle\psi|\mathbf{E}_a}{\langle\psi|\mathbf{E}_a|\psi\rangle} \tag{3.12}$$

with probability

$$\text{Prob}(a) = \langle\psi|\mathbf{E}_a|\psi\rangle. \tag{3.13}$$

The measurement outcomes can be described by a density matrix obtained by summing over all possible outcomes weighted by the probability of that outcome (rather than by choosing one particular outcome) in which case the measurement modifies the initial pure state according to

$$|\psi\rangle\langle\psi| \to \sum_a \mathbf{E}_a|\psi\rangle\langle\psi|\mathbf{E}_a. \tag{3.14}$$

This is the *ensemble* of pure states describing the measurement outcomes – it is the description we would use if we knew a measurement had been performed, but we did not know the result. Hence, the initial pure state has become a mixed state unless the initial state happened to be an eigenstate of the observable being measured. If the initial state before the measurement were a mixed state with density matrix $\boldsymbol{\rho}$, then by expressing $\boldsymbol{\rho}$ as an ensemble of pure states we find that the effect of the measurement is

$$\boldsymbol{\rho} \to \sum_a \mathbf{E}_a \boldsymbol{\rho} \mathbf{E}_a. \tag{3.15}$$

3.1.2 Generalized measurement

We would now like to generalize the measurement concept beyond these orthogonal measurements considered by Von Neumann. One way to arrive at the idea of a generalized measurement is to suppose that our system A is extended to a tensor product $\mathcal{H}_A \otimes \mathcal{H}_B$, and that we perform orthogonal measurements in the tensor product, which will not necessarily be orthogonal measurements in A alone. At first we will follow a somewhat different course that, while not as well motivated physically, is simpler and more natural from a mathematical view point.

We will suppose that our Hilbert space $\mathcal{H}_A$ is part of a larger space that has the structure of a *direct sum*

$$\mathcal{H} = \mathcal{H}_A \oplus \mathcal{H}_A^\perp. \tag{3.16}$$

Our observers who "live" in $\mathcal{H}_A$ have access only to observables with support in $\mathcal{H}_A$, observables $\mathbf{M}_A$ such that

$$\mathbf{M}_A|\psi^\perp\rangle = 0 = \langle\psi^\perp|\mathbf{M}_A, \tag{3.17}$$

for any $|\psi^\perp\rangle \in \mathcal{H}_A^\perp$. For example, in a two-qubit world, we might imagine that our observables have support only when the second qubit is in the state $|0\rangle_2$. Then $\mathcal{H}_A = \mathcal{H}_1 \otimes |0\rangle_2$ and $\mathcal{H}_A^\perp = \mathcal{H}_1 \otimes |1\rangle_2$, where $\mathcal{H}_1$ is the Hilbert space of qubit 1. (This situation may seem a bit artificial, which is what I meant in saying that the direct sum decomposition is not so well motivated.) Anyway, when we perform orthogonal measurement in $\mathcal{H}$, preparing one of a set of mutually orthogonal states, our observer will know only about the component of that state in his space $\mathcal{H}_A$. Since these components are not necessarily orthogonal in $\mathcal{H}_A$, he will conclude that the measurement prepares one of a set or non-orthogonal states.

Let $\{|i\rangle\}$ denote a basis for $\mathcal{H}_A$ and $\{|\mu\rangle\}$ a basis for $\mathcal{H}_A^\perp$. Suppose that the initial density matrix $\boldsymbol{\rho}_A$ has support in $\mathcal{H}_A$, and that we perform an orthogonal measurement in $\mathcal{H}$. We will consider the case in which each $\mathbf{E}_a$ is a one-dimensional projector, which will be general enough for our purposes. Thus, $\mathbf{E}_a = |u_a\rangle\langle u_a|$, where $|u_a\rangle$ is a normalized vector in $\mathcal{H}$. This vector has a unique orthogonal decomposition

$$|u_a\rangle = |\tilde{\psi}_a\rangle + |\tilde{\psi}_a^\perp\rangle, \tag{3.18}$$

where $|\tilde{\psi}_a\rangle$ and $|\tilde{\psi}_a^\perp\rangle$ are (unnormalized) vectors in $\mathcal{H}_A$ and $\mathcal{H}_A^\perp$ respectively. After the measurement, the new density matrix will be $|u_a\rangle\langle u_a|$ with probability $\langle u_a|\boldsymbol{\rho}_A|u_a\rangle = \langle\tilde{\psi}_a|\boldsymbol{\rho}_A|\tilde{\psi}_a\rangle$ (since $\boldsymbol{\rho}_A$ has no support on $\mathcal{H}_A^\perp$).

But to our observer who knows nothing of $\mathcal{H}_A^\perp$, there is no physical distinction between $|u_a\rangle$ and $|\tilde{\psi}_a\rangle$ (aside from normalization). If we write $|\tilde{\psi}_a\rangle = \sqrt{\lambda_a}|\psi_a\rangle$, where $|\psi_a\rangle$ is a normalized state, then for the observer limited to observations in $\mathcal{H}_A$, we might as well say that the outcome of the measurement is $|\psi_a\rangle\langle\psi_a|$ with probability $\langle\tilde{\psi}_a|\boldsymbol{\rho}_A|\tilde{\psi}_a\rangle$.

Let us define an operator

$$\mathbf{F}_a = \mathbf{E}_A \mathbf{E}_a \mathbf{E}_A = |\tilde{\psi}_a\rangle\langle\tilde{\psi}_a| = \lambda_a|\psi_a\rangle\langle\psi_a|, \tag{3.19}$$

(where $\mathbf{E}_A$ is the orthogonal projection taking $\mathcal{H}$ to $\mathcal{H}_A$). Then we may say that the outcome a has probability tr $\mathbf{F}_a\boldsymbol{\rho}$. It is evident that each $\mathbf{F}_a$ is hermitian and nonnegative, but the $\boldsymbol{F}_a$'s are not projections unless $\lambda_a = 1$. Furthermore

$$\sum_a \boldsymbol{F}_a = \boldsymbol{E}_A\left(\sum_a \boldsymbol{E}_a\right)\boldsymbol{E}_A = \boldsymbol{E}_A = \mathbf{1}_A; \tag{3.20}$$

the $\boldsymbol{F}_a$'s sum to the identity on $\mathcal{H}_A$

A partition of unity by nonnegative operators is called a *positive operator-valued measure* (POVM). (The term measure is a bit heavy-handed in our finite-dimensional context; it becomes more apt when the index a can be continually varying.) In our discussion we have arrived at the special case of a POVM by one-dimensional operators (operators with one nonvanishing eigenvalue). In the generalized measurement theory, each outcome has a probability that can be expressed as

$$\text{Prob}(a) = \text{tr } \boldsymbol{\rho}\boldsymbol{F}_a. \tag{3.21}$$

The positivity of $\boldsymbol{F}_a$ is necessary to ensure that the probabilities are positive, and $\sum_a \boldsymbol{F}_a = \boldsymbol{1}$ ensures that the probabilities sum to unity.

How does a general POVM affect the quantum state? There is not any succinct general answer to this question that is particularly useful, but in the case of a POVM by one-dimensional operators (as just discussed), where the outcome $|\psi_a\rangle\langle\psi_a|$ occurs with probability $\mathrm{tr}(\boldsymbol{F}_a\rho)$, summing over the outcomes yields

$$\begin{aligned}\boldsymbol{\rho} \to \boldsymbol{\rho}' &= \sum_a |\psi_a\rangle\langle\psi_a|(\lambda_a\langle\psi_a|\boldsymbol{\rho}|\psi_a\rangle)\\ &= \sum_a \left(\sqrt{\lambda_a}|\psi_a\rangle\langle\psi_a|\right)\boldsymbol{\rho}\left(\sqrt{\lambda_a}|\psi_a\rangle\langle\psi_a|\right)\\ &= \sum_a \sqrt{\boldsymbol{F}_a}\boldsymbol{\rho}\sqrt{\boldsymbol{F}_a},\end{aligned} \tag{3.22}$$

(which generalizes Von Neumann's $\sum_a \boldsymbol{E}_a\boldsymbol{\rho}\boldsymbol{E}_a$ to the case where the $\boldsymbol{F}_a$'s are not projectors). Note that $\mathrm{tr}\boldsymbol{\rho}' = \mathrm{tr}\boldsymbol{\rho} = 1$ because $\sum_a \boldsymbol{F}_a = 1$.

3.1.3 One-qubit POVM

For example, consider a single qubit and suppose that $\{\hat{n}_a\}$ are N unit 3-vectors that satisfy

$$\sum_a \lambda_a\hat{n}_a = 0, \tag{3.23}$$

where the λ_a's are positive real numbers, $0 < \lambda_a < 1$, such that $\sum_a \lambda_a = 1$. Let

$$\boldsymbol{F}_a = \lambda_a(\boldsymbol{1} + \hat{n}_a\cdot\vec{\boldsymbol{\sigma}}) = 2\lambda_a\boldsymbol{E}(\hat{n}_a), \tag{3.24}$$

(where $\boldsymbol{E}(\hat{n}_a)$ is the projection $|\uparrow_{\hat{n}_a}\rangle\langle\uparrow_{\hat{n}_a}|$). Then

$$\sum_a \boldsymbol{F}_a = (\sum_a \lambda_a)\boldsymbol{1} + (\sum_a \lambda_a\hat{n}_a)\cdot\vec{\boldsymbol{\sigma}} = \boldsymbol{1}; \tag{3.25}$$

hence the $\boldsymbol{F}$'s define a POVM.

In the case $N = 2$, we have $\hat{n}_1 + \hat{n}_2 = 0$, so our POVM is just an orthogonal measurement along the $\hat{n}_1$ axis. For $N = 3$, in the symmetric case $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. We have $\hat{n}_1 + \hat{n}_2 + \hat{n}_3 = 0$, and

$$\boldsymbol{F}_a = \frac{1}{3}(\boldsymbol{1} + \hat{n}_a\cdot\vec{\boldsymbol{\sigma}}) = \frac{2}{3}\boldsymbol{E}(\hat{n}_a). \tag{3.26}$$

3.1.4 Neumark's theorem

We arrived at the concept of a POVM by considering orthogonal measurement in a space larger than $\mathcal{H}_A$. Now we will reverse our tracks, showing that any POVM can be realized in this way.

So consider an arbitrary POVM with n one-dimensional positive operators $\boldsymbol{F}_a$ satisfying $\sum_{a=1}^n \boldsymbol{F}_a = \boldsymbol{1}$. We will show that this POVM can always be realized by extending the Hilbert space to a larger space, and performing orthogonal measurement in the larger space. This statement is called *Neumark's theorem.*[1]

To prove it, consider a Hilbert space $\mathcal{H}$ with $\dim \mathcal{H} = N$, and a POVM $\{\boldsymbol{F}_a\}, a = 1,\ldots,n$, with $n \geq N$. Each one-dimensional positive operator can be written

$$\boldsymbol{F}_a = |\tilde{\psi}_a\rangle\langle\tilde{\psi}_a|, \tag{3.27}$$

where the vector $|\tilde{\psi}_a\rangle$ is not normalized. Writing out the matrix elements explicitly, the property $\sum_a \boldsymbol{F}_a = \boldsymbol{1}$ becomes

$$\sum_{a=1}^n (F_a)_{ij} = \sum_{a=1}^n \tilde{\psi}^*_{ai}\tilde{\psi}_{aj} = \delta_{ij}. \tag{3.28}$$

Now let's change our perspective on eq. (3.28). Interpret the $(\psi_a)_i$'s not as $n \geq N$ vectors in an N-dimensional space, but rather an $N \leq n$ vectors $(\psi_i^T)_a$ in an n-dimensional space. Then eq. (3.28) becomes the statement that these N vectors form an orthonormal set. Naturally, it is possible to extend these vectors to an orthonormal basis for an n-dimensional space. In other words, there is an $n\times n$ matrix u_{ai}, with $u_{ai} = \tilde{\psi}_{ai}$ for $i = 1,2,\ldots,N$, such that

$$\sum_a u^*_{ai}u_{aj} = \delta_{ij}, \tag{3.29}$$

or, in matrix form $\boldsymbol{U}^\dagger\boldsymbol{U} = \boldsymbol{1}$. It follows that $\boldsymbol{U}\boldsymbol{U}^\dagger = \boldsymbol{1}$, since

$$\boldsymbol{U}(\boldsymbol{U}^\dagger\boldsymbol{U})|\psi\rangle = (\boldsymbol{U}\boldsymbol{U}^\dagger)\boldsymbol{U}|\psi\rangle = \boldsymbol{U}|\psi\rangle \tag{3.30}$$

[1] For a discussion of POVM's and Neumark's theorem, see A. Peres, *Quantum Theory: Concepts and Methods.*

for any vector $|\psi\rangle$, and (at least for finite-dimensional matrices) the range of $\boldsymbol{U}$ is the whole n-dimension space. Returning to the component notation, we have

$$\sum_j u_{aj} u_{bj}^* = \delta_{ab}, \tag{3.31}$$

so the $(u_a)_i$ are a set of n orthonormal vectors.[2]

Now suppose that we perform an orthogonal measurement in the space of dimension $n \geq N$ defined by

$$\boldsymbol{E}_a = |u_a\rangle\langle u_a|. \tag{3.32}$$

We have constructed the $|u_a\rangle$'s so that each has an orthogonal decomposition

$$|u_a\rangle = |\tilde{\psi}_a\rangle + |\tilde{\psi}_a^{\perp}\rangle; \tag{3.33}$$

where $|\tilde{\psi}_a\rangle \in \mathcal{H}$ and $|\tilde{\psi}_a^{\perp}\rangle \in \mathcal{H}^{\perp}$. By orthogonally projecting this basis onto $\mathcal{H}$, then, we recover the POVM $\{\boldsymbol{F}_a\}$. This completes the proof of Neumark's theorem.

To illustrate Neumark's theorem in action, consider again the POVM on a single qubit with

$$\boldsymbol{F}_a = \frac{2}{3}|\uparrow_{\hat{n}_a}\rangle\langle\uparrow_{\hat{n}_a}|, \tag{3.34}$$

$a = 1, 2, 3$, where $0 = \hat{n}_1 + \hat{n}_2 + \hat{n}_3$. According to the theorem, this POVM can be realized as an orthogonal measurement on a "qutrit," a quantum system in a three-dimensional Hilbert space.

Let $\hat{n}_1 = (0, 0, 1), \hat{n}_2 = (\sqrt{3}/2, 0, -1/2), \hat{n}_3 = (-\sqrt{3}/2, 0, 0, -1/2)$, and therefore, recalling that

$$|\theta, \varphi = 0\rangle = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix} \tag{3.35}$$

we may write the three vectors $|\tilde{\psi}_a\rangle = \sqrt{2/3}|\theta_a, \varphi = 0\rangle$ (where $\theta_1, \theta_2, \theta_3 = 0, 2\pi/3, 4\pi/3$) as

$$|\tilde{\psi}_1\rangle, |\tilde{\psi}_2\rangle, |\tilde{\psi}_3\rangle = \begin{pmatrix} \sqrt{2/3} \\ 0 \end{pmatrix}, \begin{pmatrix} \sqrt{1/6} \\ \sqrt{1/2} \end{pmatrix}, \begin{pmatrix} -\sqrt{1/6} \\ \sqrt{1/2} \end{pmatrix}. \tag{3.36}$$

[2] In other words, we have shown that if the rows of an $n \times n$ matrix are orthonormal, then so are the columns.

Now, we may interpret these three two-dimensional vectors as a 2×3 matrix, and as Neumark's theorem assured us, the two rows are orthonormal. Hence we can add one more orthonormal row:

$$|u_1\rangle, |u_2\rangle, |u_3\rangle = \begin{pmatrix} \sqrt{2/3} \\ 0 \\ \sqrt{1/3} \end{pmatrix}, \begin{pmatrix} \sqrt{1/6} \\ \sqrt{1/2} \\ -\sqrt{1/3} \end{pmatrix}, \begin{pmatrix} -\sqrt{1/6} \\ \sqrt{1/2} \\ \sqrt{1/3} \end{pmatrix}, \tag{3.37}$$

and we see (as the theorem also assured us) that the columns (the $|u_a\rangle$'s) are then orthonormal as well. If we perform an orthogonal measurement onto the $|u_a\rangle$ basis, an observer cognizant of only the two-dimensional subspace will conclude that we have performed the POVM $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$. We have shown that if our qubit is secretly two components of a *qutrit*, the POVM may be realized as orthogonal measurement of the qutrit.

3.1.5 Orthogonal measurement on a tensor product

A typical qubit harbors no such secret, though. To perform a generalized measurement, we will need to provide additional qubits, and perform joint orthogonal measurements on several qubits at once.

So *now* we consider the case of two (isolated) systems A and B, described by the tensor product $\mathcal{H}_A \oplus \mathcal{H}_B$. Suppose we perform an orthogonal measurement on the tensor product, with

$$\sum_a \boldsymbol{E}_a = \boldsymbol{1}, \tag{3.38}$$

where all $\boldsymbol{E}_a$'s are mutually orthogonal projectors. Let us imagine that the initial system of the quantum system is an "uncorrelated" tensor product state

$$\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B. \tag{3.39}$$

Then outcome a occurs with probability

$$\mathrm{Prob}(a) = \mathrm{tr}_{AB}[\boldsymbol{E}_a(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B)], \tag{3.40}$$

in which case the new density matrix will be

$$\rho'_{AB}(a) = \frac{\boldsymbol{E}_a(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B)\boldsymbol{E}_a}{\mathrm{tr}_{AB}[\boldsymbol{E}_a(\rho_A \otimes \rho_B)]}. \tag{3.41}$$

To an observer who has access only to system A, the new density matrix for that system is given by the partial trace of the above, or

$$\boldsymbol{\rho}'_A(a) = \frac{\mathrm{tr}_B[\boldsymbol{E}_a(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B)\boldsymbol{E}_a]}{\mathrm{tr}_{AB}[\boldsymbol{E}_a(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B)]}. \tag{3.42}$$

The expression eq. (3.40) for the probability of outcome a can also be written

$$\mathrm{Prob}(a) = \mathrm{tr}_A[\mathrm{tr}_B(\boldsymbol{E}_a(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B))] = \mathrm{tr}_A(\boldsymbol{F}_a\boldsymbol{\rho}_A); \tag{3.43}$$

If we introduce orthonormal bases $\{|i\rangle_A\}$ for $\mathcal{H}_A$ and $|\mu\rangle_B$ for $\mathcal{H}_B$, then

$$\sum_{ij\mu\nu}(E_a)_{j\nu,i\mu}(\rho_A)_{ij}(\rho_B)_{\mu\nu} = \sum_{ij}(F_a)_{ji}(\rho_A)_{ij}, \tag{3.44}$$

or

$$(F_a)_{ji} = \sum_{\mu\nu}(E_a)_{j\nu,i\mu}(\rho_B)_{\mu\nu}. \tag{3.45}$$

It follows from eq. (3.45) that each $\boldsymbol{F}_a$ has the properties:

(1) Hermiticity:

$$\begin{aligned}(F_a)^*_{ij} &= \sum_{\mu\nu}(E_a)^*_{i\nu,j\mu}(\rho_B)^*_{\mu\nu}\\ &= \sum_{\mu\nu}(E_a)_{j\mu,i\nu}(\rho_B)_{\nu\mu} = F_{ji}\end{aligned}$$

(because $\boldsymbol{E}_a$ and $\boldsymbol{\rho}_B$ are hermitian.

(2) Positivity:

In the basis that diagonalizes $\boldsymbol{\rho}_B = \sum_\mu p_\mu |\mu\rangle_B\ {}_B\langle\mu|$, ${}_A\langle\psi|\boldsymbol{F}_a|\psi\rangle_A = \sum_\mu p_\mu({}_A\langle\psi| \otimes\ {}_B\langle\mu|)\boldsymbol{E}_a(|\psi\rangle_A \otimes |\mu\rangle_B)$

$$\geq 0 \text{ (because } \boldsymbol{E}_a \text{ is positive).}$$

(3) Completeness:

$$\sum_a \boldsymbol{F}_a = \sum_\mu p_\mu\ {}_B\langle\mu|\sum_a \boldsymbol{E}_a|\mu\rangle_B = \mathbf{1}_A$$

$$\text{(because } \sum_a \boldsymbol{E}_a = \mathbf{1}_{AB} \text{ and tr } \boldsymbol{\rho}_B = 1).$$

But the $\boldsymbol{F}_a$'s need not be mutually orthogonal. In fact, the number of $\boldsymbol{F}_a$'s is limited only by the dimension of $\mathcal{H}_A \otimes \mathcal{H}_B$, which is greater than (and perhaps *much* greater than) the dimension of $\mathcal{H}_A$.

There is no simple way, in general, to express the final density matrix $\boldsymbol{\rho}'_A(a)$ in terms of $\boldsymbol{\rho}_A$ and $\boldsymbol{F}_a$. But let us disregard how the POVM changes the density matrix, and instead address this question: Suppose that $\mathcal{H}_A$ has dimension N, and consider a POVM with n one-dimensional nonnegative $\boldsymbol{F}_a$'s satisfying $\sum_{a=1}^n \boldsymbol{F}_a = \mathbf{1}_A$. Can we choose the space $\mathcal{H}_B$, density matrix $\boldsymbol{\rho}_B$ in $\mathcal{H}_B$, and projection operators $\boldsymbol{E}_a$ in $\mathcal{H}_A \otimes \mathcal{H}_B$ (where the number or $\boldsymbol{E}_a$'s may exceed the number of $\boldsymbol{F}_a$'s) such that the probability of outcome a of the orthogonal measurement satisfies[3]

$$\mathrm{tr}\ \boldsymbol{E}_a(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B) = \mathrm{tr}(\boldsymbol{F}_a\boldsymbol{\rho}_A)\ ? \tag{3.46}$$

(Never mind how the orthogonal projections modify $\boldsymbol{\rho}_A$!) We will consider this to be a "realization" of the POVM by orthogonal measurement, because we have no interest in what the state $\boldsymbol{\rho}'_A$ is for each measurement outcome; we are only asking that the *probabilities* of the outcomes agree with those defined by the POVM.

Such a realization of the POVM is indeed possible; to show this, we will appeal once more to Neumark's theorem. Each one-dimensional $\boldsymbol{F}_a$, $a = 1, 2, \ldots, n$, can be expressed as $\boldsymbol{F}_a = |\tilde{\psi}_a\rangle\langle\tilde{\psi}_a|$. According to Neumark, there are n orthonormal n-component vectors $|u_a\rangle$ such that

$$|u_a\rangle = |\tilde{\psi}_a\rangle + |\tilde{\psi}^\perp_a\rangle. \tag{3.47}$$

Now consider, to start with, the special case $n = rN$, where r is a positive integer. Then it is convenient to decompose $|\tilde{\psi}^\perp_a\rangle$ as a direct sum of $r-1$ N-component vectors:

$$|\tilde{\psi}^\perp_a\rangle = |\tilde{\psi}^\perp_{1,a}\rangle \oplus |\tilde{\psi}^\perp_{2,a}\rangle \oplus \cdots \oplus |\tilde{\psi}^\perp_{r-1,a}\rangle; \tag{3.48}$$

Here $|\tilde{\psi}^\perp_{1,a}\rangle$ denotes the first N components of $|\tilde{\psi}^\perp_a\rangle$, $|\tilde{\psi}^\perp_{2,a}\rangle$ denotes the next N components, etc. Then the orthonormality of the $|u_a\rangle$'s implies that

$$\delta_{ab} = \langle u_a|u_b\rangle = \langle\tilde{\psi}_a|\tilde{\psi}_b\rangle + \sum_{\mu=1}^{r-1}\langle\tilde{\psi}^\perp_{\mu,a}|\tilde{\psi}^\perp_{\mu,b}\rangle\ . \tag{3.49}$$

[3]If there are more $\boldsymbol{E}_a$'s than $\boldsymbol{F}_a$'s, all but n outcomes have probability *zero*.

Now we will choose $\mathcal{H}_B$ to have dimension r and we will denote an orthonormal basis for $\mathcal{H}_B$ by

$$\{|\mu\rangle_B\}, \quad \mu = 0, 1, 2, \ldots, r-1. \tag{3.50}$$

Then it follows from Eq. (3.49) that

$$|\Phi_a\rangle_{AB} = |\tilde{\psi}_a\rangle_A|0\rangle_B + \sum_{\mu=1}^{r-1} |\tilde{\psi}^{\perp}_{\mu,a}\rangle_A|\mu\rangle_B, \quad a = 1, 2, \ldots, n, \tag{3.51}$$

is an orthonormal basis for $\mathcal{H}_A \otimes \mathcal{H}_B$.

Now suppose that the state in $\mathcal{H}_A \otimes \mathcal{H}_B$ is

$$\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes |0\rangle_B\ {}_B\langle 0|, \tag{3.52}$$

and that we perform an orthogonal projection onto the basis $\{|\Phi_a\rangle_{AB}\}$ in $\mathcal{H}_A \otimes \mathcal{H}_B$. Then, since ${}_B\langle 0|\mu\rangle_B = 0$ for $\mu \neq 0$, the outcome $|\Phi_a\rangle_{AB}$ occurs with probability

$${}_{AB}\langle\Phi_a|\boldsymbol{\rho}_{AB}|\Phi_a\rangle_{AB} = {}_A\langle\tilde{\psi}_a|\boldsymbol{\rho}_A|\tilde{\psi}_a\rangle_A\ , \tag{3.53}$$

and thus,

$$\langle\Phi_a|\boldsymbol{\rho}_{AB}|\Phi_a\rangle_{AB} = \mathrm{tr}(\boldsymbol{F}_a\boldsymbol{\rho}_A). \tag{3.54}$$

We have indeed succeeded in "realizing" the POVM $\{\boldsymbol{F}_a\}$ by performing orthogonal measurement on $\mathcal{H}_A \otimes \mathcal{H}_B$. This construction is just as efficient as the "direct sum" construction described previously; we performed orthogonal measurement in a space of dimension $n = N \cdot r$.

If outcome a occurs, then the state

$$\boldsymbol{\rho}'_{AB} = |\Phi_a\rangle_{AB}\ {}_{AB}\langle\Phi_a|, \tag{3.55}$$

is prepared by the measurement. The density matrix seen by an observer who can probe only system A is obtained by performing a partial trace over $\mathcal{H}_B$,

$$\begin{aligned}\boldsymbol{\rho}'_A &= \mathrm{tr}_B\left(|\Phi_a\rangle_{AB}\ {}_{AB}\langle\Phi_a|\right)\\ &= |\tilde{\psi}_a\rangle_A\ {}_A\langle\tilde{\psi}_a| + \sum_{\mu=1}^{r-1}|\tilde{\psi}^{\perp}_{\mu,a}\rangle_A\ {}_A\langle\tilde{\psi}^{\perp}_{\mu,a}|\end{aligned} \tag{3.56}$$

which isn't quite the same thing as what we obtained in our "direct sum" construction. In any case, there are many possible ways to realize a POVM by orthogonal measurement and eq. (3.56) applies only to the particular construction we have chosen here.

Nevertheless, this construction really is perfectly adequate for realizing the POVM in which the state $|\psi_a\rangle_A\ {}_A\langle\psi_a|$ is prepared in the event that outcome a occurs. The hard part of implementing a POVM is assuring that outcome a arises with the desired probability. It is then easy to arrange that the *result* in the event of outcome a is the state $|\psi_a\rangle_A\ {}_A\langle\psi_a|$; if we like, once the measurement is performed and outcome a is found, we can simply throw $\boldsymbol{\rho}_A$ away and proceed to prepare the desired state! In fact, in the case of the projection onto the basis $|\Phi_a\rangle_{AB}$, we can complete the construction of the POVM by projecting system B onto the $\{|\mu\rangle_B\}$ basis, and communicating the result to system A. If the outcome is $|0\rangle_B$, then no action need be taken. If the outcome is $|\mu\rangle_B$, $\mu > 0$, then the state $|\tilde{\psi}^{\perp}_{\mu,a}\rangle_A$ has been prepared, which can then be rotated to $|\psi_a\rangle_A$.

So far, we have discussed only the special case $n = rN$. But if actually $n = rN - c$, $0 < c < N$, then we need only choose the final c components of $|\tilde{\psi}^{\perp}_{r-1,a}\rangle_A$ to be zero, and the states $|\Phi\rangle_{AB}$ will still be mutually orthogonal. To complete the orthonormal basis, we may add the c states

$$|e_i\rangle_A|r-1\rangle_B, \quad i = N-c+1, N-c+2, \ldots N\ ; \tag{3.57}$$

here e_i is a vector whose only nonvanishing component is the ith component, so that $|e_i\rangle_A$ is guaranteed to be orthogonal to $|\tilde{\psi}^{\perp}_{r-1,a}\rangle_A$. In this case, the POVM is realized as an orthogonal measurement on a space of dimension $rN = n + c$.

As an example of the tensor product construction, we may consider once again the single-qubit POVM with

$$\boldsymbol{F}_a = \frac{2}{3}|\uparrow_{\hat{n}_a}\rangle_A\ {}_A\langle\uparrow_{\hat{n}_a}|, \quad a = 1, 2, 3. \tag{3.58}$$

We may realize this POVM by introducing a second qubit B. In the two-qubit Hilbert space, we may project onto the orthonormal basis[4]

$$|\Phi_a\rangle = \sqrt{\frac{2}{3}}|\uparrow_{\hat{n}_a}\rangle_A|0\rangle_B + \sqrt{\frac{1}{3}}|0\rangle_A|1\rangle_B, \quad a = 1,2,3,$$
$$|\Phi_0\rangle = |1\rangle_A|1\rangle_B. \tag{3.59}$$

If the initial state is $\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes |0\rangle_B\ {}_B\langle 0|$, we have

$$\langle\Phi_a|\boldsymbol{\rho}_{AB}|\Phi_a\rangle = \frac{2}{3}{}_A\langle\uparrow_{\hat{n}_a}|\boldsymbol{\rho}_A|\uparrow_{\hat{n}_a}\rangle_A \tag{3.60}$$

so this projection implements the POVM on $\mathcal{H}_A$. (This time we performed orthogonal measurements in a four-dimensional space; we only needed three dimensions in our earlier "direct sum" construction.)

3.1.6 GHJW with POVM's

In our discussion of the GHJW theorem, we saw that by preparing a state

$$|\Phi\rangle_{AB} = \sum_\mu \sqrt{q_\mu}|\psi_\mu\rangle_A|\beta_\mu\rangle_B, \tag{3.61}$$

we can realize the ensemble

$$\boldsymbol{\rho}_A = \sum_\mu q_\mu|\psi_\mu\rangle_A\ {}_A\langle\psi_\mu|, \tag{3.62}$$

by performing orthogonal measurements on $\mathcal{H}_B$. Moreover, if $\dim\mathcal{H}_B = n$, then for this single pure state $|\Phi\rangle_{AB}$, we can realize any preparation of $\boldsymbol{\rho}_A$ as an ensemble of up to n pure states by measuring an appropriate observable on $\mathcal{H}_B$.

But we can now see that if we are willing to allow POVM's on $\mathcal{H}_B$ rather than orthogonal measurements only, then even for $\dim\mathcal{H}_B = N$, we can realize *any* preparation of $\boldsymbol{\rho}_A$ by choosing the POVM on $\mathcal{H}_B$ appropriately. The point is that $\boldsymbol{\rho}_B$ has support on a space that is at most dimension N. We may therefore rewrite $|\Phi\rangle_{AB}$ as

$$|\Phi\rangle_{AB} = \sum_\mu \sqrt{q_\mu}|\psi_\mu\rangle_A|\tilde{\beta}_\mu\rangle_B, \tag{3.63}$$

[4]Here the phase of $|\tilde{\psi}_2\rangle = \sqrt{2/3}|\uparrow_{\hat{n}_2}\rangle$ differs by -1 from that in eq. (3.36); it has been chosen so that $\langle\uparrow_{\hat{n}_a}|\uparrow_{\hat{n}_b}\rangle = -1/2$ for $a \neq b$. We have made this choice so that the coefficient of $|0\rangle_A|1\rangle_B$ is positive in all three of $|\Phi_1\rangle, |\Phi_2\rangle, |\Phi_3\rangle$.

where $|\tilde{\beta}_\mu\rangle_B$ is the result of orthogonally projecting $|\beta_\mu\rangle_B$ onto the support of $\boldsymbol{\rho}_B$. We may now perform the POVM on the support of $\boldsymbol{\rho}_B$ with $\boldsymbol{F}_\mu = |\tilde{\beta}_\mu\rangle_B\ {}_B\langle\tilde{\beta}_\mu|$, and thus prepare the state $|\psi_\mu\rangle_A$ with probability q_μ.

3.2 Superoperators

3.2.1 The operator-sum representation

We now proceed to the next step of our program of understanding the behavior of one part of a bipartite quantum system. We have seen that a pure state of the bipartite system may behave like a mixed state when we observe subsystem A alone, and that an orthogonal measurement of the bipartite system may be a (nonorthogonal) POVM on A alone. Next we ask, if a state of the bipartite system undergoes unitary evolution, how do we describe the evolution of A alone?

Suppose that the initial density matrix of the bipartite system is a tensor product state of the form

$$\boldsymbol{\rho}_A \otimes |0\rangle_B\ {}_B\langle 0|; \tag{3.64}$$

system A has density matrix $\boldsymbol{\rho}_A$, and system B is assumed to be in a pure state that we have designated $|0\rangle_B$. The bipartite system evolves for a finite time, governed by the unitary time evolution operator

$$\mathbf{U}_{AB}\left(\boldsymbol{\rho}_A \otimes |0\rangle_B\ {}_B\langle 0|\right)\mathbf{U}_{AB}. \tag{3.65}$$

Now we perform the partial trace over $\mathcal{H}_B$ to find the final density matrix of system A,

$$\boldsymbol{\rho}'_A = \mathrm{tr}_B\left(\mathbf{U}_{AB}\left(\boldsymbol{\rho}_A \otimes |0\rangle_B\ {}_B\langle 0|\right)\mathbf{U}^\dagger_{AB}\right)$$
$$= \sum_\mu {}_B\langle\mu|\mathbf{U}_{AB}|0\rangle_B\boldsymbol{\rho}_A\ {}_B\langle 0|\mathbf{U}_{AB}|\mu\rangle_B, \tag{3.66}$$

where $\{|\mu\rangle_B\}$ is an orthonormal basis for $\mathcal{H}_{B'}$ and ${}_B\langle\mu|\mathbf{U}_{AB}|0\rangle_B$ is an operator acting on $\mathcal{H}_A$. (If $\{|i\rangle_A \otimes |\mu\rangle_B\}$ is an orthonormal basis for $\mathcal{H}_A \otimes \mathcal{H}_B$, then ${}_B\langle\mu|\mathbf{U}_{AB}|\nu\rangle_B$ denotes the operator whose matrix elements are

$${}_A\langle i|\left({}_B\langle\mu|\mathbf{U}_{AB}|\nu\rangle_B\right)|j\rangle_A$$

$$= \left({}_A\langle i| \otimes {}_B\langle \mu| \right) \mathbf{U}_{AB} \left(|j\rangle_A \otimes |\nu\rangle_B \right).) \tag{3.67}$$

If we denote

$$\mathbf{M}_\mu = {}_B\langle \mu|\mathbf{U}_{AB}|0\rangle_B, \tag{3.68}$$

then we may express $\boldsymbol{\rho}'_A$ as

$$\$(\boldsymbol{\rho}_A) \equiv \boldsymbol{\rho}'_A = \sum_\mu \mathbf{M}_\mu \boldsymbol{\rho}_A \mathbf{M}_\mu^\dagger. \tag{3.69}$$

It follows from the unitarity of $\mathbf{U}_{AB}$ that the $\mathbf{M}_\mu$'s satisfy the property

$$\begin{aligned}\sum_\mu \mathbf{M}_\mu^\dagger \mathbf{M}_\mu &= \sum_\mu {}_B\langle 0|\mathbf{U}_{AB}^\dagger|\mu\rangle_B \; {}_B\langle \mu|\mathbf{U}_{AB}|0\rangle_B \\ &= {}_B\langle 0|\mathbf{U}_{AB}^\dagger \mathbf{U}_{AB}|0\rangle_B = \mathbf{1}_A.\end{aligned} \tag{3.70}$$

Eq. (3.69) defines a linear map \$ that takes linear operators to linear operators. Such a map, if the property in eq. (3.70) is satisfied, is called a *superoperator*, and eq. (3.69) is called the operator sum representation (or *Kraus* representation) of the superoperator. A superoperator can be regarded as a linear map that takes density operators to density operators, because it follows from eq. (3.69) and eq. (3.70) that $\boldsymbol{\rho}'_A$ is a density matrix if $\boldsymbol{\rho}_A$ is:

(1) $\boldsymbol{\rho}'_A$ is hermitian: $\boldsymbol{\rho}'^\dagger_A = \sum_\mu \mathbf{M}_\mu \boldsymbol{\rho}^\dagger_A \mathbf{M}^\dagger_\mu = \boldsymbol{\rho}_A$.

(2) $\boldsymbol{\rho}'_A$ has unit trace: $\mathrm{tr}\boldsymbol{\rho}'_A = \sum_\mu \mathrm{tr}(\boldsymbol{\rho}_A \mathbf{M}^\dagger_\mu \mathbf{M}_\mu) = \mathrm{tr}\boldsymbol{\rho}_A = 1$.

(3) $\boldsymbol{\rho}'_A$ is positive: ${}_A\langle \psi|\boldsymbol{\rho}'_A|\psi\rangle_A = \sum_\mu (\langle \psi|\mathbf{M}_\mu)\boldsymbol{\rho}_A(\mathbf{M}^\dagger_\mu|\psi\rangle) \geq 0$.

We showed that the operator sum representation in eq. (3.69) follows from the "unitary representation" in eq. (3.66). But furthermore, given the operator sum representation of a superoperator, it is always possible to construct a corresponding unitary representation. We choose $\mathcal{H}_B$ to be a Hilbert space whose dimension is at least as large as the number of terms in the operator sum. If $\{|\varphi_A\}$ is any vector in $\mathcal{H}_A$, the $\{|\mu\rangle_B\}$ are orthonormal states in $\mathcal{H}_B$, and $|0\rangle_B$ is some normalized state in $\mathcal{H}_B$, define the action of $\mathbf{U}_{AB}$ by

$$\mathbf{U}_{AB}\left(|\varphi\rangle_A \otimes |0\rangle_B\right) = \sum_\mu \mathbf{M}_\mu|\varphi\rangle_A \otimes |\mu\rangle_B. \tag{3.71}$$

This action is inner product preserving:

$$\begin{aligned}&\left(\sum_\nu {}_A\langle \varphi_2|\mathbf{M}^\dagger_\nu \otimes {}_B\langle \nu|\right)\left(\sum_\mu \mathbf{M}_\mu|\varphi_1\rangle_A \otimes |\mu\rangle_B\right) \\ &= {}_A\langle \varphi_2|\sum_\mu \mathbf{M}^\dagger_\mu \mathbf{M}_\mu|\varphi_1\rangle_A = {}_A\langle \varphi_2|\varphi_1\rangle_A;\end{aligned} \tag{3.72}$$

therefore, $\mathbf{U}_{AB}$ can be extended to a unitary operator acting on all of $\mathcal{H}_A \otimes \mathcal{H}_B$. Taking the partial trace we find

$$\begin{aligned}&\mathrm{tr}_B\left(\mathbf{U}_{AB}\left(|\varphi\rangle_A \otimes |0\rangle_B\right)\left({}_A\langle \varphi| \otimes {}_B\langle 0|\right)\mathbf{U}^\dagger_{AB}\right) \\ &= \sum_\mu \mathbf{M}_\mu\left(|\varphi\rangle_A \; {}_A\langle \varphi|\right)\mathbf{M}^\dagger_\mu.\end{aligned} \tag{3.73}$$

Since any $\boldsymbol{\rho}_A$ can be expressed as an ensemble of pure states, we recover the operator sum representation acting on an arbitrary $\boldsymbol{\rho}_A$.

It is clear that the operator sum representation of a given superoperator \$ is not unique. We can perform the partial trace in any basis we please. If we use the basis $\{{}_B\langle \nu'| = \sum_\mu U_{\nu\mu} \; {}_B\langle \mu|\}$ then we obtain the representation

$$\$(\boldsymbol{\rho}_A) = \sum_\nu \mathbf{N}_\nu \boldsymbol{\rho}_A \mathbf{N}^\dagger_\nu, \tag{3.74}$$

where $\mathbf{N}_\nu = U_{\nu\mu}\mathbf{M}_\mu$. We will see shortly that *any* two operator-sum representations of the same superoperator are always related this way.

Superoperators are important because they provide us with a formalism for discussing the general theory of *decoherence*, the evolution of pure states into mixed states. *Unitary* evolution of $\boldsymbol{\rho}_A$ is the special case in which there is only one term in the operator sum. If there are two or more terms, then there are pure initial states of $\mathcal{H}_A$ that become *entangled* with $\mathcal{H}_B$ under evolution governed by $\mathbf{U}_{AB}$. That is, if the operators M_1 and M_2 appearing in the operator sum are linearly independent, then there is a vector $|\varphi\rangle_A$ such that $|\tilde{\varphi}_1\rangle_A = M_1|\varphi\rangle_A$ and $|\tilde{\varphi}_2\rangle_A = M_2|\varphi\rangle_A$ are linearly independent, so that the state $|\tilde{\varphi}_1\rangle_A|1\rangle_B + |\tilde{\varphi}_2\rangle_A|2\rangle_B + \cdots$ has Schmidt number greater than one. Therefore, the pure state $|\varphi\rangle_A \; {}_A\langle \varphi|$ evolves to the *mixed* final state $\boldsymbol{\rho}'_A$.

Two superoperators $\$_1$ and $\$_2$ can be composed to obtain another superoperator $\$_2 \circ \$_1$; if $\$_1$ describes evolution from yesterday to today, and $\$_2$

describes evolution from today to tomorrow, then $\$_2 \circ \$_1$ describes the evolution from yesterday to tomorrow. But is the inverse of a superoperator also a superoperator; that is, is there a superoperator that describes the evolution from today to yesterday? In fact, you will show in a homework exercise that a superoperator is invertible only if it is unitary.

Unitary evolution operators form a group, but superoperators define a dynamical *semigroup*. When decoherence occurs, there is an arrow of time; even at the microscopic level, one can tell the difference between a movie that runs forwards and one running backwards. Decoherence causes an irrevocable loss of quantum information — once the (dead) cat is out of the bag, we can't put it back in again.

3.2.2 Linearity

Now we will broaden our viewpoint a bit and consider the essential properties that should be satisfied by any "reasonable" time evolution law for density matrices. We will see that any such law admits an operator-sum representation, so in a sense the dynamical behavior we extracted by considering part of a bipartite system is actually the most general possible.

A mapping $\$: \boldsymbol{\rho} \to \boldsymbol{\rho}'$ that takes an initial density matrix $\boldsymbol{\rho}$ to a final density matrix $\boldsymbol{\rho}'$ is a mapping of operators to operators that satisfies

(1) $\$$ *preserves hermiticity*: $\boldsymbol{\rho}'$ hermitian if $\boldsymbol{\rho}$ is.

(2) $\$$ is *trace preserving*: $\mathrm{tr}\boldsymbol{\rho}' = 1$ if $\mathrm{tr}\boldsymbol{\rho} = 1$.

(3) $\$$ is *positive*: $\boldsymbol{\rho}'$ is nonnegative if $\boldsymbol{\rho}$ is.

It is also customary to assume

(0) $\$$ is linear.

While (1), (2), and (3) really are necessary if $\boldsymbol{\rho}'$ is to be a density matrix, (0) is more open to question. Why linearity?

One possible answer is that nonlinear evolution of the density matrix would be hard to reconcile with any ensemble interpretation. If

$$\$\left(\boldsymbol{\rho}(\lambda)\right) \equiv \$\left(\lambda\boldsymbol{\rho}_1 + (1-\lambda)\boldsymbol{\rho}_2\right) = \lambda\$(\boldsymbol{\rho}_1) + (1-\lambda)\$(\boldsymbol{\rho}_2), \tag{3.75}$$

then time evolution is faithful to the probabilistic interpretation of $\boldsymbol{\rho}(\lambda)$: either (with probability λ) $\boldsymbol{\rho}_1$ was initially prepared and evolved to $\$(\boldsymbol{\rho}_1)$, or (with probability $1-\lambda$) $\boldsymbol{\rho}_2$ was initially prepared and evolved to $\$(\boldsymbol{\rho}_2)$. But a nonlinear $\$$ typically has consequences that are seemingly paradoxical.

Consider, for example, a single qubit evolving according to

$$\$(\boldsymbol{\rho}) = \exp\left[i\pi\boldsymbol{\sigma}_1\mathrm{tr}(\boldsymbol{\sigma}_1\boldsymbol{\rho})\right]\ \boldsymbol{\rho}\ \exp\left[-i\pi\boldsymbol{\sigma}_1\mathrm{tr}(\boldsymbol{\sigma}_1\boldsymbol{\rho})\right]\ . \tag{3.76}$$

One can easily check that $\$$ is positive and trace-preserving. Suppose that the initial density matrix is $\boldsymbol{\rho} = \frac{1}{2}\mathbf{1}$, realized as the ensemble

$$\boldsymbol{\rho} = \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\downarrow_z\rangle\langle\downarrow_z|. \tag{3.77}$$

Since $\mathrm{tr}(\boldsymbol{\sigma}_1\boldsymbol{\rho}) = 0$, the evolution of $\boldsymbol{\rho}$ is trivial, and both representatives of the ensemble are unchanged. If the spin was prepared as $|\uparrow_z\rangle$, it remains in the state $|\uparrow_z\rangle$.

But now imagine that, immediately after preparing the ensemble, we do nothing if the state has been prepared as $|\uparrow_z\rangle$, but we rotate it to $|\uparrow_x\rangle$ if it has been prepared as $|\downarrow_z\rangle$. The density matrix is now

$$\boldsymbol{\rho}' = \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\uparrow_x\rangle|\uparrow_x\rangle, \tag{3.78}$$

so that $\mathrm{tr}\boldsymbol{\rho}'\boldsymbol{\sigma}_1 = \frac{1}{2}$. Under evolution governed by $\$$, this becomes $\$(\boldsymbol{\rho}') = \boldsymbol{\sigma}_1\boldsymbol{\rho}'\boldsymbol{\sigma}_1$. In this case then, if the spin was prepared as $|\uparrow_z\rangle$, it evolves to the orthogonal state $|\downarrow_z\rangle$.

The state initially prepared as $|\uparrow_z\rangle$ evolves differently under these two scenarios. But what is the difference between the two cases? The difference was that *if* the spin was initially prepared as $|\downarrow_z\rangle$, we took different actions: doing nothing in case (1) but rotating the spin in case (2). Yet we have found that the spin behaves differently in the two cases, even if it was initially prepared as $|\uparrow_z\rangle$!

We are accustomed to saying that $\boldsymbol{\rho}$ describes two (or more) different alternative pure state preparations, only one of which is actually realized each time we prepare a qubit. But we have found that what happens if we prepare $|\uparrow_z\rangle$ actually *depends on what we would have done* if we had prepared $|\downarrow_x\rangle$ instead. It is no longer sensible, apparently, to regard the two possible preparations as mutually exclusive alternatives. Evolution of the alternatives actually depends on the other alternatives that supposedly were not realized.

Joe Polchinski has called this phenomenon the "Everett phone," because the different "branches of the wave function" seem to be able to "communicate" with one another.

Nonlinear evolution of the density matrix, then, can have strange, perhaps even absurd, consequences. Even so, the argument that nonlinear evolution should be excluded is not completely compelling. Indeed Jim Hartle has argued that there are versions of "generalized quantum mechanics" in which nonlinear evolution is permitted, yet a consistent probability interpretation can be salvaged. Nevertheless, we will follow tradition here and demand that \$ be linear.

3.2.3 Complete positivity

It would be satisfying were we able to conclude that any \$ satisfying (0) - (3) has an operator-sum representation, and so can be realized by unitary evolution of a suitable bipartite system. Sadly, this is not quite possible. Happily, though, it turns out that by adding one more rather innocuous sounding assumption, we *can* show that \$ has an operator-sum representation.

The additional assumption we will need (really a stronger version of (3)) is

(3') *\$ is completely positive.*

Complete positivity is defined as follows. Consider any possible extension of $\mathcal{H}_A$ to the tensor product $\mathcal{H}_A \otimes \mathcal{H}_B$; then $\$_A$ is completely positive on $\mathcal{H}_A$ if $\$_A \otimes I_B$ is positive for all such extensions.

Complete positivity is surely a reasonable property to demand on physical grounds. If we are studying the evolution of system A, we can never be certain that there is no system B, totally uncoupled to A, of which we are unaware. Complete positivity (combined with our other assumptions) is merely the statement that, if system A evolves and system B does not, any initial density matrix of the combined system evolves to another density matrix.

We will prove that assumptions (0), (1), (2), ($3'$) are sufficient to ensure that \$ is a superoperator (has an operator-sum representation). (Indeed, properties (0) - ($3'$) can be taken as an alternative definition of a superoperator.) Before proceeding with the proof, though, we will attempt to clarify the concept of complete positivity by giving an example of a positive operator that is not completely positive. The example is the transposition operator

$$T : \boldsymbol{\rho} \to \boldsymbol{\rho}^T. \tag{3.79}$$

T preserves the eigenvalues of $\boldsymbol{\rho}$ and so clearly is positive.

But is T *completely* positive (is $T_A \otimes I_B$ necessarily positive)? Let us choose $\dim(\mathcal{H}_B) = \dim(\mathcal{H}_A) = N$, and consider the maximally entangled state

$$|\Phi\rangle_{AB} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |i\rangle_A \otimes |i'\rangle_B, \tag{3.80}$$

where $\{|i\rangle_A\}$ and $\{|i'\rangle_B\}$ are orthonormal bases for $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively. Then

$$\begin{aligned} T_A \otimes I_B : \boldsymbol{\rho} = |\Phi\rangle_{AB}\ {}_{AB}\langle\Phi| &= \frac{1}{N}\sum_{i,j} (|i\rangle_A\ {}_A\langle j|) \otimes (|i'\rangle_B\ {}_B\langle j'|) \\ \to \boldsymbol{\rho}' &= \frac{1}{N}\sum_{i,j} (|j\rangle_A\ {}_A\langle i|) \otimes (|i'\rangle_B\ {}_B\langle j'|). \end{aligned} \tag{3.81}$$

We see that the operator $N\boldsymbol{\rho}'$ acts as

$$\begin{aligned} N\boldsymbol{\rho}' :& (\sum_i a_i |i\rangle_A) \otimes (\sum_j b_j |j'\rangle_B) \\ &\to (\sum_i a_i |i'\rangle_B) \otimes (\sum_j b_j |j\rangle_A), \end{aligned} \tag{3.82}$$

or

$$N\boldsymbol{\rho}'(|\varphi\rangle_A \otimes |\psi\rangle_B) = |\psi\rangle_A \otimes |\varphi\rangle_B. \tag{3.83}$$

Hence $N\boldsymbol{\rho}'$ is a *swap operator* (which squares to the identity). The eigenstates of $N\boldsymbol{\rho}'$ are states *symmetric* under the interchange $A \leftrightarrow B$, with eigenvalue 1, and *antisymmetric* states with eigenvalue -1. Since $\boldsymbol{\rho}'$ has negative eigenvalues, it is not positive, and (since $\boldsymbol{\rho}$ is certainly positive), therefore, $T_A \otimes I_B$ does not preserve positivity. We conclude that T_A, while positive, is *not* completely positive.

3.2.4 POVM as a superoperator

A unitary transformation that entangles A with B, followed by an orthogonal measurement of B, can be described as a POVM in A. In fact, the positive operators comprising the POVM can be constructed from the Kraus operators. If $|\varphi\rangle_A$ evolves as

$$|\varphi\rangle_A |0\rangle_B \to \sum_\mu \boldsymbol{M}_\mu |\varphi\rangle_A |\mu\rangle_B, \tag{3.84}$$

then the measurement in B that projects onto the $\{|\mu\rangle_E\}$ basis has outcome μ with probability

$$\mathrm{Prob}(\mu) = {}_A\langle\varphi|\boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu|\varphi\rangle_A . \tag{3.85}$$

Expressing $\boldsymbol{\rho}_A$ as an ensemble of pure states, we find the probability

$$\mathrm{Prob}(\mu) = tr(\boldsymbol{F}_\mu \boldsymbol{\rho}_A), \quad \boldsymbol{F}_\mu = \boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu , \tag{3.86}$$

for outcome μ; evidently $\boldsymbol{F}_\mu$ is positive, and $\sum_\mu \boldsymbol{F}_\mu = \mathbf{1}$ follows from the normalization of the Kraus operators. So this is indeed a realization of a POVM.

In particular, a POVM that modifies a density matrix according to

$$\boldsymbol{\rho} \to \sum_\mu \sqrt{\boldsymbol{F}_\mu}\boldsymbol{\rho}\sqrt{\boldsymbol{F}_\mu}, \tag{3.87}$$

is a special case of a superoperator. Since each $\sqrt{\boldsymbol{F}_\mu}$ is hermitian, the requirement

$$\sum_\mu \boldsymbol{F}_\mu = 1, \tag{3.88}$$

is just the operator-sum normalization condition. Therefore, the POVM has a "unitary representation;" there is a unitary $\mathbf{U}_{AB}$ that acts as

$$\mathbf{U}_{AB} : |\varphi\rangle_A \otimes |0\rangle_B \to \sum_\mu \sqrt{\boldsymbol{F}_\mu}|\varphi\rangle_A \otimes |\mu\rangle_B , \tag{3.89}$$

where $|\varphi\rangle_A$ is a pure state of system A. Evidently, then, by performing an orthogonal measurement in system B that projects onto the basis $\{|\mu\rangle_B\}$, we can realize the POVM that prepares

$$\boldsymbol{\rho}'_A = \frac{\sqrt{\boldsymbol{F}_\mu}\boldsymbol{\rho}_A\sqrt{\boldsymbol{F}_\mu}}{tr(\boldsymbol{F}_\mu \boldsymbol{\rho}_A)} \tag{3.90}$$

with probability

$$\mathrm{Prob}(\mu) = tr(\boldsymbol{F}_\mu \boldsymbol{\rho}_A). \tag{3.91}$$

This implementation of the POVM is not the most efficient possible (we require a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ of dimension $N \cdot n$, if the POVM has n possible outcomes) but it is in some ways the most convenient. A POVM is the most general measurement we can perform in system A by first entangling system A with system B, and then performing an orthogonal measurement in system B.

3.3 The Kraus Representation Theorem

Now we are almost ready to prove that any \$ satisfying the conditions $(0), (1), (2)$, and $(3')$ has an operator-sum representation (the Kraus representation theorem).[5] But first we will discuss a useful trick that will be employed in the proof. It is worthwhile to describe the trick separately, because it is of wide applicability.

The trick (which we will call the "relative-state method") is to *completely* characterize an operator $\mathbf{M}_A$ acting on $\mathcal{H}_A$ by describing how $\mathbf{M}_A \otimes \mathbf{1}_B$ acts on a single pure maximally entangled state[6] in $\mathcal{H}_A \otimes \mathcal{H}_B$ (where $\dim(\mathcal{H}_B) \geq \dim(\mathcal{H}_A) \equiv N$). Consider the state

$$|\tilde{\psi}\rangle_{AB} = \sum_{i=1}^N |i\rangle_A \otimes |i'\rangle_B \tag{3.92}$$

where $\{|i\rangle_A\}$ and $\{|i'\rangle_B\}$ are orthonormal bases of $\mathcal{H}_A$ and $\mathcal{H}_B$. (We have chosen to normalize $|\tilde{\psi}\rangle_{AB}$ so that ${}_{AB}\langle\tilde{\psi}|\tilde{\psi}\rangle_{AB} = N$; this saves us from writing various factors of $\sqrt{N}$ in the formulas below.) Note that any vector

$$|\varphi\rangle_A = \sum_i a_i|i\rangle_A, \tag{3.93}$$

in $\mathcal{H}_A$ may be expressed as a "partial" inner product

$$|\varphi\rangle_A = {}_B\langle\varphi^*|\tilde{\psi}\rangle_{AB}, \tag{3.94}$$

where

$$|\varphi^*\rangle_B = \sum_i a_i^*|i'\rangle_B . \tag{3.95}$$

We say that $|\varphi\rangle_A$ is the "relative state" of the "index state" $|\varphi^*\rangle_B$. The map

$$|\varphi\rangle_A \to |\varphi^*\rangle_B, \tag{3.96}$$

is evidently *antilinear*, and it is in fact an antiunitary map from $\mathcal{H}_A$ to a subspace of $\mathcal{H}_B$. The operator $\boldsymbol{M}_A \otimes \mathbf{1}_B$ acting on $|\tilde{\psi}\rangle_{AB}$ gives

$$(\boldsymbol{M}_A \otimes \mathbf{1}_B)|\tilde{\psi}\rangle_{AB} = \sum_i \boldsymbol{M}_A|i\rangle_A \otimes |i'\rangle_B . \tag{3.97}$$

[5] The argument given here follows B. Schumacher, quant-ph/9604023 (see Appendix A of that paper.).

[6] We say that the state $|\psi\rangle_{AB}$ is *maximally entangled* if $tr_B(|\psi\rangle_{AB}\,{}_{AB}\langle\psi|) \propto \mathbf{1}_A$.

From this state we can extract $\boldsymbol{M}_A|\psi\rangle_A$ as a relative state:

$$_B\langle\varphi^*|(\boldsymbol{M}_A\otimes \mathbf{1}_B)|\tilde{\psi}\rangle_{AB} = \boldsymbol{M}_A|\varphi\rangle_A. \tag{3.98}$$

We may interpret the relative-state formalism by saying that we can realize an ensemble of pure states in $\mathcal{H}_A$ by performing measurements in $\mathcal{H}_B$ on an entangled state – the state $|\varphi\rangle_A$ is prepared when the measurement in $\mathcal{H}_B$ has the outcome $|\varphi^*\rangle_B$. If we intend to apply an operator in $\mathcal{H}_A$, we have found that it makes no difference whether we first prepare the state and then apply the operator or we first apply the operator and then prepare the state. Of course, this conclusion makes physical sense. We could even imagine that the preparation and the operation are spacelike separated events, so that the temporal ordering has no invariant (observer-independent) meaning.

We will show that $\$_A$ has an operator-sum representation by applying the relative-state method to superoperators rather than operators. *Because* we assume that $\$_A$ is completely positive, we know that $\$_A\otimes I_B$ is positive. Therefore, if we apply $\$_A\otimes I_B$ to $\tilde{\boldsymbol{\rho}}_{AB} = |\tilde{\psi}\rangle_{AB}\ {}_{AB}\langle\tilde{\psi}|$, the result is a positive operator, an (unconventionally normalized) density matrix $\tilde{\boldsymbol{\rho}}'_{AB}$ in $\mathcal{H}_A\otimes\mathcal{H}_B$. Like any density matrix, $\tilde{\boldsymbol{\rho}}'_{AB}$ can be expanded as an ensemble of pure states. Hence we have

$$(\$_A\otimes I_B)(|\tilde{\psi}\rangle_{AB}\ {}_{AB}\langle\tilde{\psi}|) = \sum_\mu q_\mu|\tilde{\Phi}_\mu\rangle_{AB}\ {}_{AB}\langle\tilde{\Phi}_\mu|, \tag{3.99}$$

(where $q_\mu > 0$, $\sum_\mu q_\mu = 1$, and each $|\tilde{\Phi}_\mu\rangle$, like $|\tilde{\psi}\rangle_{AB}$, is normalized so that $\langle\tilde{\Phi}_\mu|\tilde{\Phi}_\mu\rangle = N$). Invoking the relative-state method, we have

$$\begin{aligned}\$_A(|\varphi\rangle_A\ {}_A\langle\varphi|) &= {}_B\langle\varphi^*|(\$_A\otimes I_B)(|\tilde{\psi}\rangle_{AB}\ {}_{AB}\langle\tilde{\psi}|)|\varphi^*\rangle_B\\ &= \sum_\mu q_\mu\ {}_B\langle\varphi^*|\tilde{\Phi}_\mu\rangle_{AB}\ {}_{AB}\langle\tilde{\Phi}_\mu|\varphi^*\rangle_B.\end{aligned} \tag{3.100}$$

Now we are almost done; we define an operator $\boldsymbol{M}_\mu$ on $\mathcal{H}_A$ by

$$\boldsymbol{M}_\mu : |\varphi\rangle_A \to \sqrt{q_\mu}\ {}_B\langle\varphi^*|\tilde{\Phi}_\mu\rangle_{AB}. \tag{3.101}$$

We can check that:

1. $\boldsymbol{M}_\mu$ is *linear*, because the map $|\varphi\rangle_A \to |\varphi^*\rangle_B$ is *antilinear*.

2. $\$_A(|\varphi\rangle_A\ {}_A\langle\varphi|) = \sum_\mu \boldsymbol{M}_\mu(|\varphi\rangle_A\ {}_A\langle\varphi|)\boldsymbol{M}_\mu^\dagger$, for any pure state $|\varphi\rangle_A\in\mathcal{H}_A$.

3. $\$_A(\boldsymbol{\rho}_A) = \sum_\mu \boldsymbol{M}_\mu\boldsymbol{\rho}_A\boldsymbol{M}_\mu^\dagger$ for any density matrix $\boldsymbol{\rho}_A$, because $\boldsymbol{\rho}_A$ can be expressed as an ensemble of pure states, and $\$_A$ is linear.

4. $\sum_\mu \boldsymbol{M}_\mu^\dagger\boldsymbol{M}_\mu = \mathbf{1}_A$, because $\$_A$ is trace preserving for any $\boldsymbol{\rho}_A$.

Thus, we have constructed an operator-sum representation of $\$_A$.

Put succinctly, the argument went as follows. Because $\$_A$ is completely positive, $\$_A\otimes I_B$ takes a maximally entangled density matrix on $\mathcal{H}_A\otimes\mathcal{H}_B$ to another density matrix. This density matrix can be expressed as an ensemble of pure states. With each of these pure states in $\mathcal{H}_A\otimes\mathcal{H}_B$, we may associate (via the relative-state method) a term in the operator sum.

Viewing the operator-sum representation this way, we may quickly establish two important corollaries:

How many Kraus operators? Each $\boldsymbol{M}_\mu$ is associated with a state $|\Phi_\mu\rangle$ in the ensemble representation of $\tilde{\boldsymbol{\rho}}'_{AB}$. Since $\tilde{\boldsymbol{\rho}}'_{AB}$ has a rank at most N^2 (where $N = \dim\mathcal{H}_A$), $\$_A$ always has an operator-sum representation with at most N^2 Kraus operators.

How ambiguous? We remarked earlier that the Kraus operators

$$\mathbf{N}_a = \boldsymbol{M}_\mu U_{\mu a}, \tag{3.102}$$

(where $U_{\mu a}$ is unitary) represent the same superoperator $\$$ as the $\boldsymbol{M}_\mu$'s. Now we can see that *any* two Kraus representations of $\$$ must always be related in this way. (If there are more $\mathbf{N}_a$'s than $\boldsymbol{M}_\mu$'s, then it is understood that some zero operators are added to the $\boldsymbol{M}_\mu$'s so that the two operator sets have the same cardinality.) This property may be viewed as a consequence of the GHJW theorem.

The relative-state construction described above established a $1-1$ correspondence between ensemble representations of the (unnormalized) density matrix $(\$_A\otimes I_B)\left(|\tilde{\psi}\rangle_{AB}\ {}_{AB}\langle\tilde{\psi}|\right)$ and operator-sum representations of $\$_A$. (We explicitly described how to proceed from the ensemble representation to the operator sum, but we can clearly go the other way, too: If

$$\$_A(|i\rangle_A\ {}_A\langle j|) = \sum_\mu \boldsymbol{M}_\mu|i\rangle_A\ {}_A\langle j|\boldsymbol{M}_\mu^\dagger, \tag{3.103}$$

then

$$\begin{aligned}(\$_A\otimes I_B)(|\tilde{\psi}\rangle_{AB}\ {}_{AB}\langle\tilde{\psi}|) &= \sum_{i,j}(\boldsymbol{M}_\mu|i\rangle_A|i'\rangle_B)({}_A\langle j|\boldsymbol{M}_\mu^\dagger\ {}_B\langle j'|)\\ &= \sum_\mu q_\mu|\tilde{\Phi}_\mu\rangle_{AB}\ {}_{AB}\langle\tilde{\Phi}_\mu|,\end{aligned} \tag{3.104}$$

where

$$\sqrt{q_\mu}|\tilde{\Phi}_\mu\rangle_{AB} = \sum_i \boldsymbol{M}_\mu |i\rangle_A |i'\rangle_B . \,) \tag{3.105}$$

Now consider two such ensembles (or correspondingly two operator-sum representations of $\$_A$), $\{\sqrt{q_\mu}|\tilde{\Phi}_\mu\rangle_{AB}\}$ and $\{\sqrt{p_a}|\tilde{\Upsilon}_a\rangle_{AB}\}$. For each ensemble, there is a corresponding "purification" in $\mathcal{H}_{AB}\otimes\mathcal{H}_C$:

$$\sum_\mu \sqrt{q_\mu}|\tilde{\Phi}_\mu\rangle_{AB}|\alpha_\mu\rangle_C$$

$$\sum_a \sqrt{p_a}|\tilde{\Upsilon}_a\rangle_{AB}|\beta_a\rangle_C , \tag{3.106}$$

where $\{(\alpha_\mu\rangle_C\}$ and $\{|\beta_a\rangle_C\}$ are two different orthonormal sets in $\mathcal{H}_c$. The GHJW theorem asserts that these two purifications are related by $\mathbf{1}_{AB}\otimes\mathbf{U}'_C$, a unitary transformation on $\mathcal{H}_C$. Therefore,

$$\sum_a \sqrt{p_a}|\tilde{\Upsilon}_a\rangle_{AB}|\beta_a\rangle_C$$

$$\begin{aligned} &= \sum_\mu \sqrt{q_\mu}|\tilde{\Phi}_\mu\rangle_{AB}\mathbf{U}'_C|\alpha_\mu\rangle_C \\ &= \sum_{\mu,a} \sqrt{q_\mu}|\tilde{\Phi}_\mu\rangle_{AB}U_{\mu a}|\beta_a\rangle_C \ , \end{aligned} \tag{3.107}$$

where, to establish the second equality we note that the orthonormal bases $\{|\alpha_\mu\rangle_C\}$ and $\{|\beta_a\rangle_C\}$ are related by a unitary transformation, and that a product of unitary transformations is unitary. We conclude that

$$\sqrt{p_a}|\tilde{\Upsilon}_a\rangle_{AB} = \sum_\mu \sqrt{q_\mu}|\tilde{\Phi}_\mu\rangle_{AB}U_{\mu a}, \tag{3.108}$$

(where $U_{\mu a}$ is unitary) from which follows

$$\boldsymbol{N}_a = \sum_\mu \boldsymbol{M}_\mu U_{\mu a}. \tag{3.109}$$

Remark. Since we have already established that we can proceed from an operator-sum representation of \$ to a unitary representation, we have now found that any "reasonable" evolution law for density operators on $\mathcal{H}_A$ can

be realized by a unitary transformation $\mathbf{U}_{AB}$ that acts on $\mathcal{H}_A\otimes\mathcal{H}_B$ according to

$$\mathbf{U}_{AB} : |\psi\rangle_A \otimes |0\rangle_B \to \sum_\mu |\varphi\rangle_A \otimes |\mu\rangle_B . \tag{3.110}$$

Is this result surprising? Perhaps it is. We may interpret a superoperator as describing the evolution of a system (A) that interacts with its environment (B). The general states of system plus environment are entangled states. But in eq. (3.110), we have assumed an initial state of A and B that is unentangled. Apparently though a real system is bound to be entangled with its surroundings, for the purpose of describing the evolution of its density matrix there is no loss of generality if we *imagine* that there is no pre-existing entanglement when we begin to track the evolution!

Remark: The operator-sum representation provides a very convenient way to express any completely positive \$. But a positive \$ does not admit such a representation if it is not completely positive. As far as I know, there is no convenient way, comparable to the Kraus representation, to express the most general *positive* \$.

3.4 Three Quantum Channels

The best way to familiarize ourselves with the superoperator concept is to study a few examples. We will now consider three examples (all interesting and useful) of superoperators for a single qubit. In deference to the traditions and terminology of (classical) communication theory. I will refer to these superoperators as *quantum channels*. If we wish, we may imagine that \$ describes the fate of quantum information that is transmitted with some loss of fidelity from a sender to a receiver. Or, if we prefer, we may imagine (as in our previous discussion), that the transmission is in time rather than space; that is, \$ describes the evolution of a quantum system that interacts with its environment.

3.4.1 Depolarizing channel

The *depolarizing channel* is a model of a decohering qubit that has particularly nice symmetry properties. We can describe it by saying that, with probability $1-p$ the qubit remains intact, while with probability p an "error" occurs. The error can be of any one of three types, where each type of error is equally likely. If $\{|0\rangle, |1\rangle\}$ is an orthonormal basis for the qubit, the three types of errors can be characterized as:

1. Bit flip error: $\begin{smallmatrix}|0\rangle\to|1\rangle\\|1\rangle\to|0\rangle\end{smallmatrix}$ or $|\psi\rangle \to \boldsymbol{\sigma}_1|\psi\rangle, \boldsymbol{\sigma}_1 = \left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$,

2. Phase flip error: $\begin{smallmatrix}|0\rangle\to|0\rangle\\|1\rangle\to-|1\rangle\end{smallmatrix}$ or $|\psi\rangle \to \boldsymbol{\sigma}_3|\psi\rangle, \boldsymbol{\sigma}_3 = \left(\begin{smallmatrix}1&0\\0&-1\end{smallmatrix}\right)$,

3. Both: $\begin{smallmatrix}|0\rangle\to+i|1\rangle\\|1\rangle\to-i|0\rangle\end{smallmatrix}$ or $|\psi\rangle \to \boldsymbol{\sigma}_2|\psi\rangle, \boldsymbol{\sigma}_2 = \left(\begin{smallmatrix}0&-i\\i&0\end{smallmatrix}\right)$.

If an error occurs, then $|\psi\rangle$ evolves to an ensemble of the three states $\boldsymbol{\sigma}_1|\psi\rangle, \boldsymbol{\sigma}_2|\psi\rangle, \boldsymbol{\sigma}_3|\psi\rangle$, all occuring with equal likelihood.

Unitary representation

The depolarizing channel can be represented by a unitary operator acting on $\mathcal{H}_A \otimes \mathcal{H}_E$, where $\mathcal{H}_E$ has dimension 4. (I am calling it $\mathcal{H}_E$ here to encourage you to think of the auxiliary system as the environment.) The unitary operator $\mathbf{U}_{AE}$ acts as

$$\begin{aligned}\mathbf{U}_{AE} &: |\psi\rangle_A \otimes |0\rangle_E \\ &\to \sqrt{1-p}|\psi\rangle \otimes |0\rangle_E + \sqrt{\frac{p}{3}}\bigg[\boldsymbol{\sigma}_1|\psi\rangle_A \otimes |1\rangle_E \\ &+ \boldsymbol{\sigma}_2|\psi\rangle \otimes |2\rangle_E + \boldsymbol{\sigma}_3|\psi\rangle \otimes |3\rangle_E\bigg].\end{aligned} \tag{3.111}$$

(Since $\mathbf{U}_{AE}$ is inner product preserving, it has a unitary extension to all of $\mathcal{H}_A \otimes \mathcal{H}_E$.) The environment evolves to one of four mutually orthogonal states that "keep a record" of what transpired; if we could only measure the environment in the basis $\{|\mu\rangle_E, \mu = 0,1,2,3\}$, we would know what kind of error had occurred (and we would be able to intervene and reverse the error).

Kraus representation

To obtain an operator-sum representation of the channel, we evaluate the partial trace over the environment in the $\{|\mu\rangle_E\}$ basis. Then

$$\boldsymbol{M}_\mu = {}_E\langle\mu|\boldsymbol{U}_{AE}|0\rangle_E, \tag{3.112}$$

so that

$$\boldsymbol{M}_0 = \sqrt{1-p}\;\mathbf{1},\;\; \boldsymbol{M}, = \sqrt{\frac{p}{3}}\;\boldsymbol{\sigma}_1,\;\; \boldsymbol{M}_2 = \sqrt{\frac{p}{3}}\;\boldsymbol{\sigma}_2,\;\; \boldsymbol{M}_3 = \sqrt{\frac{p}{3}}\;\boldsymbol{\sigma}_3. \tag{3.113}$$

Using $\boldsymbol{\sigma}_i^2 = \mathbf{1}$, we can readily check the normalization condition

$$\sum_\mu \boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu = \left[(1-p) + 3\frac{p}{3}\right]\mathbf{1} = \mathbf{1}. \tag{3.114}$$

A general initial density matrix $\boldsymbol{\rho}_A$ of the qubit evolves as

$$\begin{aligned}\boldsymbol{\rho} \to \boldsymbol{\rho}' = (1-p)\boldsymbol{\rho} + \\ \frac{p}{3}\left(\boldsymbol{\sigma}_1\boldsymbol{\rho}\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2\boldsymbol{\rho}\boldsymbol{\sigma}_2 + \boldsymbol{\sigma}_3\boldsymbol{\rho}\boldsymbol{\sigma}_3\right).\end{aligned} \tag{3.115}$$

where we are summing over the four (in principle distinguishable) ways that the environment could evolve.

Relative-state representation

We can also characterize the channel by describing how a maximally-entangled state of two qubits evolves, when the channel acts only on the first qubit. There are four mutually orthogonal maximally entangled states, which may be denoted

$$\begin{aligned}|\phi^+\rangle_{AB} &= \frac{1}{\sqrt{2}}(|00\rangle_{AB} + |11\rangle_{AB}),\\ |\phi^-\rangle_{AB} &= \frac{1}{\sqrt{2}}(|00\rangle_{AB} - |11\rangle_{AB}),\\ |\psi^+\rangle_{AB} &= \frac{1}{\sqrt{2}}(|01\rangle_{AB} + |10\rangle_{AB}),\\ |\psi^-\rangle_{AB} &= \frac{1}{\sqrt{2}}(|01\rangle_{AB} - |10\rangle_{AB}).\end{aligned} \tag{3.116}$$

If the initial state is $|\phi^+\rangle_{AB}$, then when the depolarizing channel acts on the first qubit, the entangled state evolves as

$$|\phi^+\rangle\langle\phi^+| \to (1-p)|\phi^+\rangle\langle\phi^+|$$

$$+\frac{p}{3}\left(|\psi^+\rangle\langle\psi^+| + |\psi^-\rangle\langle\psi^-| + |\phi^-\rangle\langle\phi^-|\right). \tag{3.117}$$

The "worst possible" quantum channel has $p = 3/4$ for in that case the initial entangled state evolves as

$$|\phi^+\rangle\langle\phi^+| \to \frac{1}{4}\left(|\phi^+\rangle\langle\phi^+| + |\phi^-\rangle\langle\phi^-| \right.$$
$$\left. +|\psi^+\rangle\langle\psi^+| + |\psi^-\rangle\langle\psi^-|\right) = \frac{1}{4}\mathbf{1}_{AB}; \tag{3.118}$$

it becomes the totally random density matrix on $\mathcal{H}_A \otimes \mathcal{H}_B$. By the relative-state method, then, we see that a pure state $|\varphi\rangle_A$ of qubit A evolves as

$$|\varphi\rangle_A\ {}_A\langle\varphi| \to\ {}_B\langle\varphi^*|2\left(\frac{1}{4}\mathbf{1}_{AB}\right)|\varphi^*\rangle_B = \frac{1}{2}\mathbf{1}_A; \tag{3.119}$$

it becomes the random density matrix on $\mathcal{H}_A$, irrespective of the value of the initial state $|\varphi\rangle_A$. It is as though the channel threw away the initial quantum state, and replaced it by completely random junk.

An alternative way to express the evolution of the maximally entangled state is

$$|\phi^+\rangle\langle\phi^+| \to \left(1 - \frac{4}{3}p\right)|\phi^+\rangle\langle\phi^+| + \frac{4}{3}p\left(\frac{1}{4}\mathbf{1}_{AB}\right). \tag{3.120}$$

Thus instead of saying that an error occurs with probability p, with errors of three types all equally likely, we could instead say that an error occurs with probability $4/3p$, where the error completely "randomizes" the state (at least we can say that for $p \leq 3/4$). The existence of two natural ways to define an "error probability" for this channel can sometimes cause confusion and misunderstanding.

One useful measure of how well the channel preserves the original quantum information is called the "entanglement fidelity" F_e. It quantifies how "close" the final density matrix is to the original maximally entangled state $|\phi^+\rangle$:

$$F_e = \langle\phi^+|\boldsymbol{\rho}'|\phi^+\rangle. \tag{3.121}$$

For the depolarizing channel, we have $F_e = 1 - p$, and we can interpret F_e as the probability that no error occured.

Block-sphere representation

It is also instructive to see how the depolarizing channel acts on the Bloch sphere. An arbitrary density matrix for a single qubit can be written as

$$\boldsymbol{\rho} = \frac{1}{2}\left(\mathbf{1} + \vec{P}\cdot\vec{\boldsymbol{\sigma}}\right), \tag{3.122}$$

where $\vec{P}$ is the "spin polarization" of the qubit. Suppose we rotate our axes so that $\vec{P} = P_3\hat{e}_3$ and $\boldsymbol{\rho} = \frac{1}{2}(\mathbf{1} + P_3\boldsymbol{\sigma}_3)$. Then, since $\boldsymbol{\sigma}_3\boldsymbol{\sigma}_3\boldsymbol{\sigma}_3 = \boldsymbol{\sigma}_3$ and $\boldsymbol{\sigma}_1\boldsymbol{\sigma}_3\boldsymbol{\sigma}_1 = -\boldsymbol{\sigma}_3 = \boldsymbol{\sigma}_2\boldsymbol{\sigma}_3\boldsymbol{\sigma}_2$, we find

$$\boldsymbol{\rho}' = \left(1 - p + \frac{p}{3}\right)\frac{1}{2}(\mathbf{1} + P_3\boldsymbol{\sigma}_3) + \frac{2p}{3}\frac{1}{2}(\mathbf{1} - P_3\boldsymbol{\sigma}_3), \tag{3.123}$$

or $P_3' = \left(1 - \frac{4}{3}p\right)P_3$. From the rotational symmetry, we see that

$$\vec{P}' = \left(1 - \frac{4}{3}p\right)\vec{P}, \tag{3.124}$$

irrespective of the direction in which P points. Hence, the Bloch sphere contracts uniformly under the action of the channel; the spin polarization is reduced by the factor $1 - \frac{4}{3}p$ (which is why we call it the depolarizing channel). This result was to be expected in view of the observation above that the spin is totally "randomized" with probability $\frac{4}{3}p$.

Invertibility?

Why do we say that the superoperator is not invertible? Evidently we can reverse a uniform contraction of the sphere with a uniform inflation. But the trouble is that the inflation of the Bloch sphere is not a superoperator, because it is not positive. Inflation will take values of $\vec{P}$ with $|\vec{P}| \leq 1$ to values with $|\vec{P}| > 1$, and so will take a density operator to an operator with a negative eigenvalue. Decoherence can shrink the ball, but no physical process can blow it up again! A superoperator running backwards in time is *not* a superoperator.

3.4.2 Phase-damping channel

Our next example is the *phase-damping channel*. This case is particularly instructive, because it provides a revealing caricature of decoherence in realistic physical situations, with all inessential mathematical details stripped away.

Unitary representation

A unitary representation of the channel is

$$
\begin{aligned}
|0\rangle_A|0\rangle_E &\to \sqrt{1-p}\,|0\rangle_A|0\rangle_E + \sqrt{p}\,|0\rangle_A|1\rangle_E,\\
|1\rangle_A|0\rangle_E &\to \sqrt{1-p}\,|1\rangle_A|0\rangle_E + \sqrt{p}\,|1\rangle_A|2\rangle_E.
\end{aligned} \tag{3.125}
$$

In this case, unlike the depolarizing channel, qubit A does not make any transitions. Instead, the environment "scatters" off of the qubit occasionally (with probability p) being kicked into the state $|1\rangle_E$ if A is in the state $|0\rangle_A$ and into the state $|2\rangle_E$ if A is in the state $|1\rangle_A$. Furthermore, also unlike the depolarizing channel, the channel picks out a preferred basis for qubit A; the basis $\{|0\rangle_A, |1\rangle_A\}$ is the only basis in which bit flips never occur.

Kraus operators

Evaluating the partial trace over $\mathcal{H}_E$ in the $\{|0\rangle_E, |1\rangle_E, |2\rangle_E\}$basis, we obtain the Kraus operators

$$
\boldsymbol{M}_0 = \sqrt{1-p}\,\mathbf{1}, \boldsymbol{M}_1 = \sqrt{p}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \boldsymbol{M}_2 = \sqrt{p}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.126}
$$

it is easy to check that $\boldsymbol{M}_0^2 + \boldsymbol{M}_1^2 + \boldsymbol{M}_2^2 = \mathbf{1}$. In this case, three Kraus operators are not really needed; a representation with two Kraus operators is possible, as you will show in a homework exercise.

An initial density matrix $\boldsymbol{\rho}$ evolves to

$$
\begin{aligned}
\$(\boldsymbol{\rho}) &= \boldsymbol{M}_0\boldsymbol{\rho}\boldsymbol{M}_0 + \boldsymbol{M}_1\boldsymbol{\rho}\boldsymbol{M}_1 + \boldsymbol{M}_2\boldsymbol{\rho}\boldsymbol{M}_2\\
&= (1-p)\boldsymbol{\rho} + p\begin{pmatrix} \rho_{00} & 0 \\ 0 & \rho_{11} \end{pmatrix} = \begin{pmatrix} \rho_{00} & (1-p)\;\;\rho_{01} \\ (1-p)\rho_{10} & \rho_{11} \end{pmatrix};
\end{aligned} \tag{3.127}
$$

thus the on-diagonal terms in $\boldsymbol{\rho}$ remain invariant while the off-diagonal terms decay.

Now suppose that the probability of a scattering event per unit time is Γ, so that $p = \Gamma\Delta t \ll 1$ when time Δt elapses. The evolution over a time

$t = n\Delta t$ is governed by $\$^n$, so that the off-diagonal terms are suppressed by $(1-p)^n = (1-\Gamma\Delta t)^{t/\Delta t} \to e^{-\Gamma t}$ (as $\Delta t \to 0$). Thus, if we prepare an initial pure state $a|0\rangle + b|1\rangle$, then after a time $t \gg \Gamma^{-1}$, the state decays to the incoherent superposition $\boldsymbol{\rho}' = |a|^2|0\rangle\langle 0| + |b|^2|1\rangle\langle 1|$. Decoherence occurs, in the preferred basis $\{|0\rangle, |1\rangle\}$.

Bloch-sphere representation

This will be worked out in a homework exercise.

Interpretation

We might interpret the phase-damping channel as describing a heavy "classical" particle (e.g., an interstellar dust grain) interacting with a background gas of light particles (e.g., the 3^0K microwave photons). We can imagine that the dust is initially prepared in a superposition of position eigenstates $|\psi\rangle = \frac{1}{\sqrt{2}}(|x\rangle + |-x\rangle)$ (or more generally a superposition of position-space wavepackets with little overlap). We might be able to monitor the behavior of the dust particle, but it is hopeless to keep track of the quantum state of all the photons that scatter from the particle; for our purposes, the quantum state of the particle is described by the density matrix $\boldsymbol{\rho}$ obtained by tracing over the photon degrees of freedom.

Our analysis of the phase damping channel indicates that if photons are scattered by the dust particle at a rate Γ, then the off-diagonal terms in $\boldsymbol{\rho}$ decay like $\exp(-\Gamma t)$, and so become completely negligible for $t \gg \Gamma^{-1}$. At that point, the coherence of the superposition of position eigenstates is completely lost – there is no chance that we can recombine the wavepackets and induce them to interfere. (If we attempt to do a double-slit interference pattern with dust grains, we will not see any interference pattern if it takes a time $t \gg \Gamma^{-1}$ for the grain to travel from the source to the screen.)

The dust grain is heavy. Because of its large inertia, its state of motion is little affected by the scattered photons. Thus, there are two disparate time scales relevant to its dynamics. On the one hand, there is a damping time scale, the time for a significant amount of the particle's momentum to be transfered to the photons; this is a long time if the particle is heavy. On the other hand, there is the decoherence time scale. In this model, the time scale for decoherence is of order Γ, the time for a *single* photon to be scattered by the dust grain, which is far shorter than the damping time scale. For a macroscopic object, decoherence is *fast*.

As we have already noted, the phase-damping channel picks out a preferred basis for decoherence, which in our "interpretation" we have assumed to be the position-eigenstate basis. Physically, decoherence prefers the spatially localized states of the dust grain because the *interactions* of photons and grains are localized in space. Grains in distinguishable positions tend to scatter the photons of the environment into mutually orthogonal states.

Even if the separation between the "grains" were so small that it could not be resolved very well by the scattered photons, the decoherence process would still work in a similar way. Perhaps photons that scatter off grains at positions x and $-x$ are not mutually orthogonal, but instead have an overlap

$$\langle\gamma+|\gamma-\rangle = 1-\varepsilon, \varepsilon \ll 1. \tag{3.128}$$

The phase-damping channel would still describe this situation, but with p replaced by $p\varepsilon$ (if p is still the probability of a scattering event). Thus, the decoherence rate would become $\Gamma_{\mathbf{dec}} = \varepsilon\Gamma_{\mathbf{scat}}$, where $\Gamma_{\mathbf{scat}}$ is the scattering rate (see the homework).

The intuition we distill from this simple model applies to a vast variety of physical situations. A coherent superposition of macroscopically distinguishable states of a "heavy" object decoheres very rapidly compared to its damping rate. The spatial locality of the interactions of the system with its environment gives rise to a preferred "local" basis for decoherence. Presumably, the same principles would apply to the decoherence of a "cat state" $\frac{1}{\sqrt{2}}(|\text{ dead}\rangle + |\text{ alive}\rangle)$, since "deadness" and "aliveness" can be distinguished by localized probes.

3.4.3 Amplitude-damping channel

The *amplitude-damping channel* is a schematic model of the decay of an excited state of a (two-level) atom due to spontaneous emission of a photon. By detecting the emitted photon ("observing the environment") we can perform a POVM that gives us information about the initial preparation of the atom.

Unitary representation

We denote the atomic ground state by $|0\rangle_A$ and the excited state of interest by $|1\rangle_A$. The "environment" is the electromagnetic field, assumed initially to be in its vacuum state $|0\rangle_E$. After we wait a while, there is a probability p

that the excited state has decayed to the ground state and a photon has been emitted, so that the environment has made a transition from the state $|0\rangle_E$ ("no photon") to the state $|1\rangle_E$ ("one photon"). This evolution is described by a unitary transformation that acts on atom and environment according to

$$\begin{aligned} |0\rangle_A|0\rangle_E &\to |0\rangle_A|0\rangle_E \\ |1\rangle_A|0\rangle_E &\to \sqrt{1-p}|1\rangle_A|0\rangle_E + \sqrt{p}|0\rangle_A|1\rangle_E. \end{aligned} \tag{3.129}$$

(Of course, if the atom starts out in its ground state, and the environment is at zero temperature, then there is no transition.)

Kraus operators

By evaluating the partial trace over the environment in the basis $\{|0\rangle_E, |1\rangle_E\}$, we find the kraus operators

$$\boldsymbol{M}_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}, \boldsymbol{M}_1 = \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix}, \tag{3.130}$$

and we can check that

$$\boldsymbol{M}_0^\dagger\boldsymbol{M}_0 + \boldsymbol{M}_1^\dagger\boldsymbol{M}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1-p \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & p \end{pmatrix} = \mathbf{1}. \tag{3.131}$$

The operator $\boldsymbol{M}_1$ induces a "quantum jump" – the decay from $|1\rangle_A$ to $|0\rangle_A$, and $\boldsymbol{M}_0$ describes how the state evolves if no jump occurs. The density matrix evolves as

$$\begin{aligned} \boldsymbol{\rho} \to \$(\rho) &= \boldsymbol{M}_0\boldsymbol{\rho}\boldsymbol{M}_0^\dagger + \boldsymbol{M}_1\boldsymbol{\rho}\boldsymbol{M}_1^\dagger \\ &= \begin{pmatrix} \rho_{00} & \sqrt{1-p}\rho_{01} \\ \sqrt{1-p}\rho_{10} & (1-p)\rho_{11} \end{pmatrix} + \begin{pmatrix} p\rho_{11} & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \rho_{00}+p\rho_{11} & \sqrt{1-p}\rho_{01} \\ \sqrt{1-p}\rho_{10} & (1-p)\rho_{11} \end{pmatrix}. \end{aligned} \tag{3.132}$$

If we apply the channel n times in succession, the ρ_{11} matrix element decays as

$$\rho_{11} \to (1-p)^n\rho_{11}; \tag{3.133}$$

so if the probability of a transition in time interval Δt is $\Gamma\Delta t$, then the probability that the excited state persists for time t is $(1-\Gamma\Delta t)^{t/\Delta t} \to e^{-\Gamma t}$, the expected exponential decay law.

As $t \to \infty$, the decay probability approaches unity, so

$$\$(\boldsymbol{\rho}) \to \begin{pmatrix} \rho_{00}+\rho_{11} & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.134}$$

The atom always winds up in its ground state. This example shows that it is sometimes possible for a superoperator to take a mixed initial state, e.g.,

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{00} & 0 \\ 0 & \rho_{11} \end{pmatrix}, \tag{3.135}$$

to a pure final state.

Watching the environment

In the case of the decay of an excited atomic state via photon emission, it may not be impractical to monitor the environment with a photon detector. The measurement of the environment prepares a pure state of the atom, and so in effect prevents the atom from decohering.

Returning to the unitary representation of the amplitude-damping channel, we see that a coherent superposition of the atomic ground and excited states evolves as

$$\begin{aligned} &(a|0\rangle_A + b|1\rangle_A)|0\rangle_E \\ &\quad \to (a|0\rangle_A + b\sqrt{1-p}|1\rangle)|0\rangle_E + \sqrt{p}|0\rangle_A|1\rangle_E. \end{aligned} \tag{3.136}$$

If we detect the photon (and so project out the state $|1\rangle_E$ of the environment), then we have prepared the state $|0\rangle_A$ of the atom. In fact, we have prepared a state in which we know with certainty that the initial atomic state was the excited state $|1\rangle_A$ – the ground state could not have decayed.

On the other hand, if we detect no photon, and our photon detector has perfect efficiency, then we have projected out the state $|0\rangle_E$ of the environment, and so have prepared the atomic state

$$a|0\rangle_A + b\sqrt{1-p}|1\rangle_A. \tag{3.137}$$

The atomic state has evolved due to our failure to detect a photon – it has become more likely that the initial atomic state was the ground state!

As noted previously, a unitary transformation that entangles A with E, followed by an orthogonal measurement of E, can be described as a POVM in A. If $|\varphi\rangle_A$ evolves as

$$|\varphi\rangle_A|0\rangle_E \to \sum_\mu \boldsymbol{M}_\mu|\varphi\rangle_A|\mu\rangle_E, \tag{3.138}$$

then an orthogonal measurement in E that projects onto the $\{|\mu\rangle_E\}$ basis realizes a POVM with

$$\text{Prob}(\mu) = tr(\boldsymbol{F}_\mu\boldsymbol{\rho}_A), \quad \boldsymbol{F}_\mu = \boldsymbol{M}_\mu^\dagger\boldsymbol{M}_\mu, \tag{3.139}$$

for outcome μ. In the case of the amplitude damping channel, we find

$$\boldsymbol{F}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1-p \end{pmatrix}, \quad \boldsymbol{F}_1 = \begin{pmatrix} 0 & 0 \\ 0 & p \end{pmatrix}, \tag{3.140}$$

where $\boldsymbol{F}_1$ determines the probability of a successful photon detection, and $\boldsymbol{F}_0$ the complementary probability that no photon is detected.

If we wait a time $t \gg \Gamma^{-1}$, so that p approaches 1, our POVM approaches an orthogonal measurement, the measurement of the initial atomic state in the $\{|0\rangle_A, |1\rangle_A\}$ basis. A peculiar feature of this measurement is that we can project out the state $|0\rangle_A$ by *not* detecting a photon. This is an example of what Dicke called "interaction-free measurement" – because *no change* occured in the state of the environment, we can infer what the atomic state must have been. The term "interaction-free measurement" is in common use, but it is rather misleading; obviously, if the Hamiltonian of the world did not include a coupling of the atom to the electromagnetic field, the measurement could not have been possible.

3.5 Master Equation

3.5.1 Markovian evolution

The superoperator formalism provides us with a general description of the evolution of density matrices, including the evolution of pure states to mixed states (decoherence). In the same sense, unitary transformations provide a general description of coherent quantum evolution. But in the case of coherent evolution, we find it very convenient to characterize the dynamics of a quantum system with a *Hamiltonian*, which describes the evolution over an infinitesimal time interval. The dynamics is then described by a differential equation, the *Schrödinger equation*, and we may calculate the evolution over a finite time interval by integrating the equation, that is, by piecing together the evolution over many infinitesimal intervals.

It is often possible to describe the (not necessarily coherent) evolution of a density matrix, at least to a good approximation, by a differential equation. This equation, the *master equation*, will be our next topic.

In fact, it is not at all obvious that there need be a differential equation that describes decoherence. Such a description will be possible only if the evolution of the quantum system is "Markovian," or in other words, *local* in time. If the evolution of the density operator $\boldsymbol{\rho}(t)$ is governed by a (first-order) differential equation in t, then that means that $\boldsymbol{\rho}(t+dt)$ is completely determined by $\boldsymbol{\rho}(t)$.

We have seen that we can always describe the evolution of density operator $\boldsymbol{\rho}_A$ in Hilbert space $\mathcal{H}_A$ if we imagine that the evolution is actually unitary in the extended Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_E$. But even if the evolution in $\mathcal{H}_A \otimes \mathcal{H}_E$ is governed by a Schrdinger equation, this is not sufficient to ensure that the evolution of $\boldsymbol{\rho}_A(t)$ will be local in t. Indeed, if we know only $\boldsymbol{\rho}_A(t)$, we do not have complete initial data for the Schrodinger equation; we need to know the state of the "environment," too. Since we know from the general theory of superoperators that we are entitled to insist that the quantum state in $\mathcal{H}_A \otimes \mathcal{H}_E$ at time $t=0$ is

$$\boldsymbol{\rho}_A \otimes |0\rangle_E \; {}_E\langle 0|, \tag{3.141}$$

a sharper statement of the difficulty is that the density operator $\boldsymbol{\rho}_A(t+dt)$ depends not only on $\boldsymbol{\rho}_A(t)$, but also on $\boldsymbol{\rho}_A$ at earlier times, because the reservoir E[7] retains a *memory* of this information for a while, and can transfer it back to system A.

This quandary arises because information flows on a two-way street. An open system (whether classical or quantum) is *dissipative* because information can flow from the system to the reservoir. But that means that information can also flow back from reservoir to system, resulting in non-Markovian

[7]In discussions of the mater equation, the environment is typically called the *reservoir*, in deference to the deeply ingrained conventions of statistical physics.

fluctuations of the system.[8]

Except in the case of coherent (unitary) evolution, then, fluctuations are inevitable, and an exact Markovian description of quantum dynamics is impossible. Still, in many contexts, a Markovian description is a very good approximation. The key idea is that there may be a clean separation between the typical correlation time of the fluctuations and the time scale of the evolution that we want to follow. Crudely speaking, we may denote by $(\Delta t)_{\rm res}$ the time it takes for the reservoir to "forget" information that it acquired from the system — after time $(\Delta t)_{\rm res}$ we can regard that information as forever lost, and neglect the possibility that the information may feed back again to influence the subsequent evolution of the system.

Our description of the evolution of the system will incorporate "coarse-graining" in time; we perceive the dynamics through a filter that screens out the high frequency components of the motion, with $\omega \gg (\Delta t_{\rm coarse})^{-1}$. An approximately Markovian description should be possible, then, if $(\Delta t)_{\rm res} \ll (\Delta t)_{\rm coarse}$; we can neglect the memory of the reservoir, because we are unable to resolve its effects. This "Markovian approximation" will be *useful* if the time scale of the dynamics that we want to observe is long compared to $(\Delta t)_{\rm coarse}$, e.g., if the *damping* time scale $(\Delta t)_{\rm damp}$ satisfies

$$(\Delta t)_{\rm damp} \gg (\Delta t)_{\rm coarse} \gg (\Delta t)_{\rm res}. \tag{3.142}$$

This condition often applies in practice, for example in atomic physics, where $(\Delta t)_{\rm res} \sim \hbar/kT \sim 10^{-14}$ s (T is the temperature) is orders of magnitude larger than the typical lifetime of an excited atomic state.

An instructive example to study is the case where the system A is a single harmonic oscillator ($\mathbf{H}_A = \omega \boldsymbol{a}^\dagger \boldsymbol{a}$), and the reservoir R consists of many oscillators ($\boldsymbol{H}_R = \sum_i \omega_i \boldsymbol{b}_i^\dagger \boldsymbol{b}_i$, weakly coupled to the system by a perturbation

$$\mathbf{H}' = \sum_i \lambda_i (\boldsymbol{a}\boldsymbol{b}_i^\dagger + \boldsymbol{a}^\dagger \boldsymbol{b}_i). \tag{3.143}$$

The reservoir Hamiltonian could represent the (free) electromagnetic field, and then $\boldsymbol{H}'$, in lowest nontrivial order of perturbation theory induces transitions in which the oscillator emits or absorbs a single photon, with its occupation number $n = \boldsymbol{a}^\dagger \boldsymbol{a}$ decreasing or increasing accordingly.

[8]This inescapable connection underlies the *fluctuation-dissipation theorem*, a powerful tool in statistical physics.

We could arrive at the master equation by analyzing this system using time-dependent perturbation theory, and carefully introducing a finite frequency cutoff. The details of that analysis can be found in the book "An Open Systems Approach to Quantum Optics," by Howard Carmichael. Here, though, I would like to short-circuit that careful analysis, and leap to the master equation by a more heuristic route.

3.5.2 The Lindbladian

Under unitary evolution, the time evolution of the density matrix is governed by the Schrödinger equation

$$\dot{\boldsymbol{\rho}} = -i[\boldsymbol{H}, \boldsymbol{\rho}], \tag{3.144}$$

which we can solve formally to find

$$\boldsymbol{\rho}(t) = e^{-i\boldsymbol{H}t}\boldsymbol{\rho}_{(0)}e^{i\boldsymbol{H}t}, \tag{3.145}$$

if $\boldsymbol{H}$ is time independent. Our goal is to generalize this equation to the case of Markovian but nonunitary evolution, for which we will have

$$\dot{\boldsymbol{\rho}} = \mathcal{L}[\boldsymbol{\rho}]. \tag{3.146}$$

The linear operator $\mathcal{L}$, which generates a finite superoperator in the same sense that a Hamiltonian $\boldsymbol{H}$ generates unitary time evolution, will be called the *Lindbladian*. The formal solution to eq. (3.146) is

$$\boldsymbol{\rho}(t) = e^{\mathcal{L}t}[\boldsymbol{\rho}(0)], \tag{3.147}$$

if $\mathcal{L}$ is t-independent.

To compute the Lindbladian, we could start with the Schrödinger equation for the coupled system and reservoir

$$\dot{\boldsymbol{\rho}}_A = tr_R(\dot{\boldsymbol{\rho}}_{AR}) = tr_R(-i[\boldsymbol{H}_{AR}, \boldsymbol{\rho}_{AR}]), \tag{3.148}$$

but as we have already noted, we cannot expect that this formula for $\dot{\boldsymbol{\rho}}_A$ can be expressed in terms of $\boldsymbol{\rho}_A$ alone. To obtain the Lindbladian, we need to explicitly invoke the Markovian approximation (as Carmichael does). On the other hand, suppose we *assume* that the Markov approximation applies. We already know that a *general* superoperator has a Kraus representation

$$\boldsymbol{\rho}(t) = \$_t(\boldsymbol{\rho}(0)) = \sum_\mu \boldsymbol{M}_\mu(t)\boldsymbol{\rho}(0)\boldsymbol{M}_\mu^\dagger(t), \tag{3.149}$$

and that $\$_{t=0} = I$. If the elapsed time is the infinitesimal interval dt, and

$$\boldsymbol{\rho}(dt) = \boldsymbol{\rho}(0) + O(dt), \tag{3.150}$$

then one of the Kraus operators will be $\boldsymbol{M}_0 = \mathbf{1} + O(dt)$, and all the others will be of order $\sqrt{dt}$. The operators $\boldsymbol{M}_\mu, \mu > 0$ describe the "quantum jumps" that the system might undergo, all occuring with a probability of order dt. We may, therefore, write

$$\begin{aligned} \boldsymbol{M}_\mu &= \sqrt{dt}\,\mathbf{L}_\mu, \quad \mu = 1, 2, 3, \ldots \\ \boldsymbol{M}_0 &= \mathbf{1} + (-i\boldsymbol{H} + \mathbf{K})dt, \end{aligned} \tag{3.151}$$

where $\boldsymbol{H}$ and $\boldsymbol{K}$ are both hermitian and $\boldsymbol{L}_\mu, \boldsymbol{H}$, and $\boldsymbol{K}$ are all zeroth order in dt. In fact, we can determine $\mathbf{K}$ by invoking the Kraus normalization condition:

$$\mathbf{1} = \sum_\mu \boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu = \mathbf{1} + dt(2\boldsymbol{K} + \sum_{\mu>0} \boldsymbol{L}_\mu^\dagger \boldsymbol{L}_\mu), \tag{3.152}$$

or

$$\boldsymbol{K} = -\frac{1}{2}\sum_{\mu>0} \boldsymbol{L}_\mu^\dagger \boldsymbol{L}_\mu. \tag{3.153}$$

Substituting into eq. (3.149), expressing $\boldsymbol{\rho}(dt) = \boldsymbol{\rho}(0) + dt\dot{\boldsymbol{\rho}}(0)$, and equating terms of order dt, we obtain Lindblad's equation:

$$\dot{\boldsymbol{\rho}} \equiv \mathcal{L}[\boldsymbol{\rho}] = -i[\boldsymbol{H}, \boldsymbol{\rho}] + \sum_{\mu>0}\left(\boldsymbol{L}_\mu\boldsymbol{\rho}\boldsymbol{L}_\mu^\dagger - \frac{1}{2}\boldsymbol{L}_\mu^\dagger\boldsymbol{L}_\mu\boldsymbol{\rho} - \frac{1}{2}\boldsymbol{\rho}\boldsymbol{L}_\mu^\dagger\boldsymbol{L}_\mu\right). \tag{3.154}$$

The first term in $\mathcal{L}[\boldsymbol{\rho}]$ is the usual Schrodinger term that generates unitary evolution. The other terms describe the possible transitions that the system may undergo due to interactions with the reservoir. The operators $\boldsymbol{L}_\mu$ are called *Lindblad operators* or *quantum jump operators*. Each $\boldsymbol{L}_\mu\boldsymbol{\rho}\boldsymbol{L}_\mu^\dagger$ term induces one of the possible quantum jumps, while the $-1/2\boldsymbol{L}_\mu^\dagger\boldsymbol{L}_\mu\boldsymbol{\rho} - 1/2\boldsymbol{\rho}\boldsymbol{L}_\mu^\dagger\boldsymbol{L}_\mu$ terms are needed to normalize properly the case in which no jumps occur.

Lindblad's eq (3.154) is what we were seeking – the general form of (completely positive) Markovian evolution of a density matrix: that is, the master equation. It follows from the Kraus representation that we started with that Lindblad's equation preserves density matrices: $\boldsymbol{\rho}(t + dt)$ is a density matrix if $\boldsymbol{\rho}(t)$ is. Indeed, we can readily check, using eq. (3.154), that $\dot{\boldsymbol{\rho}}$ is Hermitian and $tr\dot{\boldsymbol{\rho}} = 0$. That $\mathcal{L}[\boldsymbol{\rho}]$ preserves positivity is somewhat less manifest but, as already noted, follows from the Kraus representation.

If we recall the connection between the Kraus representation and the unitary representation of a superoperator, we clarify the interpretation of the master equation. We may imagine that we are continuously monitoring the reservoir, projecting it in each instant of time onto the $|\mu\rangle_R$ basis. With probability $1 - 0(dt)$, the reservoir remains in the state $|0\rangle_R$, but with probability of order dt, the reservoir makes a quantum jump to one of the states $|\mu\rangle_R$, $\mu > 0$. When we say that the reservoir has "forgotten" the information it acquired from the system (so that the Markovian approximation applies), we mean that these transitions occur with probabilities that increase linearly with time. Recall that this is *not* automatic in time-dependent perturbation theory. At a small time t the probability of a particular transition is proportional to t^2; we obtain a rate (in the derivation of "Fermi's golden rule") only by summing over a continuum of possible final states. Because the number of accessible states actually decreases like $1/t$, the probability of a transition, summed over final states, is proportional to t. By using a Markovian description of dynamics, we have implicitly assumed that our $(\Delta t)_{\rm coarse}$ is long enough so that we can assign rates to the various possible transitions that might be detected when we monitor the environment. In practice, this is where the requirement $(\Delta t)_{\rm coarse} \gg (\Delta t)_{\rm res}$ comes from.

3.5.3 Damped harmonic oscillator

As an example to illustrate the master equation, we consider the case of a harmonic oscillator interacting with the electromagnetic field, coupled as

$$\boldsymbol{H}' = \sum_i \lambda_i (\boldsymbol{a}\boldsymbol{b}_i^\dagger + \boldsymbol{a}^\dagger \boldsymbol{b}_i). \tag{3.155}$$

Let us also suppose that the reservoir is at zero temperature; then the excitation level of the oscillator can cascade down by successive emission of photons, but no absorption of photons will occur. Hence, there is only one jump operator:

$$\boldsymbol{L}_1 = \sqrt{\Gamma}\boldsymbol{a}. \tag{3.156}$$

Here Γ is the rate for the oscillator to decay from the first excited ($n = 1$) state to the ground ($n = 0$) state; because of the form of $\boldsymbol{H}$, the rate for

the decay from level n to $n - I$ is $n\Gamma$.[9] The master equation in the Lindblad form becomes

$$\dot{\boldsymbol{\rho}} = -i[\boldsymbol{H}_0, \boldsymbol{\rho}] + \Gamma(\boldsymbol{a}\boldsymbol{\rho}\boldsymbol{a}^\dagger - \frac{1}{2}\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho} - \frac{1}{2}\boldsymbol{\rho}\boldsymbol{a}^\dagger\boldsymbol{a}). \tag{3.157}$$

where $\boldsymbol{H}_0 = \omega\boldsymbol{a}^\dagger\boldsymbol{a}$ is the Hamiltonian of the oscillator. This is the same equation obtained by Carmichael from a more elaborate analysis. (The only thing we have missed is the *Lamb shift*, a radiative renormalization of the frequency of the oscillator that is of the same order as the jump terms in $\mathcal{L}[\boldsymbol{\rho}]$.)

The jump terms in the master equation describe the *damping* of the oscillator due to photon emission.[10] To study the effect of the jumps, it is convenient to adopt the *interaction picture*; we define interaction picture operators $\boldsymbol{\rho}_I$ and $\boldsymbol{a}_I$ by

$$\begin{aligned} \boldsymbol{\rho}(t) &= e^{-i\boldsymbol{H}_0 t}\boldsymbol{\rho}_I(t)e^{i\boldsymbol{H}_0 t}, \\ \boldsymbol{a}(t) &= e^{-i\boldsymbol{H}_0 t}\boldsymbol{a}_I(t)e^{i\boldsymbol{H}_0 t}, \end{aligned} \tag{3.158}$$

so that

$$\dot{\boldsymbol{\rho}}_I = \Gamma(\boldsymbol{a}_I\boldsymbol{\rho}_I\boldsymbol{a}_I^\dagger - \frac{1}{2}\boldsymbol{a}_I^\dagger\boldsymbol{a}_I\boldsymbol{\rho} - \frac{1}{2}\boldsymbol{\rho}_I\boldsymbol{a}_I^\dagger\boldsymbol{a}_I). \tag{3.159}$$

where in fact $\boldsymbol{a}_I(t) = \boldsymbol{a}e^{-i\omega t}$ so we can replace $\boldsymbol{a}_I$ by $\boldsymbol{a}$ on the right-hand side. The variable $\tilde{\boldsymbol{a}} = e^{-iH_0 t}\boldsymbol{a}e^{+iH_0 t} = e^{i\omega t}\boldsymbol{a}$ remains constant in the absence of damping. With damping, $\tilde{\boldsymbol{a}}$ decays according to

$$\frac{d}{dt}\langle\tilde{\boldsymbol{a}}\rangle = \frac{d}{dt}tr(\boldsymbol{a}\boldsymbol{\rho}_I) = tr\boldsymbol{a}\dot{\boldsymbol{\rho}} \ , \tag{3.160}$$

and from eq. (3.159) we have

$$tr\boldsymbol{a}\dot{\boldsymbol{\rho}} = \Gamma tr\left(\boldsymbol{a}^2\boldsymbol{\rho}_I\boldsymbol{a}^\dagger - \frac{1}{2}\boldsymbol{a}\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho}_I - \frac{1}{2}\boldsymbol{a}\boldsymbol{\rho}_I\boldsymbol{a}^\dagger\boldsymbol{a}\right)$$

[9] The nth level of excitation of the oscillator may be interpreted as a state of n noninteracting particles; the rate is $n\Gamma$ because any one of the n particles can decay.

[10] This model extends our discussion of the amplitude-damping channel to a damped oscillator rather than a damped qubit.

$$= \Gamma tr\left(\frac{1}{2}[\boldsymbol{a}^\dagger, \boldsymbol{a}]\boldsymbol{a}\boldsymbol{\rho}_I\right) = -\frac{\Gamma}{2} tr(\boldsymbol{a}\boldsymbol{\rho}_I) = -\frac{\Gamma}{2}\langle\tilde{\boldsymbol{a}}\rangle. \tag{3.161}$$

Integrating this equation, we obtain

$$\langle\tilde{\boldsymbol{a}}(t)\rangle = e^{-\Gamma t/2}\langle\tilde{\boldsymbol{a}}(0)\rangle. \tag{3.162}$$

Similarly, the occupation number of the oscillator $n \equiv \boldsymbol{a}^\dagger\boldsymbol{a} = \tilde{\boldsymbol{a}}^\dagger\tilde{\boldsymbol{a}}$ decays according to

$$\frac{d}{dt}\langle n\rangle = \frac{d}{dt}\langle\tilde{\boldsymbol{a}}^\dagger\tilde{\boldsymbol{a}}\rangle = tr(\boldsymbol{a}^\dagger\boldsymbol{a}\dot{\boldsymbol{\rho}}_I)$$

$$\begin{aligned} &= \Gamma tr\left(\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{a}\boldsymbol{\rho}_I\boldsymbol{a}^\dagger - \frac{1}{2}\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho}_I - \frac{1}{2}\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho}_I\boldsymbol{a}^\dagger\boldsymbol{a}\right) \\ &= \Gamma tr\boldsymbol{a}^\dagger[\boldsymbol{a}^\dagger, \boldsymbol{a}]\boldsymbol{a}\boldsymbol{\rho}_I = -\Gamma tr\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho}_I = -\Gamma\langle n\rangle, \end{aligned} \tag{3.163}$$

which integrates to

$$\langle n(t)\rangle = e^{-\Gamma t}\langle n(0)\rangle. \tag{3.164}$$

Thus Γ is the damping rate of the oscillator. We can interpret the nth excitation state of the oscillator as a state of n noninteracting particles, each with a decay probability Γ per unit time; hence eq. (3.164) is just the exponential law satisfied by the population of decaying particles.

More interesting is what the master equation tells us about decoherence. The details of that analysis will be a homework exercise. But we will analyze here a simpler problem – an oscillator undergoing phase damping.

3.5.4 Phase damping

To model phase damping of the oscillator, we adopt a different coupling of the oscillator to the reservoir:

$$\boldsymbol{H}' = \left(\sum_i \lambda_i \boldsymbol{b}_i^\dagger\boldsymbol{b}_i\right)\boldsymbol{a}^\dagger\boldsymbol{a}. \tag{3.165}$$

Thus, there is just one Lindblad operator, and the master equation in the interaction picture is.

$$\dot{\boldsymbol{\rho}}_I = \Gamma\left(\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho}_I\boldsymbol{a}^\dagger\boldsymbol{a} - \frac{1}{2}(\boldsymbol{a}^\dagger\boldsymbol{a})^2\boldsymbol{\rho}_I - \frac{1}{2}\boldsymbol{\rho}_I(\boldsymbol{a}^\dagger\boldsymbol{a})^2\right). \tag{3.166}$$

Here Γ can be interpreted as the rate at which reservoir photons are *scattered* when the oscillator is singly occupied. If the occupation number is n then the scattering rate becomes Γn^2. The reason for the factor of n^2 is that the contributions to the scattering amplitude due to each of n oscillator "particles" all add coherently; the amplitude is proportional to n and the rate to n^2.

It is easy to solve for $\dot{\boldsymbol{\rho}}_I$ in the occupation number basis. Expanding

$$\boldsymbol{\rho}_I = \sum_{n,m}\rho_{nm}|n\rangle\langle m|, \tag{3.167}$$

(where $\boldsymbol{a}^\dagger\boldsymbol{a}|n\rangle = n|n\rangle$), the master equation becomes

$$\begin{aligned} \dot{\rho}_{nm} &= \Gamma\left(nm - \frac{1}{2}n^2 - \frac{1}{2}m^2\right)\rho_{nm} \\ &= -\frac{\Gamma}{2}(n-m)^2\rho_{nm}, \end{aligned} \tag{3.168}$$

which integrates to

$$\rho_{nm}(t) = \rho_{nm}(0)\exp\left[-\frac{1}{2}\Gamma t(n-m)^2\right]. \tag{3.169}$$

If we prepare a "cat state" like

$$|\mathrm{cat}\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |n\rangle), \quad n \gg 1, \tag{3.170}$$

a superposition of occupation number eigenstates with much different values of n, the off-diagonal terms in the density matrix decay like $\exp(-\frac{1}{2}\Gamma n^2 t)$. In fact, this is just the same sort of behavior we found when we analyzed phase damping for a single qubit. The rate of decoherence is Γn^2 because this is the rate for reservoir photons to scatter off the excited oscillator in the state $|n\rangle$. We also see, as before, that the phase decoherence chooses a preferred basis. Decoherence occurs in the number-eigenstate basis because it is the occupation number that appears in the coupling $\boldsymbol{H}'$ of the oscillator to the reservoir.

Return now to amplitude damping. In our amplitude damping model, it is the annihilation operator $\boldsymbol{a}$ (and its adjoint) that appear in the coupling $\boldsymbol{H}'$ of oscillator to reservoir, so we can anticipate that decoherence will occur in the basis of $\boldsymbol{a}$ eigenstates. The *coherent state*

$$|\alpha\rangle = e^{-|\alpha|^2/2}e^{\alpha a^\dagger}|0\rangle = e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}|n\rangle, \tag{3.171}$$

is the normalized eigenstate of $\boldsymbol{a}$ with complex eigenvalue α. Two coherent states with distinct eigenvalues α_1 and α_2 are not orthogonal; rather

$$\begin{aligned}|\langle\alpha_1|\alpha_2\rangle|^2 &= e^{-|\alpha_1|^2}e^{-|\alpha_2|^2}e^{2Re(\alpha_1^*\alpha_2)}\\ &= \exp(-|\alpha_1-\alpha_2|^2),\end{aligned} \tag{3.172}$$

so the overlap is very small when $|\alpha_1-\alpha_2|$ is large.

Imagine that we prepare a cat state

$$|\mathrm{cat}\rangle = \frac{1}{\sqrt{2}}(|\alpha_1\rangle + |\alpha_2\rangle), \tag{3.173}$$

a superposition of coherent states with $|\alpha_1-\alpha_2| \gg 1$. You will show that the off diagonal terms in $\boldsymbol{\rho}$ decay like

$$\exp\left(-\frac{\Gamma t}{2}|\alpha_1-\alpha_2|^2\right) \tag{3.174}$$

(for $\Gamma t << 1$). Thus the decoherence rate

$$\Gamma_{\mathrm{dec}} = \frac{1}{2}|\alpha_1-\alpha_2|^2\Gamma_{\mathrm{damp}}, \tag{3.175}$$

is enormously fast compared to the damping rate. Again, this behavior is easy to interpret. The expectation value of the occupation number in a coherent state is $\langle\alpha|\boldsymbol{a}^\dagger\boldsymbol{a}|\alpha\rangle = |\alpha|^2$. Therefore, if $\alpha_{1,2}$ have comparable moduli but significantly different phases (as for a superposition of minimum uncertainty wave packets centered at x and $-x$), the decoherence rate is of the order of the rate for emission of a *single* photon. This rate is very large compared to the rate for a significant fraction of the oscillator energy to be dissipated.

We can also consider an oscillator coupled to a reservoir with a finite temperature. Again, the decoherence rate is roughly the rate for a single photon to be emitted or absorbed, but the rate is much faster than at zero temperature. Because the photon modes with frequency comparable to the oscillator frequency ω have a thermal occupation number

$$n_\gamma = \frac{T}{\hbar\omega}, \tag{3.176}$$

(for $T \gg \hbar\omega$), the interaction rate is further enhanced by the factor n_γ. We have then

$$\begin{aligned}\frac{\Gamma_{\mathrm{dec}}}{\Gamma_{\mathrm{damp}}} &\sim n_{\mathrm{osc}}n_\gamma \sim \frac{E}{\hbar\omega}\frac{T}{\hbar\omega}\\ &\sim \frac{m\omega^2x^2}{\hbar\omega}\frac{T}{\hbar\omega} \sim x^2\frac{mT}{\hbar^2} \sim \frac{x^2}{\lambda_T^2},\end{aligned} \tag{3.177}$$

where x is the amplitude of oscillation and λ_T is the thermal de Broglie wavelength. Decoherence is *fast*.

3.6 What is the problem? (Is there a problem?)

Our survey of the foundations of quantum theory is nearly complete. But before we proceed with our main business, let us briefly assess the status of these foundations. Is quantum theory in good shape, or is there a fundamental problem at its roots still in need of resolution?

One potentially serious issue, first visited in §2.1, is the *measurement problem*. We noted the odd dualism inherent in our axioms of quantum theory. There are two ways for the quantum state of a system to change: *unitary* evolution, which is *deterministic*, and *measurement*, which is *probabilistic*. But why should measurement be fundamentally different than any other physical process? The dualism has led some thoughtful people to suspect that our current formulation of quantum theory is still not complete.

In this chapter, we have learned more about measurement. In §3.1.1, we discussed how unitary evolution can establish correlations (entanglement) between a system and the *pointer* of an apparatus. Thus, a pure state of the system can evolve to a mixed state (after we trace over the pointer states), and that mixed state admits an interpretation as an *ensemble* of mutually orthogonal pure states (the eigenstates of the density operator of the system), each occuring with a specified probability. Thus, already in this simple observation, we find the seeds of a deeper understanding of how the "collapse" of a state vector can arise from unitary evolution alone. But on the other hand, we discussed in §2.5 now the ensemble interpretation of a density matrix is ambiguous, and we saw particularly clearly in §2.5.5 that, if we are able to measure the pointer in any basis we please, then we can prepare the system in any one of many "weird" states, superpositions of eigenstates of the system's $\boldsymbol{\rho}$ (the GHJW theorem). Collapse, then (which *destroys* the relative phases of the states in a superposition), cannot be explained by entanglement alone.

In §3.4 and §3.5, we studied another important element of the measurement process – *decoherence*. The key idea is that, for macroscopic systems, we cannot hope to keep track of all microscopic degrees of freedom. We must be content with a *coarse-grained* description, obtained by tracing over the many unobserved variables. In the case of a macroscopic measurement apparatus, we must trace over the degrees of freedom of the environment with which the apparatus inevitably interacts. We then find that the apparatus decoheres exceedingly rapidly in a certain preferred basis, a basis determined by the nature of the coupling of the apparatus to the environment. It seems to be a feature of the Hamiltonian of the world that fundamental interactions are well localized in space, and therefore the basis selected by decoherence is a basis of states that are well localized spatially. The cat is either alive or dead – it is not in the state $1/\sqrt{2}(|\text{Alive}\rangle + |\text{Dead}\rangle)$.

By tracing over the degrees of freedom of the environment, we obtain a more complete picture of the measurement process, of "collapse." Our system becomes entangled with the apparatus, which is in turn entangled with the environment. If we regard the microstate of the environment as forever inaccessible, then we are well entitled to say that a measurement has taken place. The relative phases of the basis states of the system have been lost irrevocably – its state vector has collapsed.

Of course, as a matter of principle, no phase information has really been lost. The evolution of system + apparatus + environment is unitary and deterministic. In principle, we could, perhaps, perform a highly nonlocal measurement of the environment, and restore to the system the phase information that was allegedly destroyed. In this sense, our explanation of collapse is, as John Bell put it, merely FAPP (for all practical purposes). After the "measurement," the coherence of the system basis states could still be restored in principle (we could reverse the measurement by "quantum erasure"), but undoing a measurement is extremely improbable. True, collapse is merely FAPP (though perhaps we might argue, in a cosmological context, that some measurements really *are* irreversible in principle), but isn't FAPP good enough?

Our goal in physics is to account for observed phenomena with a model that is as simple as possible. We should not postulate two fundamental processes (unitary evolution *and* measurement) if only one (unitary evolution) will suffice. Let us then accept, at least provisionally, this hypothesis:

> The evolution of a *closed* quantum system is *always unitary*.

Of course, we have seen that not all superoperators are unitary. The point of the hypothesis is that nonunitary evolution in an *open* system, including

the collapse that occurs in the measurement process, always arises from disregarding some of the degrees of freedom of a larger system. This is the view promulgated by Hugh Everett, in 1957. According to this view, the evolution of the quantum state of "the universe" is actually deterministic!

But even if we accept that collapse is explained by decoherence in a system that is truly deterministic, we have not escaped all the puzzles of quantum theory. For the wave function of the universe is in fact a superposition of a state in which the cat is dead and a state in which the cat is alive. Yet each time I look at a cat, it is always either dead or alive. Both outcomes are possible, but only one is realized in *fact*. Why is that?

Your answer to this question may depend on what you think quantum theory is about. There are (at least) two reasonable schools of thought.

Platonic : Physics describes *reality*. In quantum theory, the "wave function of the universe" is a complete description of physical reality.

Positivist : Physics describes our *perceptions*. The wave function encodes our state of knowledge, and the task of quantum theory is to make the best possible predictions about the future, given our current state of knowledge.

I believe in reality. My reason, I think, is a pragmatic one. As a physicist, I seek the most economical model that "explains" what I perceive. To this physicist, at least, the simplest assumption is that my perceptions (and yours, too) are correlated with an underlying reality, external to me. This ontology may seem hopelessly naive to a serious philosopher. But I choose to believe in reality because that assumption seems to be the simplest one that might successfully account for my perceptions. (In a similar vein, I chose to believe that science is more than just a social consensus. I believe that science makes progress, and comes ever closer to a satisfactory understanding of Nature – the laws of physics are discovered, not invented. I believe this because it is the simplest explanation of how scientists are so successful at reaching consensus.)

Those who hold the contrary view (that, even if there is an underlying reality, the state vector only encodes a state of knowledge rather than an underlying reality) tend to believe that the current formulation of quantum theory is not fully satisfactory, that there is a deeper description still awaiting discovery. To me it seems more economical to assume that the wavefunction does describe reality, unless and until you can dissuade me.

If we believe that the wavefunction describes reality and *if* we accept Everett's view that all evolution is unitary, then we must accept that all possible outcomes of a measurement have an equal claim to being "real." How then, are we to understand why, when we do an experiment, only *one* outcome is actually realized – the cat is either alive or dead.

In fact there is no paradox here, but only if we are willing (consistent with the spirit of the Everett interpretation) to include *ourselves* in the quantum system described by the wave function. This wave function describes all the possible correlations among the subsystems, including the correlations between the cat and my mental state. If we prepare the cat state and then look at the cat, the density operator (after we trace over other extraneous degrees of freedom) becomes

$$|\text{Decay}\rangle_{\text{atom}}\ |\text{Dead}\rangle_{\text{cat}}\ |\text{Know it's Dead}\rangle_{\text{me}} \quad \left(\text{Prob} = \frac{1}{2}\right)$$

$$|\text{No decay}\rangle_{\text{atom}}\ |\text{Alive}\rangle_{\text{cat}}\ |\text{Know it's Alive}\rangle_{\text{me}} \quad \left(\text{Prob} = \frac{1}{2}\right) \tag{3.178}$$

This $\boldsymbol{\rho}$ describes two alternatives, but for either alternative, I am certain about the health of the cat. I *never* see a cat that is half alive and half dead. (I am in an eigenstate of the "certainty operator," in accord with experience.)

By assuming that the wave function describes reality and that all evolution is unitary, we are led to the "many-worlds interpretation" of quantum theory. In this picture, each time there is a "measurement," the wave function of the universe "splits" into two branches, corresponding to the two possible outcomes. After many measurements, there are many branches (many worlds), all with an equal claim to describing reality. This proliferation of worlds seems like an ironic consequence of our program to develop the most economical possible description. But we ourselves follow one particular branch, and for the purpose of predicting what we will see in the next instant, the many other branches are of no consequence. The proliferation of worlds comes at no cost to us. The "many worlds" may seem weird, but should we be surprised if a complete description of reality, something completely foreign to our experience, seems weird to us?

By including ourselves in the reality described by the wave function, we have understood why we perceive a definite outcome to a measurement, but there is still a further question: how does the concept of *probability* enter

into this (deterministic) formalism? This question remains troubling, for to answer it we must be prepared to state what is meant by "probability."

The word "probability" is used in two rather different senses. Sometimes *probability* means *frequency*. We say the probability of a coin coming up heads is 1/2 if we expect, as we toss the coin many times, the number of heads divided by the total number of tosses to converge to 1/2. (This is a tricky concept though; even if the probability is 1/2, the coin still *might* come up heads a trillion times in a row.) In rigorous mathematical discussions, probability theory often seems to be a branch of measure theory – it concerns the properties of infinite sequences.

But in everyday life, and also in quantum theory, probabilities typically are *not* frequencies. When we make a measurement, we do not repeat it an infinite number of times on identically prepared systems. In the Everett viewpoint, or in cosmology, there is just one universe, not many identically prepared ones.

So what *is* a probability? In practice, it is a number that quantifies the plausibility of a proposition given a state of knowledge. Perhaps surprisingly, this view can be made the basis of a well-defined mathematical theory, sometimes called the "Bayesian" view of probability. The term "Bayesian" reflects the way probability theory is typically used (both in science and in everyday life) – to test a hypothesis given some observed data. Hypothesis testing is carried out using Bayes's rule for conditional probability

$$P(A_0|B) = P(B|A_0)P(A_0)/P(B). \tag{3.179}$$

For example – suppose that A_0 is the preparation of a particular quantum state, and B is a particular outcome of a measurement of the state. We have made the measurement (obtaining B) and now we want to infer how the state was prepared (compute $P(A_0|B)$. Quantum mechanics allows us to compute $P(B|A_0)$. But it does *not* tell us $P(A_0)$ (or $P(B)$). We have to make a guess of $P(A_0)$, which is possible if we adopt a "principle of indifference" – if we have no knowledge that A_i is more or less likely than A_j we assume $P(A_i) = P(A_j)$. Once an ensemble of preparations is chosen, we can compute

$$P(B) = \sum_i P(B/A_i)P(A_i), \tag{3.180}$$

and so obtain $P(A_0|B)$ by applying Bayes's rule.

But if our attitude will be that probability theory quantifies plausibility given a state of knowledge, we are obligated to ask "whose state of knowledge?" To recover an objective theory, we must interpret probability in quantum theory not as a prediction based on our *actual* state of knowledge, but rather as a prediction based on the most complete *possible* knowledge about the quantum state. If we prepare $|\uparrow_x\rangle$ and measure $\boldsymbol{\sigma}_3$, then we say that the result is $|\uparrow_z\rangle$ with probability 1/2, not because that is the best prediction we can make based on what we know, but because it is the best prediction *anyone* can make, no matter how much they know. It is in this sense that the outcome is truly *random*; it cannot be predicted with certainty even when our knowledge is complete (in contrast to the pseudo randomness that arises in classical physics because our knowledge is incomplete).

So how, now, are we to extract probabilities from Everett's deterministic universe? Probabilities arise because we (a part of the system) cannot predict our future with certainty. I know the formalism, I know the Hamiltonian and wave function of the universe, I know my branch of the wave function. Now I am about to look at the cat. A second from now, I will be either be certain that the cat is dead or I will be certain that it is alive. Yet even with all I know, I cannot predict the future. Even with complete knowledge about the present, I cannot say what my state of knowledge will be after I look at the cat. The best I can do is assign probabilities to the outcomes. So, while the wave function of the universe *is* deterministic I, as a part of the system, can do no better than making probabilistic predictions.

Of course, as already noted, decoherence is a crucial part of this story. We may consistently assign probabilities to the alternatives Dead and Alive only if there is no (or at least negligible) possibility of interference among the alternatives. Probabilities make sense only when we can identify an exhaustive set of mutually exclusive alternatives. Since the issue is really whether interference might arise at a later time, we cannot decide whether probability theory applies by considering a quantum state at a fixed time; we must examine a set of mutually exclusive (coarse-grained) histories, or sequences of events. There is a sophisticated technology ("decoherence functionals") for adjudicating whether the various histories decohere to a sufficient extent for probabilities to be sensibly assigned.

So the Everett viewpoint *can* be reconciled with the quantum indeterminism that we observe, but there is still a troubling gap in the picture, at least as far as I can tell. I am about to look at the cat, and I know that the density matrix a second from now will be

$$|\text{Dead}\rangle_{\text{cat}}\ |\text{Know it's Dead}\rangle_{\text{me}}\ , \quad \text{Prob} = p_{\text{dead}},$$

$$|\text{Alive}\rangle_{\text{cat}}\ |\text{Know it's Alive}\rangle_{\text{me}}\ , \quad \text{Prob} = p_{\text{alive}}. \tag{3.181}$$

But how do I infer that p_{dead} and p_{alive} actually are probabilities that I (in my Bayesian posture) may assign to my future perceptions? I *still* need a rule to translate this density operator into probabilities assigned to the alternatives. It seems contrary to the Everett philosophy to *assume* such a rule; we could prefer to say that the only rule needed to define the theory is the Schrödinger equation (and perhaps a prescription to specify the initial wave function). Postulating a probability formula comes perilously close to allowing that there is a nondeterministic measurement process after all. So here is the issue regarding the foundations of theory for which I do not know a fully satisfying resolution.

Since we have not been able to remove all discomfiture concerning the origin of probability in quantum theory, it may be helpful to comment on an interesting suggestion due to Hartle. To implement his suggestion, we must return (perhaps with regret) to the frequency interpretation of probability. Hartle's insight is that we need not assume the probability interpretation as part of the measurement postulate. It is really sufficient to make a weaker assumption:

> If we prepare a quantum state $|a\rangle$, such that $\mathbf{A}|a\rangle = a|a\rangle$, and then immediately measure $\mathbf{A}$, the outcome of the measurement is a.

This seems like an assumption that a Bayesian residing in Everett's universe would accept. I am about to measure an observable, and the wavefunction will branch, but if the observable has the *same* value in every branch, then I *can* predict the outcome.

To implement a frequency interpretation of probability, we should, strictly speaking, consider an infinite number of trials. Suppose we want to make a statement about the probability of obtaining the result $|\uparrow_z\rangle$ when we measure $\boldsymbol{\sigma}_3$ in the state

$$|\psi\rangle = a|\uparrow_z\rangle + b|\downarrow_z\rangle. \tag{3.182}$$

Then we should imagine that an infinite number of copies are prepared, so the state is

$$|\psi^{(\infty)}\rangle \equiv (|\psi\rangle)^{\infty} = |\psi\rangle \otimes |\psi\rangle \otimes |\psi\rangle \otimes \cdots \tag{3.183}$$

and we imagine measuring $\boldsymbol{\sigma}_3$ for each of the copies. Formally, the case of an infinite number of trials can be formulated as the $N \to \infty$ limit of N trials.

Hartle's idea is to consider an "average spin" operator

$$\bar{\boldsymbol{\sigma}}_3 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\sigma}_3^{(i)}, \tag{3.184}$$

and to argue that $(|\psi\rangle)^N$ becomes an *eigenstate* of $\bar{\boldsymbol{\sigma}}_3$ with eigenvalue $|a|^2 - |b|^2$, as $N \to \infty$. Then we can invoke the weakened measurement postulate to infer that a measurement of $\bar{\boldsymbol{\sigma}}_3$ will yield the result $|a|^2 - |b|^2$ with certainty, and that the fraction of all the spins that point up is therefore $|a|^2$. In this sense, $|a|^2$ is the probability that the measurement of $\boldsymbol{\sigma}_3$ yields the outcome $|\uparrow_z\rangle$.

Consider, for example, the special case

$$|\psi_x^{(N)}\rangle \equiv (|\uparrow_x\rangle)^N = \left[\frac{1}{\sqrt{2}}(|\uparrow_z\rangle + |\downarrow_z\rangle)\right]^N. \tag{3.185}$$

We can compute

$$\begin{aligned} &\langle\psi_x^{(N)}|\bar{\boldsymbol{\sigma}}_3|\psi_x^{(N)}\rangle = 0 \ , \\ &\langle\psi_x^{(N)}|\bar{\boldsymbol{\sigma}}_3^2|\psi_x^{(N)}\rangle \\ &= \frac{1}{N^2}\langle\psi_x^{(N)}|\sum_{ij}\boldsymbol{\sigma}_3^{(i)}\boldsymbol{\sigma}_3^{(j)}|\psi_x^{(N)}\rangle \\ &= \frac{1}{N^2}\sum_{ij}\delta^{ij} = \frac{N}{N^2} = \frac{1}{N}. \end{aligned} \tag{3.186}$$

Taking the formal $N \to \infty$ limit, we conclude that $\bar{\boldsymbol{\sigma}}_3$ has vanishing dispersion about its mean value $\langle\bar{\boldsymbol{\sigma}}_3\rangle = 0$, and so at least in this sense $|\psi_x^{(\infty)}\rangle$ is an "eigenstate" of $\bar{\boldsymbol{\sigma}}_3$ with eigenvalue zero.

Coleman and Lesniewski have noted that one can take Hartle's argument a step further, and argue that the measurement outcome $|\uparrow_z\rangle$ not only occurs with the right frequency, but also that the $|\uparrow_z\rangle$ outcomes are *randomly distributed*. To make sense of this statement, we must formulate a definition of randomness. We say that an infinite string of bits is random if the string is *incompressible*; there is no simpler way to generate the first N bits than simply writing them out. We formalize this idea by considering the length

of the shortest computer program (on a certain universal computer) that generates the first N bits of the sequence. Then, for a random string

$$\text{Length of shortest program } > N - \text{ const.} \tag{3.187}$$

where the constant may depend on the particular computer used or on the particular sequence, but not on N.

Coleman and Lesniewski consider an orthogonal projection operator $\boldsymbol{E}_{\text{random}}$ that, acting on a state $|\psi\rangle$ that is an eigenstate of each $\boldsymbol{\sigma}_3^{(i)}$, satisfies

$$\boldsymbol{E}_{\text{random}}|\psi\rangle = |\psi\rangle, \tag{3.188}$$

if the sequence of eigenvalues of $\boldsymbol{\sigma}_3^{(i)}$ is random, and

$$\boldsymbol{E}_{\text{random}}|\psi\rangle = 0, \tag{3.189}$$

if the sequence is not random. This property alone is not sufficient to determine how $\boldsymbol{E}_{\text{random}}$ acts on all of $(\mathcal{H}_2)^\infty$, but with an additional technical assumption, they find that $\boldsymbol{E}_{\text{random}}$ exists, is unique, and has the property

$$\boldsymbol{E}_{\text{random}}|\psi_x^{(\infty)}\rangle = |\psi_x^{(\infty)}\rangle. \tag{3.190}$$

Thus, we "might as well say" that $|\psi_x^{(\infty)}\rangle$ is random, with respect to $\boldsymbol{\sigma}_3$ measurements – a procedure for distinguishing the random states from nonrandom ones that works properly for strings of $\boldsymbol{\sigma}_3$ eigenstates, will inevitably identify $|\psi_x^{(\infty)}\rangle$ as random, too.

These arguments are interesting, but they do not leave me completely satisfied. The most disturbing thing is the need to consider infinite sequences (a feature of any frequency interpretation probability). For any finite N, we are unable to apply Hartle's weakened measurement postulate, and even in the limit $N \to \infty$, applying the postulate involves subtleties. It would be preferable to have *a stronger* weakened measurement postulate that could be applied at finite N, but I am not sure how to formulate that postulate or how to justify it.

In summary then: Physics should describe the objective physical world, and the best representation of physical reality that we know about is the quantum-mechanical wave function. Physics should aspire to explain all observed phenomena as economically as possible – it is therefore unappealing to postulate that the measurement process is governed by different dynamical principles than other processes. Fortunately, everything we know about physics is compatible with the hypothesis that all physical processes (including measurements) can be accurately modeled by the unitary evolution of a wave function (or density matrix). When a microscopic quantum system interacts with a macroscopic apparatus, decoherence drives the "collapse" of the wave function "for all practical purposes."

If we eschew measurement as a mystical primitive process, and we accept the wave function as a description of physical reality, then we are led to the Everett or "many-worlds" interpretation of quantum theory. In this view, all possible outcomes of any "measurement" are regarded as "real" — but I perceive only a specific outcome because the state of my brain (a part of the quantum system) is strongly correlated with the outcome.

Although the evolution of the wave function in the Everett interpretation is deterministic, I am unable to predict with certainty the outcome of an experiment to be performed in the future – I don't know what branch of the wavefunction I will end up on, so I am unable to predict my future state of mind. Thus, while the "global" picture of the universe is in a sense deterministic, from my own local perspective from within the system, I perceive quantum mechanical randomness.

My own view is that the Everett interpretation of quantum theory provides a satisfying explanation of measurement and of the origin of randomness, but does not yet fully explain the quantum mechanical rules for computing probabilities. A full explanation should go beyond the frequency interpretation of probability — ideally it would place the Bayesian view of probability on a secure objective foundation.

3.7 Summary

POVM. If we restrict our attention to a subspace of a larger Hilbert space, then an orthogonal (Von Neumann) measurement performed on the larger space cannot in general be described as an orthogonal measurement on the subspace. Rather, it is a *generalized measurement* or *POVM* – the outcome a occurs with a probability

$$\text{Prob}(a) = tr\left(\boldsymbol{F}_a \boldsymbol{\rho}\right) \ , \tag{3.191}$$

where $\boldsymbol{\rho}$ is the density matrix of the subsystem, each $\boldsymbol{F}_a$ is a positive hermitian operator, and the $\boldsymbol{F}_a$'s satisfy

$$\sum_a \boldsymbol{F}_a = \mathbf{1} \ . \tag{3.192}$$

A POVM in $\mathcal{H}_A$ can be realized as a unitary transformation on the tensor product $\mathcal{H}_A \otimes \mathcal{H}_B$, followed by an orthogonal measurement in $\mathcal{H}_B$.

Superoperator. Unitary evolution on $\mathcal{H}_A \otimes \mathcal{H}_B$ will not in general appear to be unitary if we restrict our attention to $\mathcal{H}_A$ alone. Rather, evolution in $\mathcal{H}_A$ will be described by a *superoperator*, (which can be inverted by another superoperator only if unitary). A general superoperator \$ has an operator-sum (Kraus) representation:

$$\$: \boldsymbol{\rho} \to \$(\boldsymbol{\rho}) = \sum_\mu \boldsymbol{M}_\mu \boldsymbol{\rho} \boldsymbol{M}_\mu^\dagger \ , \tag{3.193}$$

where

$$\sum_\mu \boldsymbol{M}_\mu^\dagger \boldsymbol{M}_\mu = \mathbf{1} \ . \tag{3.194}$$

In fact, any reasonable (linear and completely positive) mapping of density matrices to density matrices has unitary and operator-sum representations.

Decoherence. Decoherence – the decay of quantum information due to the interaction of a system with its environment – can be described by a superoperator. If the environment frequently "scatters" off the system, and the state of the environment is not monitored, then off-diagonal terms in the density matrix of the system decay rapidly in a preferred basis (typically a spatially localized basis selected by the nature of the coupling of the system to the environment). The time scale for decoherence is set by the scattering rate, which may be much larger than the damping rate for the system.

Master Equation. When the relevant dynamical time scale of an open quantum system is long compared to the time for the environment to "forget" quantum information, the evolution of the system is effectively local in time (the Markovian approximation). Much as general unitary evolution is generated by a Hamiltonian, a general Markovian superoperator is generated by a *Lindbladian* $\mathcal{L}$ as described by the *master equation*:

$$\dot{\boldsymbol{\rho}} \equiv \mathcal{L}[\boldsymbol{\rho}] = -i[\boldsymbol{H}, \boldsymbol{\rho}] + \sum_\mu \left(\boldsymbol{L}_\mu \boldsymbol{\rho} \boldsymbol{L}_\mu^\dagger - \frac{1}{2}\boldsymbol{L}_\mu^\dagger \boldsymbol{L}_\mu \boldsymbol{\rho} - \frac{1}{2}\boldsymbol{\rho} \boldsymbol{L}_\mu^\dagger \boldsymbol{L}_\mu \right) . \tag{3.195}$$

Here each *Lindblad operator* (or *quantum jump operator*) represents a "quantum jump" that could in principle be detected if we monitored the environment faithfully. By solving the master equation, we can compute the decoherence rate of an open system.

3.8 Exercises

3.1 Realization of a POVM

Consider the POVM defined by the four positive operators

$$\begin{aligned} P_1 &= \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| \ , \quad P_2 = \frac{1}{2}|\downarrow_z\rangle\langle\downarrow_z| \\ P_3 &= \frac{1}{2}|\uparrow_x\rangle\langle\uparrow_x| \ , \quad P_4 = \frac{1}{2}|\downarrow_x\rangle\langle\downarrow_x| \ . \end{aligned} \tag{3.196}$$

Show how this POVM can be realized as an orthogonal measurement in a two-qubit Hilbert space, if one ancilla spin is introduced.

3.2 Invertibility of superoperators

The purpose of this exercise is to show that a superoperator is invertible only if it is unitary. Recall that any superoperator has an operator-sum representation; it acts on a pure state as

$$\mathcal{M}(|\psi\rangle\langle\psi|) = \sum_\mu \mathbf{M}_\mu|\psi\rangle\langle\psi|\mathbf{M}_\mu^\dagger, \tag{3.197}$$

where $\sum_\mu \mathbf{M}_\mu^\dagger\mathbf{M}_\mu = \mathbf{1}$. Another superoperator $\mathcal{N}$ is said to be the inverse of $\mathcal{M}$ if $\mathcal{N}\circ\mathcal{M} = I$, or

$$\sum_{\mu,a} \mathbf{N}_a\mathbf{M}_\mu|\psi\rangle\langle\psi|\mathbf{M}_\mu^\dagger\mathbf{N}_a^\dagger = |\psi\rangle\langle\psi|, \tag{3.198}$$

for any $|\psi\rangle$. It follows that

$$\sum_{\mu,a} |\langle\psi|\mathbf{N}_a\mathbf{M}_\mu|\psi\rangle|^2 = 1. \tag{3.199}$$

a) Show, using the normalization conditions satisfied by the $\mathbf{N}_a$'s and $\mathbf{M}_\mu$'s, that $\mathcal{N}\circ\mathcal{M} = I$ implies that

$$\mathbf{N}_a\mathbf{M}_\mu = \lambda_{a\mu}\mathbf{1}, \tag{3.200}$$

for *each* a and μ; i.e., that each $\mathbf{N}_a\mathbf{M}_\mu$ is a multiple of the identity.

b) Use the result of (a) to show that $\mathbf{M}_\nu^\dagger\mathbf{M}_\mu$ is proportional to the identity for each μ and ν.

c) Show that it follows from (b) that $\mathcal{M}$ is unitary.

3.3 How many superoperators?

How many real parameters are needed to parametrize the general superoperator

$$\$: \boldsymbol{\rho} \to \boldsymbol{\rho}' \ , \tag{3.201}$$

if $\boldsymbol{\rho}$ is a density operator in a Hilbert space of dimension N? [Hint: How many real parameters parametrize an $N\times N$ Hermitian matrix? How many for a linear mapping of Hermitian matices to Hermitian matrices? How many for a *trace-preserving* mapping of Hermitian matrices to Hermitian matrices?]

3.4 How fast is decoherence?

A very good pendulum with mass $m = 1$ g and circular frequency $\omega = 1\ \mathrm{s}^{-1}$ has quality factor $Q = 10^9$. The pendulum is prepared in the "cat state"

$$|\mathrm{cat}\rangle = \frac{1}{\sqrt{2}}(|x\rangle + |-x\rangle), \tag{3.202}$$

a superposition of minimum uncertainty wave packets, each initially at rest, centered at positions $\pm x$, where $x = 1$ cm. Estimate, in order of magnitude, how long it takes for the cat state to decohere, if the environment is at

a) zero temperature.

b) room temperature.

3.5 Phase damping

In class, we obtained an operator-sum representation of the phase-damping channel for a single qubit, with Kraus operators

$$\mathbf{M}_0 = \sqrt{1-p}\ \mathbf{1}, \quad \mathbf{M}_1 = \sqrt{p}\ \frac{1}{2}(\mathbf{1}+\boldsymbol{\sigma}_3),$$

$$\mathbf{M}_2 = \sqrt{p}\ \frac{1}{2}(\mathbf{1}-\boldsymbol{\sigma}_3). \tag{3.203}$$

a) Find an alternative representation using only two Kraus operators $\mathbf{N}_0, \mathbf{N}_1$.

b) Find a unitary 3×3 matrix $U_{\mu a}$ such that your Kraus operators found in (a) (augmented by $\mathbf{N}_2 = 0$) are related to $\mathbf{M}_{0,1,2}$ by

$$\mathbf{M}_\mu = U_{\mu a}\mathbf{N}_a. \tag{3.204}$$

c) Consider a single-qubit channel with a unitary representation

$$|0\rangle_A|0\rangle_E \to \sqrt{1-p}\ |0\rangle_A|0\rangle_E + \sqrt{p}\ |0\rangle_A|\gamma_0\rangle_E$$

$$|1\rangle_A|0\rangle_E \to \sqrt{1-p}\ |1\rangle_A|0\rangle_E + \sqrt{p}\ |1\rangle_A|\gamma_1\rangle_E, \tag{3.205}$$

where $|\gamma_0\rangle_E$ and $|\gamma_1\rangle_E$ are normalized states, both orthogonal to $|0\rangle_E$, that satisfy

$${}_E\langle\gamma_0|\gamma_1\rangle_E = 1-\varepsilon, \quad 0 < \varepsilon < 1. \tag{3.206}$$

Show that this is again the phase-damping channel, and find its operator-sum representation with two Kraus operators.

d) Suppose that the channel in (c) describes what happens to the qubit when a single photon scatters from it. Find the decoherence rate Γ_{decoh} in terms of the scattering rate Γ_{scatt}.

3.6 Decoherence on the Bloch sphere

Parametrize the density matrix of a single qubit as

$$\boldsymbol{\rho} = \frac{1}{2}\left(\mathbf{1} + \vec{P}\cdot\vec{\boldsymbol{\sigma}}\right). \tag{3.207}$$

a) Describe what happens to $\vec{P}$ under the action of the phase-damping channel.

b) Describe what happens to $\vec{P}$ under the action of the amplitude-damping channel defined by the Kraus operators.

$$\mathbf{M}_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}, \quad \mathbf{M}_1 = \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix}. \tag{3.208}$$

c) The same for the "two-Pauli channel:"

$$\mathbf{M}_0 = \sqrt{1-p}\ \mathbf{1}, \quad \mathbf{M}_1 = \sqrt{\frac{p}{2}}\ \boldsymbol{\sigma}_1, \quad \mathbf{M}_2 = \sqrt{\frac{p}{2}}\ \boldsymbol{\sigma}_3. \tag{3.209}$$

3.7 Decoherence of the damped oscillator

We saw in class that, for an oscillator that can emit quanta into a zero-temperature reservoir, the interaction picture density matrix $\boldsymbol{\rho}_I(t)$ of the oscillator obeys the master equation

$$\dot{\boldsymbol{\rho}}_I = \Gamma\left(\boldsymbol{a}\boldsymbol{\rho}_I\boldsymbol{a}^\dagger - \frac{1}{2}\boldsymbol{a}^\dagger\boldsymbol{a}\boldsymbol{\rho}_I - \frac{1}{2}\boldsymbol{\rho}_I\boldsymbol{a}^\dagger\boldsymbol{a}\right), \tag{3.210}$$

where $\boldsymbol{a}$ is the annihilation operator of the oscillator.

a) Consider the quantity

$$X(\lambda,t) = tr\left[\boldsymbol{\rho}_I(t)e^{\lambda\boldsymbol{a}^\dagger}e^{-\lambda^*\boldsymbol{a}}\right], \tag{3.211}$$

(where λ is a complex number). Use the master equation to derive and solve a differential equation for $X(\lambda,t)$. You should find

$$X(\lambda,t) = X(\lambda',0), \tag{3.212}$$

where λ' is a function of λ, Γ, and t. What is this function $\lambda'(\lambda,\Gamma,t)$?

b) Now suppose that a "cat state" of the oscillator is prepared at $t=0$:

$$|\text{cat}\rangle = \frac{1}{\sqrt{2}}\left(|\alpha_1\rangle + |\alpha_2\rangle\right), \tag{3.213}$$

where $|\alpha\rangle$ denotes the coherent state

$$|\alpha\rangle = e^{-|\alpha|^2/2}e^{\alpha\boldsymbol{a}^\dagger}|0\rangle. \tag{3.214}$$

Use the result of (a) to infer the density matrix at a later time t. Assuming $\Gamma t \ll 1$, at what rate do the off-diagonal terms in $\boldsymbol{\rho}$ decay (in this coherent state basis)?

Chapter 4

Quantum Entanglement

4.1 Nonseparability of EPR pairs

4.1.1 Hidden quantum information

The deep ways that quantum information differs from classical information involve the properties, implications, and uses of *quantum entanglement.* Recall from §2.4.1 that a bipartite pure state is *entangled* if its Schmidt number is greater than one. Entangled states are interesting because they exhibit correlations that have no classical analog. We will begin the study of these correlations in this chapter.

Recall, for example, the *maximally entangled* state of two qubits defined in §3.4.1:

$$|\phi^+\rangle_{AB} = \frac{1}{\sqrt{2}}(|00\rangle_{AB} + |11\rangle_{AB}). \tag{4.1}$$

"Maximally entangled" means that when we trace over qubit B to find the density operator $\boldsymbol{\rho}_A$ of qubit A, we obtain a multiple of the identity operator

$$\boldsymbol{\rho}_A = tr_B(|\phi^+\rangle_{AB\ AB}\langle\phi^+|) = \frac{1}{2}\mathbf{1}_A, \tag{4.2}$$

(and similarly $\boldsymbol{\rho}_B = \frac{1}{2}\mathbf{1}_B$). This means that if we measure spin A along *any* axis, the result is completely random – we find spin up with probability 1/2 and spin down with probability 1/2. Therefore, if we perform any local measurement of A or B, we acquire no information about the preparation of the state, instead we merely generate a random bit. This situation contrasts

sharply with case of a single qubit in a pure state; there we can store a bit by preparing, say, either $|\uparrow_{\hat{n}}\rangle$ or $|\downarrow_{\hat{n}}\rangle$, and we can recover that bit reliably by measuring along the $\hat{n}$-axis. With two qubits, we ought to be able to store two bits, but in the state $|\phi^+\rangle_{AB}$ this information is *hidden*; at least, we can't acquire it by measuring A or B.

In fact, $|\phi^+\rangle$ is one member of a basis of four mutually orthogonal states for the two qubits, all of which are maximally entangled — the basis

$$\begin{aligned} |\phi^\pm\rangle &= \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle), \\ |\psi^\pm\rangle &= \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle), \end{aligned} \tag{4.3}$$

introduced in §3.4.1. We can choose to prepare one of these four states, thus encoding two bits in the state of the two-qubit system. One bit is the *parity* bit ($|\phi\rangle$ or $|\psi\rangle$) – are the two spins aligned or antialigned? The other is the *phase* bit ($+$ or $-$) – what superposition was chosen of the two states of like parity. Of course, we can recover the information by performing an orthogonal measurement that projects onto the $\{|\phi^+\rangle, |\phi^-\rangle, |\psi^+\rangle, |\psi^-\rangle\}$ basis. But if the two qubits are distantly separated, we cannot acquire this information locally; that is, by measuring A or measuring B.

What we *can* do locally is *manipulate* this information. Suppose that Alice has access to qubit A, but not qubit B. She may apply $\boldsymbol{\sigma}_3$ to her qubit, flipping the relative phase of $|0\rangle_A$ and $|1\rangle_A$. This action flips the phase bit stored in the entangled state:

$$\begin{aligned} |\phi^+\rangle &\leftrightarrow |\phi^-\rangle, \\ |\psi^+\rangle &\leftrightarrow |\psi^-\rangle. \end{aligned} \tag{4.4}$$

On the other hand, she can apply $\boldsymbol{\sigma}_1$, which flips her spin ($|0\rangle_A \leftrightarrow |1\rangle_A$), and also flips the parity bit of the entangled state:

$$\begin{aligned} |\phi^+\rangle &\leftrightarrow |\psi^+\rangle, \\ |\phi^-\rangle &\leftrightarrow -|\psi^-\rangle. \end{aligned} \tag{4.5}$$

Bob can manipulate the entangled state similarly. In fact, as we discussed in §2.4, either Alice or Bob can perform a local unitary transformation that changes one maximally entangled state to any other maximally entangled state.[1] What their local unitary transformations *cannot* do is alter $\boldsymbol{\rho}_A = \boldsymbol{\rho}_B = \frac{1}{2}\mathbf{1}$ – the information they are manipulating is information that neither one can read.

But now suppose that Alice and Bob are able to exchange (classical) messages about their measurement outcomes; together, then, they can learn about how their measurements are correlated. The entangled basis states are conveniently characterized as the simultaneous eigenstates of two commuting observables:

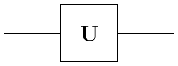
$$\begin{aligned} &\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)}, \\ &\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}; \end{aligned} \tag{4.6}$$

the eigenvalue of $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$ is the parity bit, and the eigenvalue of $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)}$ is the phase bit. Since these operators commute, they can in principle be measured simultaneously. But they cannot be measured simultaneously if Alice and Bob perform localized measurements. Alice and Bob could both choose to measure their spins along the z-axis, preparing a simultaneous eigenstate of $\boldsymbol{\sigma}_3^{(A)}$ and $\boldsymbol{\sigma}_3^{(B)}$. Since $\boldsymbol{\sigma}_3^{(A)}$ and $\boldsymbol{\sigma}_3^{(B)}$ both commute with the parity operator $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$, their orthogonal measurements do not disturb the parity bit, and they can combine their results to infer the parity bit. But $\boldsymbol{\sigma}_3^{(A)}$ and $\boldsymbol{\sigma}_3^{(B)}$ do *not* commute with phase operator $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)}$, so their measurement disturbs the phase bit. On the other hand, they could both choose to measure their spins along the x-axis; then they would learn the phase bit at the cost of disturbing the parity bit. But they can't have it both ways. To have hope of acquiring the parity bit without disturbing the phase bit, they would need to learn about the product $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$ without finding out anything about $\boldsymbol{\sigma}_3^{(A)}$ and $\boldsymbol{\sigma}_3^{(B)}$ separately. That cannot be done locally.

Now let us bring Alice and Bob together, so that they can operate on their qubits jointly. How might they acquire both the parity bit and the phase bit of their pair? By applying an appropriate unitary transformation, they can rotate the entangled basis $\{|\phi^\pm\rangle, |\psi^\pm\rangle\}$ to the unentangled basis $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$. Then they can measure qubits A and B separately to acquire the bits they seek. How is this transformation constructed?

[1]But of course, this does not suffice to perform an arbitrary unitary transformation on the four-dimensional space $\mathcal{H}_A \otimes \mathcal{H}_B$, which contains states that are not maximally entangled. The maximally entangled states are *not* a subspace – a superposition of maximally entangled states typically is *not* maximally entangled.

This is a good time to introduce notation that will be used heavily later in the course, the quantum circuit notation. Qubits are denoted by horizontal lines, and the single-qubit unitary transformation $\mathbf{U}$ is denoted:



A particular single-qubit unitary we will find useful is the *Hadamard transform*

$$\boldsymbol{H} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{\sqrt{2}}(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_3), \tag{4.7}$$

which has the properties

$$\boldsymbol{H}^2 = \mathbf{1}, \tag{4.8}$$

and

$$\begin{aligned} \boldsymbol{H}\boldsymbol{\sigma}_1\boldsymbol{H} &= \boldsymbol{\sigma}_3, \\ \boldsymbol{H}\boldsymbol{\sigma}_3\boldsymbol{H} &= \boldsymbol{\sigma}_1. \end{aligned} \tag{4.9}$$

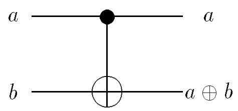
(We can envision $\boldsymbol{H}$ (up to an overall phase) as a $\theta = \pi$ rotation about the axis $\hat{n} = \frac{1}{\sqrt{2}}(\hat{n}_1 + \hat{n}_3)$ that rotates $\hat{x}$ to $\hat{z}$ and vice-versa; we have

$$R(\hat{n}, \theta) = \mathbf{1}\cos\frac{\theta}{2} + i\hat{n}\cdot\vec{\boldsymbol{\sigma}}\sin\frac{\theta}{2} = i\frac{1}{\sqrt{2}}(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_3) = i\boldsymbol{H}.) \tag{4.10}$$

Also useful is the two-qubit transformation known as the XOR or controlled-NOT transformation; it acts as

$$\mathbf{CNOT} : |a, b\rangle \to |a, a \oplus b\rangle, \tag{4.11}$$

on the basis states $a, b = 0, 1$, where $a \oplus b$ denotes addition modulo 2, and is denoted:

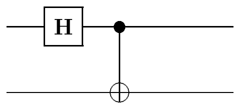


Thus this gate flips the second bit if the first is 1, and acts trivially if the first bit is 0; we see that

$$(\mathbf{CNOT})^2 = \mathbf{1}. \tag{4.12}$$

We call a the *control* (or source) bit of the **CNOT**, and b the *target* bit.

By composing these "primitive" transformations, or quantum *gates*, we can build other unitary transformations. For example, the "circuit"



(to be read from left to right) represents the product of $\boldsymbol{H}$ applied to the first qubit followed by **CNOT** with the first bit as the source and the second bit as the target. It is straightforward to see that this circuit transforms the standard basis to the entangled basis,

$$\begin{aligned}
|00\rangle &\to \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)|0\rangle \to |\phi^+\rangle,\\
|01\rangle &\to \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)|1\rangle \to |\psi^+\rangle,\\
|10\rangle &\to \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)|0\rangle \to |\phi^-\rangle,\\
|11\rangle &\to \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)|1\rangle \to |\psi^-\rangle,
\end{aligned} \tag{4.13}$$

so that the first bit becomes the phase bit in the entangled basis, and the second bit becomes the parity bit.

Similarly, we can invert the transformation by running the circuit backwards (since both **CNOT** and $\boldsymbol{H}$ square to the identity); if we apply the inverted circuit to an entangled state, and then measure both bits, we can learn the value of both the phase bit and the parity bit.

Of course, $\boldsymbol{H}$ acts on only one of the qubits; the "nonlocal" part of our circuit is the controlled-NOT gate – this is the operation that establishes or removes entanglement. If we could only perform an "interstellar **CNOT**," we would be able to create entanglement among distantly separated pairs, or

extract the information encoded in entanglement. But we can't. To do its job, the **CNOT** gate must act on its target without revealing the value of its source. Local operations and classical communication will not suffice.

4.1.2 Einstein locality and hidden variables

Einstein was disturbed by quantum entanglement. Eventually, he along with Podolsky and Rosen sharpened their discomfort into what they regarded as a paradox. As later reinterpreted by Bohm, the situation they described is really the same as that discussed in §2.5.3. Given a maximally entangled state of two qubits shared by Alice and Bob, Alice can choose one of several possible measurements to perform on her spin that will realize different possible ensemble interpretations of Bob's density matrix; for example, she can prepare either $\boldsymbol{\sigma}_1$ or $\boldsymbol{\sigma}_3$ eigenstates.

We have seen that Alice and Bob are unable to exploit this phenomenon for faster-than-light communication. Einstein knew this but he was still dissatisfied. He felt that in order to be considered a *complete* description of physical reality a theory should meet a stronger criterion, that might be called Einstein locality:

> Suppose that A and B are spacelike separated systems. Then in a *complete* description of physical reality an action performed on system A must not modify the description of system B.

But if A and B are entangled, a measurement of A is performed and a *particular* outcome is known to have been obtained, then the density matrix of B *does* change. Therefore, by Einstein's criterion, the description of a quantum system by a wavefunction cannot be considered complete.

Einstein seemed to envision a more complete description that would remove the indeterminacy of quantum mechanics. A class of theories with this feature are called *local hidden-variable theories*. In a hidden variable theory, measurement is actually fundamentally deterministic, but appears to be probabilistic because some degrees of freedom are not precisely known. For example, perhaps when a spin is prepared in what quantum theory would describe as the pure state $|\uparrow_{\hat{z}}\rangle$, there is actually a deeper theory in which the state prepared is parametrized as $(\hat{z}, \lambda)$ where λ $(0 \le \lambda \le 1)$ is the hidden variable. Suppose that with present-day experimental technique, we have no control over λ, so when we prepare the spin state, λ might take any value – the probability distribution governing its value is uniform on the unit interval.

Now suppose that when we measure the spin along an axis rotated by θ from the $\hat{z}$ axis, the outcome will be

$$
\begin{aligned}
&|\uparrow_\theta\rangle, \text{ for } 0 \leq \lambda \leq \cos^2\frac{\theta}{2}\\
&|\downarrow_\theta\rangle, \text{ for } \cos^2\frac{\theta}{2} < \lambda \leq 1. \qquad (4.14)
\end{aligned}
$$

If we know λ, the outcome is deterministic, but if λ is completely unknown, then the probability distribution governing the measurement will agree with the predictions of quantum theory.

Now, what about entangled states? When we say that a hidden variable theory is *local*, we mean that it satisfies the Einstein locality constraint. A measurement of A does not modify the values of the variables that govern the measurements of B. This seems to be what Einstein had in mind when he envisioned a more complete description.

4.1.3 Bell Inequalities

John Bell's fruitful idea was to test Einstein locality by considering the quantitative properties of the correlations between measurement outcomes obtained by Bob and Alice.[2] Let's first examine the predictions of quantum mechanics regarding these correlations.

Note that the state $|\psi^-\rangle$ has the properties

$$
(\vec{\boldsymbol{\sigma}}^{(A)} + \vec{\boldsymbol{\sigma}}^{(B)})|\psi^-\rangle = 0, \qquad (4.15)
$$

as we can see by explicit computation. Now consider the expectation value

$$
\langle\phi^-|(\vec{\boldsymbol{\sigma}}^{(A)}\cdot\hat{n})(\vec{\boldsymbol{\sigma}}^{(B)}\cdot\hat{m})|\psi^-\rangle. \qquad (4.16)
$$

Since we can replace $\vec{\boldsymbol{\sigma}}^{(B)}$ by $-\vec{\boldsymbol{\sigma}}^{(A)}$ acting on $|\psi^-\rangle$, this can be expressed as

$$
\begin{aligned}
&-\langle(\vec{\boldsymbol{\sigma}}^{(A)}\cdot\hat{n})(\vec{\boldsymbol{\sigma}}^{(A)}\cdot\hat{m})\rangle =\\
&-n_i m_j tr(\boldsymbol{\rho}_A\boldsymbol{\sigma}_i^{(A)}\boldsymbol{\sigma}_j^{(A)}) = -n_i m_j\delta_{ij} = -\hat{n}\cdot\hat{m} = -\cos\theta, \qquad (4.17)
\end{aligned}
$$

[2] A good reference on Bell inequalities is A. Peres, *Quantum Theory: Concepts and Methods*, chapter 6.

where θ is the angle between the axes $\hat{n}$ and $\hat{m}$. Thus we find that the measurement outcomes are always perfectly anticorrelated when we measure both spins along the same axis $\hat{n}$, and we have also obtained a more general result that applies when the two axes are different. Since the projection operator onto the spin up (spin down) states along $\hat{n}$ is $\boldsymbol{E}(\hat{n},\pm) = \frac{1}{2}(\mathbf{1}\pm\hat{n}\cdot\vec{\boldsymbol{\sigma}})$, we also obtain

$$
\begin{aligned}
&\langle\psi^-|\boldsymbol{E}^{(A)}(\hat{n},+)\boldsymbol{E}^{(B)}(\hat{m},+)|\psi^-\rangle\\
&= \langle\psi^-|\boldsymbol{E}^{(A)}(\hat{n},-)\boldsymbol{E}^{(B)}(\hat{m},-)|\psi^-\rangle = \frac{1}{4}(1-\cos\theta),\\
&\langle\psi^-|\boldsymbol{E}^{(A)}(\hat{n},+)\boldsymbol{E}^{(B)}(\hat{m},-)|\psi^-\rangle\\
&= \langle\psi^-|\boldsymbol{E}^{(A)}(\hat{n},-)\boldsymbol{E}^{(B)}(\hat{m},+)|\psi-\rangle = \frac{1}{4}(1+\cos\theta); \qquad (4.18)
\end{aligned}
$$

The probability that the outcomes are opposite is $\frac{1}{2}(1+\cos\theta)$, and the probability that the outcomes are the same is $\frac{1}{2}(1-\cos\theta)$.

Now suppose Alice will measure her spin along one of the three axes in the $x-z$ plane,

$$
\begin{aligned}
\hat{n}_1 &= (0,0,1)\\
\hat{n}_2 &= \left(\frac{\sqrt{3}}{2},0,-\frac{1}{2}\right)\\
\hat{n}_3 &= \left(-\frac{\sqrt{3}}{2},0,-\frac{1}{2}\right). \qquad (4.19)
\end{aligned}
$$

Once she performs the measurement, she disturbs the state of the spin, so she won't have a chance to find out what would have happened if she had measured along a different axis. Or will she? If she shares the state $|\psi^-\rangle$ with Bob, then Bob can help her. If Bob measures along, say, $\hat{n}_2$, and sends the result to Alice, then Alice knows what would have happened *if* she had measured along $\hat{n}_2$, since the results are perfectly anticorrelated. Now she can go ahead and measure along $\hat{n}_1$ as well. According to quantum mechanics, the probability that measuring along $\hat{n}_1$, and $\hat{n}_2$ give the *same* result is

$$
P_{same} = \frac{1}{2}(1-\cos\theta) = \frac{1}{4}. \qquad (4.20)
$$

(We have $\cos\theta = 1/2$ because Bob measures along $-\hat{n}_2$ to obtain Alice's result for measuring along $\hat{n}_2$). In the same way, Alice and Bob can work together to determine outcomes for the measurement of Alice's spin along any two of the axes $\hat{n}_1, \hat{n}_2$, and $\hat{n}_3$.

It is as though three coins are resting on a table; each coin has either the heads (H) or tails (T) side facing up, but the coins are covered, at first, so we don't know which. It is possible to reveal two of the coins (measure the spin along two of the axes) to see if they are H or T, but then the third coin always disappears before we get a chance to uncover it (we can't measure the spin along the third axis).

Now suppose that there are actually local hidden variables that provide a *complete* description of this system, and the quantum correlations are to arise from a probability distribution governing the hidden variables. Then, in this context, the Bell inequality is the statement

$$P_{same}(1,2) + P_{same}(1,3) + P_{same}(2,3) \geq 1, \tag{4.21}$$

where $P_{same}(i,j)$ denotes the probability that coins i and j have the *same* value (HH or TT). This is satisfied by any probability distribution for the three coins because no matter what the values of the coins, there will always be two that are the same. But in quantum mechanics,

$$P_{same}(1,2) + P_{same}(1,3) + P_{same}(2,3) = 3 \cdot \frac{1}{4} = \frac{3}{4} < 1. \tag{4.22}$$

We have found that the correlations predicted by quantum theory are incompatible with the local hidden variable hypothesis.

What are the implications? To some people, the peculiar correlations unmasked by Bell's theorem call out for a deeper explanation than quantum mechanics seems to provide. They see the EPR phenomenon as a harbinger of new physics awaiting discovery. But they may be wrong. We have been waiting over 60 years since EPR, and so far no new physics.

Perhaps we have learned that it can be dangerous to reason about what might have happened, but didn't actually happen. (Of course, we do this all the time in our everyday lives, and we usually get away with it, but sometimes it gets us into trouble.) I claimed that Alice knew what would happen when she measured along $\hat{n}_2$, because Bob measured along $\hat{n}_2$, and every time we have ever checked, their measurement outcomes are always perfectly anticorrelated. But Alice did *not* measure along $\hat{n}_2$; she measured along $\hat{n}_1$ instead. We got into trouble by trying to assign probabilities to the outcomes of measurements along $\hat{n}_1, \hat{n}_2$, and $\hat{n}_3$, even though we can only

perform one of those measurements. This turned out to lead to mathematical inconsistencies, so we had better not do it. From this viewpoint we have affirmed Bohr's principle of *complementary* — we are forbidden to consider simultaneously the possible outcomes of two mutually exclusive experiments.

Another common attitude is that the violations of the Bell inequalities (confirmed experimentally) have exposed an essential nonlocality built into the quantum description of Nature. One who espouses this view has implicitly rejected the complementarity principle. *If* we do insist on talking about outcomes of mutually exclusive experiments *then* we are forced to conclude that Alice's choice of measurement actually exerted a subtle *influence* on the outcome of Bob's measurement. This is what is meant by the "nonlocality" of quantum theory.

By ruling out local hidden variables, Bell demolished Einstein's dream that the indeterminacy of quantum theory could be eradicated by adopting a more complete, yet still local, description of Nature. If we accept locality as an inviolable principle, then we are forced to accept randomness as an unavoidable and intrinsic feature of quantum measurement, rather than a consequence of incomplete knowledge.

The human mind seems to be poorly equipped to grasp the correlations exhibited by entangled quantum states, and so we speak of the weirdness of quantum theory. But whatever your attitude, experiment forces you to accept the existence of the weird correlations among the measurement outcomes. There is no big mystery about how the correlations were established – we saw that it was necessary for Alice and Bob to get together at some point to create entanglement among their qubits. The novelty is that, even when A and B are distantly separated, we cannot accurately regard A and B as two separate qubits, and use classical information to characterize how they are correlated. They are more than just correlated, they are a single *inseparable* entity. They are *entangled*.

4.1.4 Photons

Experiments that test the Bell inequality are done with entangled photons, not with spin$-\frac{1}{2}$ objects. What are the quantum-mechanical predictions for photons?

Suppose, for example, that an excited atom emits two photons that come out back to back, with vanishing angular momentum and even parity. If $|x\rangle$ and $|y\rangle$ are horizontal and vertical linear polarization states of the photon, then we have seen that

$$
\begin{aligned}
|+\rangle &= \frac{1}{\sqrt{2}}(|x\rangle + i|y\rangle),\\
|-\rangle &= \frac{1}{\sqrt{2}}(i|x\rangle + |y\rangle),
\end{aligned}
\tag{4.23}
$$

are the eigenstates of *helicity* (angular momentum along the axis of propagation $\hat{z}$. For two photons, one propagating in the $+\hat{z}$ direction, and the other in the $-\hat{z}$ direction, the states

$$
\begin{aligned}
&|+\rangle_A|-\rangle_B\\
&|-\rangle_A|+\rangle_B
\end{aligned}
\tag{4.24}
$$

are invariant under rotations about $\hat{z}$. (The photons have opposite values of J_z, but the *same* helicity, since they are propagating in opposite directions.) Under a reflection in the $y-z$ plane, the polarization states are modified according to

$$
\begin{aligned}
|x\rangle \to -|x\rangle, &\quad |+\rangle \to +i|-\rangle,\\
|y\rangle \to |y\rangle, &\quad |-\rangle \to -i|+\rangle;
\end{aligned}
\tag{4.25}
$$

therefore, the parity eigenstates are *entangled* states

$$
\frac{1}{\sqrt{2}}(|+\rangle_A|-\rangle_B \pm |-\rangle_A|+\rangle_B). \tag{4.26}
$$

The state with $J_z = 0$ and even parity, then, expressed in terms of the linear polarization states, is

$$
\begin{aligned}
&-\frac{i}{\sqrt{2}}(|+\rangle_A|-\rangle_B + |-\rangle_A|+\rangle_B)\\
&= \frac{1}{\sqrt{2}}(|xx\rangle_{AB} + |yy\rangle_{AB})_n = |\phi^+\rangle_{AB}.
\end{aligned}
\tag{4.27}
$$

Because of invariance under rotations about $\hat{z}$, the state has this form irrespective of how we orient the x and y axes.

We can use a polarization analyzer to measure the linear polarization of either photon along any axis in the xy plane. Let $|x(\theta)\rangle$ and $|y(\theta)\rangle$ denote

the linear polarization eigenstates along axes rotated by angle θ relative to the canonical x and y axes. We may define an operator (the analog of $\vec{\boldsymbol{\sigma}} \cdot \hat{n}$)

$$
\boldsymbol{\tau}(\theta) = |x(\theta)\rangle\langle x(\theta)| - |y(\theta)\rangle\langle y(\theta)|, \tag{4.28}
$$

which has these polarization states as eigenstates with respective eigenvalues ± 1. Since

$$
|x(\theta)\rangle = \begin{pmatrix} \cos & \theta \\ \sin & \theta \end{pmatrix}, \quad |y(\theta)\rangle = \begin{pmatrix} -\sin & \theta \\ \cos & \theta \end{pmatrix}, \tag{4.29}
$$

in the $|x\rangle, |y\rangle$ basis, we can easily compute the expectation value

$$
{}_{AB}\langle\phi^+|\boldsymbol{\tau}^{(A)}(\theta_1)\boldsymbol{\tau}^{(B)}(\theta_2)|\phi^+\rangle_{AB}. \tag{4.30}
$$

Using rotational invariance:

$$
\begin{aligned}
&= {}_{AB}\langle\phi^+|\boldsymbol{\tau}^{(A)}(0)\boldsymbol{\tau}^{(B)}(\theta_2-\theta_1)|\phi^+\rangle_{AB}\\
&= \frac{1}{2}\,{}_B\langle x|\boldsymbol{\tau}^{(B)}(\theta_2-\theta_1)|x\rangle_B - \frac{1}{2}\,{}_B\langle y|\boldsymbol{\tau}^{(B)}(\theta_2-\theta_1)|y\rangle_B\\
&= \cos^2(\theta_2-\theta_1) - \sin^2(\theta_2-\theta_1) = \cos[2(\theta_2-\theta_1)].
\end{aligned}
\tag{4.31}
$$

(For spin-$\frac{1}{2}$ objects, we would obtain

$$
{}_{AB}\langle\phi^+|(\vec{\boldsymbol{\sigma}}^{(A)}\cdot\hat{n}_1)(\vec{\boldsymbol{\sigma}}^{(B)}\cdot\hat{n}_2) = \hat{n}_1\cdot\hat{n}_2 = \cos(\theta_2-\theta_1); \tag{4.32}
$$

the argument of the cosine is different than in the case of photons, because the half angle $\theta/2$ appears in the formula analogous to eq. (4.29).)

4.1.5 More Bell inequalities

So far, we have considered only one (particularly interesting) case of the Bell inequality. Here we will generalize the result.

Consider a correlated pair of photons, A and B. We may choose to measure the polarization of photon A along either one of two axes, α or α'. The corresponding observables are denoted

$$
\begin{aligned}
\boldsymbol{a} &= \boldsymbol{\tau}^{(A)}(\alpha)\\
\boldsymbol{a}' &= \boldsymbol{\tau}^{(A)}(\alpha').
\end{aligned}
\tag{4.33}
$$

Similarly, we may choose to measure photon B along either axis β or axis β'; the corresponding observables are

$$\begin{aligned} \boldsymbol{b} &= \boldsymbol{\tau}^{(B)}(\beta) \\ \boldsymbol{b} &= \boldsymbol{\tau}^{(B)}(\beta'). \end{aligned} \tag{4.34}$$

We will, to begin with, consider the special case $\alpha' = \beta' \equiv \gamma$.

Now, if we make the local hidden variable hypothesis, what can be infer about the correlations among these observables? We'll assume that the prediction of quantum mechanics is satisfied if we measure $\boldsymbol{a}'$ and $\boldsymbol{b}'$, namely

$$\langle \boldsymbol{a}'\boldsymbol{b}' \rangle = \langle \boldsymbol{\tau}^{(B)}(\gamma)\boldsymbol{\tau}^{(B)}(\gamma) \rangle = 1; \tag{4.35}$$

when we measure both photons along the same axes, the outcomes *always* agree. Therefore, these two observables have exactly the *same* functional dependence on the hidden variables – they are really the *same* observable, with we will denote $\boldsymbol{c}$.

Now, let $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ be *any* three observables with the properties

$$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} = \pm 1; \tag{4.36}$$

i.e., they are functions of the hidden variables that take only the two values ± 1. These functions satisfy the identity

$$\boldsymbol{a}(\boldsymbol{b} - \boldsymbol{c}) = \boldsymbol{ab}(1 - \boldsymbol{bc}). \tag{4.37}$$

(We can easily verify the identity by considering the cases $\boldsymbol{b} - \boldsymbol{c} = 0, 2, -2$.) Now we take expectation values by integrating over the hidden variables, weighted by a nonnegative probability distribution:

$$\langle \boldsymbol{ab} \rangle - \langle \boldsymbol{ac} \rangle = \langle \boldsymbol{ab}(1 - \boldsymbol{bc}) \rangle. \tag{4.38}$$

Furthermore, since $\boldsymbol{ab} = \pm 1$, and $1 - \boldsymbol{bc}$ is nonnegative, we have

$$\begin{aligned} &|\langle \boldsymbol{ab}(1 - \boldsymbol{bc}) \rangle| \\ &\leq |\langle 1 - \boldsymbol{bc} \rangle| = 1 - \langle \boldsymbol{bc} \rangle. \end{aligned} \tag{4.39}$$

We conclude that

$$|\langle \boldsymbol{ab} \rangle - \langle \boldsymbol{ac} \rangle| \leq 1 - \langle \boldsymbol{bc} \rangle. \tag{4.40}$$

This is the Bell inequality.

To make contact with our earlier discussion, consider a pair of spin-$\frac{1}{2}$ objects in the state $|\phi^+\rangle$, where α, β, γ are separated by successive 60^o angles. Then quantum mechanics predicts

$$\begin{aligned} \langle \boldsymbol{ab} \rangle &= \frac{1}{2} \\ \langle \boldsymbol{bc} \rangle &= \frac{1}{2} \\ \langle \boldsymbol{ac} \rangle &= -\frac{1}{2}, \end{aligned} \tag{4.41}$$

which violates the Bell inequality:

$$1 = \frac{1}{2} + \frac{1}{2} \not\leq 1 - \frac{1}{2} = \frac{1}{2}. \tag{4.42}$$

For photons, to obtain the same violation, we halve the angles, so α, β, γ are separated by 30^o angles.

Return now to the more general case $\alpha' \neq \beta'$. We readily see that $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{b}, \boldsymbol{b}' = \pm 1$ implies that

$$(\boldsymbol{a} + \boldsymbol{a}')\boldsymbol{b} - (\boldsymbol{a} - \boldsymbol{a}')\boldsymbol{b}' = \pm 2, \tag{4.43}$$

(by considering the two cases $\boldsymbol{a} + \boldsymbol{a}' = 0$ and $\boldsymbol{a} - \boldsymbol{a}' = 0$), or

$$\langle \boldsymbol{ab} \rangle + \langle \boldsymbol{a}'\boldsymbol{b} \rangle + \langle \boldsymbol{a}'\boldsymbol{b}' \rangle - \langle \boldsymbol{ab}' \rangle = \langle \boldsymbol{\theta} \rangle, \tag{4.44}$$

where $\boldsymbol{\theta} = \pm 2$. Evidently

$$|\langle \boldsymbol{\theta} \rangle| \leq 2, \tag{4.45}$$

so that

$$|\langle \boldsymbol{ab} \rangle + \langle \boldsymbol{a}'\boldsymbol{b} \rangle + \langle \boldsymbol{a}'\boldsymbol{b}' \rangle - \langle \boldsymbol{ab}' \rangle| \leq 2. \tag{4.46}$$

This result is called the CHSH (Clauser-Horne-Shimony-Holt) inequality. To see that quantum mechanics violates it, consider the case for photons where $\alpha, \beta, \alpha', \beta'$ are separated by successive $22.5°$ angles, so that the quantum-mechanical predictions are

$$\begin{aligned} \langle \boldsymbol{ab} \rangle = \langle \boldsymbol{a}'\boldsymbol{b} \rangle = \langle \boldsymbol{a}'\boldsymbol{b}' \rangle &= \cos\frac{\pi}{4} = \frac{1}{\sqrt{2}}, \\ \langle \boldsymbol{ab}' \rangle &= \cos\frac{3\pi}{4} = -\frac{1}{\sqrt{2}}, \end{aligned} \tag{4.47}$$

while

$$2\sqrt{2} \not\leq 2. \tag{4.48}$$

4.1.6 Maximal violation

We can see that the case just considered ($\alpha, \beta, \alpha', \beta'$ separated by successive 22.5^o angles) provides the largest possible quantum mechanical violation of the CHSH inequality. In quantum theory, suppose that $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{b}, \boldsymbol{b}'$ are observables that satisfy

$$\boldsymbol{a}^2 = \boldsymbol{a}'^2 = \boldsymbol{b}^2 = \boldsymbol{b}'^2 = \boldsymbol{1}, \tag{4.49}$$

and

$$0 = [\boldsymbol{a}, \boldsymbol{b}] = [\boldsymbol{a}, \boldsymbol{b}'] = [\boldsymbol{a}', \boldsymbol{b}] = [\boldsymbol{a}', \boldsymbol{b}']. \tag{4.50}$$

Let

$$\boldsymbol{C} = \boldsymbol{ab} + \boldsymbol{a}'\boldsymbol{b} + \boldsymbol{a}'\boldsymbol{b}' - \boldsymbol{ab}'. \tag{4.51}$$

Then

$$\boldsymbol{C}^2 = 4 + \boldsymbol{aba}'\boldsymbol{b}' - \boldsymbol{a}'\boldsymbol{bab}' + \boldsymbol{a}'\boldsymbol{b}'\boldsymbol{ab} - \boldsymbol{ab}'\boldsymbol{a}'\boldsymbol{b}. \tag{4.52}$$

(You can check that the other terms cancel)

$$= 4 + [\boldsymbol{a}, \boldsymbol{a}'][\boldsymbol{b}, \boldsymbol{b}']. \tag{4.53}$$

The *sup norm* $\parallel \boldsymbol{M} \parallel$ of a bounded operator $\boldsymbol{M}$ is defined by

$$\parallel \boldsymbol{M} \parallel = \begin{matrix} sup \\ |\psi\rangle \end{matrix} \left(\frac{\parallel \boldsymbol{M}|\psi\rangle \parallel}{\parallel |\psi\rangle \parallel} \right); \tag{4.54}$$

it is easy to verify that the sup norm has the properties

$$\begin{aligned} &\parallel \boldsymbol{MN} \parallel \leq \parallel \boldsymbol{M} \parallel \parallel \boldsymbol{N} \parallel, \\ &\parallel \boldsymbol{M} + \boldsymbol{N} \parallel \leq \parallel \boldsymbol{M} \parallel + \parallel \boldsymbol{N} \parallel, \end{aligned} \tag{4.55}$$

and therefore

$$\parallel [\boldsymbol{M}, \boldsymbol{N}] \parallel \leq \parallel \boldsymbol{MN} \parallel + \parallel \boldsymbol{NM} \parallel \leq 2 \parallel \boldsymbol{M} \parallel \parallel \boldsymbol{N} \parallel . \tag{4.56}$$

We conclude that

$$\parallel \boldsymbol{C}^2 \parallel \leq 4 + 4 \parallel \boldsymbol{a} \parallel \cdot \parallel \boldsymbol{a}' \parallel \cdot \parallel \boldsymbol{b} \parallel \cdot \parallel \boldsymbol{b}' \parallel = 8, \tag{4.57}$$

or

$$\parallel \boldsymbol{C} \parallel \leq 2\sqrt{2} \tag{4.58}$$

(Cirel'son's inequality). Thus, the expectation value of $\boldsymbol{C}$ cannot exceed $2\sqrt{2}$, precisely the value that we found to be attained in the case where $\alpha, \beta, \alpha', \beta'$ are separated by successive 22.5^o angles. The violation of the CHSH inequality that we found is the largest violation allowed by quantum theory.

4.1.7 The Aspect experiment

The CHSH inequality was convincingly tested for the first time by Aspect and collaborators in 1982. Two entangled photons were produced in the decay of an excited calcium atom, and each photon was directed by a switch to one of two polarization analyzers, chosen pseudo-randomly. The photons were detected about 12m apart, corresponding to a light travel time of about 40 ns. This time was considerably longer than either the cycle time of the switch, or the difference in the times of arrival of the two photons. Therefore the "decision" about which observable to measure was made after the photons were already in flight, and the events that selected the axes for the measurement of photons A and B were spacelike separated. The results were consistent with the quantum predictions, and violated the CHSH inequality by five standard deviations. Since Aspect, many other experiments have confirmed this finding.

4.1.8 Nonmaximal entanglement

So far, we have considered the Bell inequality violations predicted by quantum theory for a maximally entangled state such as $|\phi^+\rangle$. But what about more general states such as

$$|\phi\rangle = \alpha|00\rangle + \beta|11\rangle? \tag{4.59}$$

(Any pure state of two qubits can be expressed this way in the Schmidt basis; by adopting suitable phase conventions, we may assume that α and β are real and nonnegative.)

Consider first the extreme case of separable pure states, for which

$$\langle \boldsymbol{ab} \rangle = \langle \boldsymbol{a} \rangle \langle \boldsymbol{b} \rangle. \tag{4.60}$$

In this case, it is clear that no Bell inequality violation can occur, because we have already seen that a (local) hidden variable theory *does* exist that correctly reproduces the predictions of quantum theory for a pure state of a single qubit. Returning to the spin-$\frac{1}{2}$ notation, suppose that we measure the spin of each particle along an axis $\hat{n} = (\sin\theta, 0, \cos\theta)$ in the xz plane. Then

$$\boldsymbol{a} = (\boldsymbol{\sigma}^{(A)} \cdot \hat{n}_1) = \begin{pmatrix} \cos\theta_1 & \sin\theta_1 \\ \sin\theta_1 & -\cos\theta_1 \end{pmatrix}^{(A)},$$
$$\boldsymbol{b} = (\boldsymbol{\sigma}^{(B)} \cdot \hat{n}_2) = \begin{pmatrix} \cos\theta_2 & \sin\theta_2 \\ \sin\theta_2 & -\cos\theta_2 \end{pmatrix}^{(B)}, \tag{4.61}$$

so that quantum mechanics predicts

$$\begin{aligned} \langle \boldsymbol{ab} \rangle &= \langle \phi | \boldsymbol{ab} | \phi \rangle \\ &= \cos\theta_1 \cos\theta_2 + 2\alpha\beta \sin\theta_1 \sin\theta_2 \end{aligned} \tag{4.62}$$

(and we recover $\cos(\theta_1 - \theta_2)$ in the maximally entangled case $\alpha = \beta = 1/\sqrt{2}$). Now let us consider, for simplicity, the (nonoptimal!) special case

$$\theta_A = 0, \quad \theta'_A = \frac{\pi}{2}, \quad \theta'_B = -\theta_B, \tag{4.63}$$

so that the quantum predictions are:

$$\begin{aligned} \langle \boldsymbol{ab} \rangle &= \cos\theta_B = \langle \boldsymbol{ab}' \rangle \\ \langle \boldsymbol{a}'\boldsymbol{b} \rangle &= 2\alpha\beta \sin\theta_B = -\langle \boldsymbol{a}'\boldsymbol{b}' \rangle \end{aligned} \tag{4.64}$$

Plugging into the CHSH inequality, we obtain

$$|\cos\theta_B - 2\alpha\beta \sin\theta_B| \le 1, \tag{4.65}$$

and we easily see that violations occur for θ_B close to 0 or π. Expanding to linear order in θ_B, the left hand side is

$$\simeq 1 - 2\alpha\beta\theta_B, \tag{4.66}$$

which surely exceeds 1 for θ_B negative and small.

We have shown, then, that *any* entangled pure state of two qubits violates some Bell inequality. It is not hard to generalize the argument to an arbitrary bipartite pure state. For bipartite pure states, then, "entangled" is equivalent to "Bell-inequality violating." For bipartite mixed states, however, we will see shortly that the situation is more subtle.

4.2 Uses of Entanglement

After Bell's work, quantum entanglement became a subject of intensive study among those interested in the foundations of quantum theory. But more recently (starting less than ten years ago), entanglement has come to be viewed not just as a tool for exposing the weirdness of quantum mechanics, but as a potentially valuable *resource*. By exploiting entangled quantum states, we can perform tasks that are otherwise difficult or impossible.

4.2.1 Dense coding

Our first example is an application of entanglement to communication. Alice wants to send messages to Bob. She might send classical bits (like dots and dashes in Morse code), but let's suppose that Alice and Bob are linked by a *quantum* channel. For example, Alice can prepare qubits (like photons) in any polarization state she pleases, and send them to Bob, who measures the polarization along the axis of his choice. Is there any advantage to sending qubits instead of classical bits?

In principle, if their quantum channel has perfect fidelity, and Alice and Bob perform the preparation and measurement with perfect efficiency, then they are no *worse* off using qubits instead of classical bits. Alice can prepare, say, either $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$, and Bob can measure along $\hat{z}$ to infer the choice she made. This way, Alice can send one classical bit with each qubit. But in fact, that is the best she can do. Sending one qubit at a time, no matter how she prepares it and no matter how Bob measures it, no more than one classical bit can be carried by each qubit. (This statement is a special case of a bound proved by Kholevo (1973) on the classical information capacity of a quantum channel.)

But now, let's change the rules a bit – let's suppose that Alice and Bob share an entangled pair of qubits in the state $|\phi^+\rangle_{AB}$. The pair was prepared last year; one qubit was shipped to Alice and the other to Bob, anticipating that the shared entanglement would come in handy someday. Now, use of the quantum channel is very expensive, so Alice can afford to send only one qubit to Bob. Yet it is of the utmost importance for Alice to send Bob *two* classical bits of information.

Fortunately, Alice remembers about the entangled state $|\phi^+\rangle_{AB}$ that she shares with Bob, and she carries out a protocol that she and Bob had arranged for just such an emergency. On her member of the entangled pair, she can perform one of four possible unitary transformations:

1) $\mathbf{1}$ (she does nothing),

2) $\boldsymbol{\sigma}_1$ (180^o rotation about $\hat{x}$-axis),

3) $\boldsymbol{\sigma}_2$ (180^o rotation about $\hat{y}$-axis),

4) $\boldsymbol{\sigma}_3$ (180^o rotation about $\hat{z}$-axis).

As we have seen, by doing so, she transforms $|\phi^+\rangle_{AB}$ to one of 4 mutually orthogonal states:

1) $|\phi^+\rangle_{AB}$,

2) $|\psi^+\rangle_{AB}$,

3) $|\psi^-\rangle_{AB}$,

4) $|\phi^-\rangle_{AB}$.

Now, she sends her qubit to Bob, who receives it and then performs an orthogonal collective measurement on the pair that projects onto the maximally entangled basis. The measurement outcome unambiguously distinguishes the four possible actions that Alice could have performed. Therefore the single qubit sent from Alice to Bob has successfully carried 2 bits of classical information! Hence this procedure is called "dense coding."

A nice feature of this protocol is that, if the message is highly confidential, Alice need not worry that an eavesdropper will intercept the transmitted qubit and decipher her message. The transmitted qubit has density matrix $\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1}_A$, and so carries no information at all. All the information is in the correlations between qubits A and B, and this information is inaccessible unless the adversary is able to obtain both members of the entangled pair. (Of course, the adversary *can* "jam" the channel, preventing the information but reaching Bob.)

From one point of view, Alice and Bob really *did* need to use the channel twice to exchange two bits of information – a qubit had to be transmitted for them to establish their entangled pair in the first place. (In effect, Alice has merely sent to Bob two qubits chosen to be in one of the four mutually orthogonal entangled states.) But the first transmission could have taken place a long time ago. The point is that when an emergency arose and two bits had to be sent immediately while only one use of the channel was possible, Alice and Bob could exploit the pre-existing entanglement to communicate more efficiently. They used entanglement as a resource.

4.2.2 EPR Quantum Key Distribution

Everyone has secrets, including Alice and Bob. Alice needs to send a highly private message to Bob, but Alice and Bob have a very nosy friend, Eve, who they know will try to listen in. Can they communicate with assurance that Eve is unable to eavesdrop?

Obviously, they should use some kind of code. Trouble is, aside from being very nosy, Eve is also very smart. Alice and Bob are not confident that they are clever enough to devise a code that Eve cannot break.

Except there is one coding scheme that is surely unbreakable. If Alice and Bob share a *private key*, a string of random bits known only to them, then Alice can convert her message to ASCII (a string of bits no longer than the key) *add* each bit of her message (module 2) to the corresponding bit of the key, and send the result to Bob. Receiving this string, Bob adds the key to it to extract Alice's message.

This scheme is secure because even if Eve should intercept the transmission, she will not learn anything because the transmitted string itself carries no information – the message is encoded in a correlation between the transmitted string and the *key* (which Eve doesn't know).

There is still a problem, though, because Alice and Bob need to establish a shared random key, and they must ensure that Eve can't know the key. They could meet to exchange the key, but that might be impractical. They could entrust a third party to transport the key, but what if the intermediary is secretly in cahoots with Eve? They could use "public key" distribution protocols, but these are not guaranteed to be secure.

Can Alice and Bob exploit *quantum* information (and specifically entanglement) to solve the key exchange problem? They can! This observation is the basis of what is sometimes called "quantum cryptography." But since quantum mechanics is really used for key exchange rather than for encoding, it is more properly called "quantum key distribution."

Let's suppose that Alice and Bob share a supply of entangled pairs, each prepared in the state $|\psi^-\rangle$. To establish a shared private key, they may carry out this protocol.

For each qubit in her/his possession, Alice and Bob decide to measure either $\boldsymbol{\sigma}_1$ or $\boldsymbol{\sigma}_3$. The decision is pseudo-random, each choice occuring with probability 1/2. Then, after the measurements are performed, both Alice and Bob publicly announce what observables they measured, but do not reveal the outcomes they obtained. For those cases (about half) in which they measured their qubits along different axes, their results are discarded (as Alice and Bob obtained uncorrelated outcomes). For those cases in which they measured along the same axis, their results, though random, are *perfectly (anti-)correlated.* Hence, they have established a shared random key.

But, is this protocol really invulnerable to a sneaky attack by Eve? In particular, Eve might have clandestinely tampered with the pairs at some time and in the past. Then the pairs that Alice and Bob possess might be (unbeknownst to Alice and Bob) not perfect $|\psi^-\rangle$'s, but rather pairs that are entangled with qubits in Eve's possession. Eve can then wait until Alice and Bob make their public announcements, and proceed to measure her qubits in a manner designed to acquire maximal information about the results that Alice and Bob obtained. Alice and Bob must protect themselves against this type of attack.

If Eve has indeed tampered with Alice's and Bob's pairs, then the most general possible state for an AB pair and a set of E qubits has the form

$$\begin{aligned} |\Upsilon\rangle_{ABE} = |00\rangle_{AB}|e_{00}\rangle_E + |01\rangle_{AB}|e_{01}\rangle_E \\ + |10\rangle_{AB}|e_{10}\rangle_E + |11\rangle_{AB}|e_{11}\rangle_E. \end{aligned} \tag{4.67}$$

But now recall that the defining property or $|\psi^-\rangle$ is that it is an eigenstate with eigenvalue -1 of both $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)}$ and $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$. Suppose that A and B are able to verify that the pairs in their possession have this property. To satisfy $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)} = -1$, we must have

$$|\Upsilon\rangle_{AB} = |01\rangle_{AB}|e_{01}\rangle_E + |10\rangle_{AB}|e_{10}\rangle_E, \tag{4.68}$$

and to also satisfy $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)} = -1$, we must have

$$|\Upsilon\rangle_{ABE} = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)|e\rangle_E = |\psi^-\rangle|e\rangle. \tag{4.69}$$

We see that it is possible for the AB pairs to be eigenstates of $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)}$ and $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$ only if they are completely unentangled with Eve's qubits.

Therefore, Eve will not be able to learn anything about Alice's and Bob's measurement results by measuring her qubits. The random key is secure.

To verify the properties $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)} = -1 = \boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$, Alice and Bob can sacrifice a portion of their shared key, and publicly compare their measurement outcomes. They should find that their results are indeed perfectly correlated. If so they will have high statistical confidence that Eve is unable to intercept the key. If not, they have detected Eve's nefarious activity. They may then discard the key, and make a fresh attempt to establish a secure key.

As I have just presented it, the quantum key distribution protocol seems to require entangled pairs shared by Alice and Bob, but this is not really so. We might imagine that Alice prepares the $|\psi^-\rangle$ pairs herself, and then measures one qubit in each pair before sending the other to Bob. This is completely equivalent to a scheme in which Alice prepares one of the four states

$$|\uparrow_z\rangle, |\downarrow_z\rangle, |\uparrow_x\rangle, |\downarrow_x\rangle, \tag{4.70}$$

(chosen at random, each occuring with probability 1/4) and sends the qubit to Bob. Bob's measurement and the verification are then carried out as before. This scheme (known as BB84 in quantum key distribution jargon) is just as secure as the entanglement-based scheme.[3]

Another intriguing variation is called the "time-reversed EPR" scheme. Here both Alice and Bob prepare one of the four states in eq. (4.70), and they both send their qubits to Charlie. Then Charlie performs a Bell measurement on the pair, orthogonally projecting out one of $|\phi^\pm\rangle|\psi^\pm\rangle$, and he publicly announces the result. Since all four of these states are simultaneous eigenstates of $\boldsymbol{\sigma}_1^{(A)}\boldsymbol{\sigma}_1^{(B)}$ and $\boldsymbol{\sigma}_3^{(A)}\boldsymbol{\sigma}_3^{(B)}$, when Alice and Bob both prepared their spins along the same axis (as they do about half the time) they share a single bit.[4] Of course, Charlie could be allied with Eve, but Alice and Bob can verify that Charlie has acquired no information as before, by comparing a portion of their key. This scheme has the advantage that Charlie could

[3] Except that in the EPR scheme, Alice and Bob can wait until just before they need to talk to generate the key, thus reducing the risk that Eve might at some point burglarize Alice's safe to learn what states Alice prepared (and so infer the key).

[4] Until Charlie does his measurement, the states prepared by Bob and Alice are totally uncorrelated. A definite correlation (or anti-correlation) is established after Charlie performs his measurement.

operate a central switching station by storing qubits received from many parties, and then perform his Bell measurement when two of the parties request a secure communication link. A secure key can be established even if the quantum communication line is down temporarily, as long as both parties had the foresight to send their qubits to Charlie on an earlier occasion (when the quantum channel was open.)

So far, we have made the unrealistic assumption that the quantum communication channel is perfect, but of course in the real world errors will occur. Therefore even if Eve has been up to no mischief, Alice and Bob will sometimes find that their verification test will fail. But how are they to distinguish errors due to imperfections of the channel from errors that occur because Eve has been eavesdropping?

To address this problem, Alice and Bob must enhance their protocol in two ways. First they must implement (classical) error correction to reduce the effective error rate. For example, to establish each bit of their shared key they could actually exchange a block of three random bits. If the three bits are not all the same, Alice can inform Bob which of the three is different than the other two; Bob can flip that bit in his block, and *then* use majority voting to determine a bit value for the block. This way, Alice and Bob share the same key bit even if an error occured for one bit in the block of three.

However, error correction alone does not suffice to ensure that Eve has acquired negligible information about the key – error correction must be supplemented by (classical) privacy amplification. For example, after performing error correction so that they are confident that they share the same key, Alice and Bob might extract a bit of "superkey" as the *parity* of n key bits. To know *anything* about the parity of n bits, Eve would need to know *something* about each of the bits. Therefore, the parity bit is considerably more secure, on the average, than each of the individual key bits.

If the error rate of the channel is low enough, one can hope to show that quantum key distribution, supplemented by error correction and privacy amplification, is invulnerable to any attack that Eve might muster (in the sense that the information acquired by Eve can be guaranteed to be arbitrarily small). Whether this has been established is, at the moment, a matter of controversy.

4.2.3 No cloning

The security of quantum key distribution is based on an essential difference between quantum information and classical information. It is not possible to acquire information that *distinguishes* between nonorthogonal quantum states without *disturbing* the states.

For example, in the BB84 protocol, Alice sends to Bob any one of the four states $|\uparrow_z\rangle |\downarrow_z\rangle |\uparrow_x\rangle |\downarrow_x\rangle$, and Alice and Bob are able to verify that none of their states are perturbed by Eve's attempt at eavesdropping. Suppose, more generally, that $|\varphi\rangle$ and $|\psi\rangle$ are two nonorthogonal states in $\mathcal{H}$ ($\langle\psi|\varphi\rangle \neq 0$) and that a unitary transformation U is applied to $\mathcal{H}\otimes\mathcal{H}_E$ (where $\mathcal{H}_E$ is a Hilbert space accessible to Eve) that leaves both $|\psi\rangle$ and $|\varphi\rangle$ undisturbed. Then

$$\begin{aligned} U: \quad & |\psi\rangle \otimes |0\rangle_E \to |\psi\rangle \otimes |e\rangle_E, \\ & |\varphi\rangle \otimes |0\rangle_E \to |\varphi\rangle \otimes |f\rangle_E, \end{aligned} \tag{4.71}$$

and unitarity implies that

$$\begin{aligned} \langle\psi|\phi\rangle &= ({}_E\langle 0| \otimes \langle\psi|)(|\varphi\rangle \otimes |0\rangle_E) \\ &= ({}_E\langle e| \otimes \langle\psi|)(|\varphi\rangle \otimes |f\rangle_E) \\ &= \langle\psi|\varphi\rangle_E\langle e|f\rangle_E. \end{aligned} \tag{4.72}$$

Hence, for $\langle\psi|\varphi\rangle \neq 0$, we have ${}_E\langle e|f\rangle_E = 1$, and therefore since the states are normalized, $|e\rangle = |f\rangle$. This means that no measurement in $\mathcal{H}_E$ can reveal any information that distinguishes $|\psi\rangle$ from $|\varphi\rangle$. In the BB84 case this argument shows that the state in $\mathcal{H}_E$ will be the same irrespective of which of the four states $|\uparrow_z\rangle, |\downarrow_z\rangle, |\uparrow_x\rangle, |\downarrow_x\rangle$ is sent by Alice, and therefore Eve learns nothing about the key shared by Alice and Bob. On the other hand, if Alice is sending to Bob one of the two orthogonal states $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$, there is nothing to prevent Eve from acquiring a copy of the information (as with classical bits).

We have noted earlier that if we have many identical copies of a qubit, then it is possible to measure the mean value of noncommuting observables like $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$, and $\boldsymbol{\sigma}_3$ to completely determine the density matrix of the qubit. Inherent in the conclusion that nonorthogonal state cannot be distinguished without disturbing them, then, is the implicit provision that it is not possible to make a perfect copy of a qubit. (If we could, we would make as many copies as we need to find $\langle\boldsymbol{\sigma}_1\rangle, \langle\boldsymbol{\sigma}_2\rangle$, and $\langle\boldsymbol{\sigma}_3\rangle$ to any specified accuracy.) Let's now make this point explicit: there is no such thing as a perfect quantum Xerox machine.

Orthogonal quantum states (like classical information) *can* be reliably copied. For example, the unitary transformation that acts as

$$\begin{aligned} U: \quad & |0\rangle_A|0\rangle_B \rightarrow |0\rangle_A|0\rangle_B \\ & |1\rangle_A|0\rangle_B \rightarrow |1\rangle_A|1\rangle_B, \end{aligned} \tag{4.73}$$

copies the first qubit onto the second if the first qubit is in one of the states $|0\rangle_A$ or $|1\rangle_A$. But if instead the first qubit is in the state $|\psi\rangle = a|0\rangle_A + b|1\rangle_A$, then

$$\begin{aligned} U: \quad & (a|0\rangle_A + b|1\rangle_A)|0\rangle_B \\ & \rightarrow a|0\rangle_A|0\rangle_B + b|1\rangle_A|1\rangle_B. \end{aligned} \tag{4.74}$$

Thus is *not* the state $|\psi\rangle\otimes|\psi\rangle$ (a tensor product of the original and the copy); rather it is something very different – an entangled state of the two qubits.

To consider the most general possible quantum Xerox machine, we allow the full Hilbert space to be larger than the tensor product of the space of the original and the space of the copy. Then the most general "copying" unitary transformation acts as

$$\begin{aligned} U: \quad & |\psi\rangle_A|0\rangle_B|0\rangle_E \rightarrow |\psi\rangle_A|\psi\rangle_B|e\rangle_E \\ & |\varphi\rangle_A|0\rangle_B|0\rangle_E \rightarrow |\varphi\rangle_A|\varphi\rangle_B|f\rangle_E. \end{aligned} \tag{4.75}$$

Unitarity then implies that

$$_A\langle\psi|\varphi\rangle_A = \ _A\langle\psi|\varphi\rangle_A \ _B\langle\psi|\varphi\rangle_B \ _E\langle e|f\rangle_E; \tag{4.76}$$

therefore, if $\langle\psi|\varphi\rangle \neq 0$, then

$$1 = \langle\psi|\varphi\rangle \ _E\langle e|f\rangle_E. \tag{4.77}$$

Since the states are normalized, we conclude that

$$|\langle\psi|\varphi\rangle| = 1, \tag{4.78}$$

so that $|\psi\rangle$ and $|\varphi\rangle$ actually represent the same ray. No unitary machine can make a copy of both $|\varphi\rangle$ and $|\psi\rangle$ if $|\varphi\rangle$ and $|\psi\rangle$ are *distinct*, *nonorthogonal* states. This result is called the no-cloning theorem.

4.2.4 Quantum teleportation

In dense coding, we saw a case where quantum information could be exploited to enhance the transmission of classical information. Now let's address a closely related issue: Can we use classical information to realize transmission of quantum information?

Alice has a qubit, but she doesn't know it's state. Bob needs this qubit desperately. But that darn quantum channel is down again! Alice can send only *classical* information to Bob.

She could try measuring $\vec{\boldsymbol{\sigma}}\cdot\hat{n}$, projecting her qubit to either $|\uparrow_{\hat{n}}\rangle$ or $|\downarrow_{\hat{n}}\rangle$. She could send the measurement outcome to Bob who could then proceed to prepare the state Alice found. But you showed in a homework exercise that Bob's qubit will not be a perfect copy of Alice's; on the average we'll have

$$F = |_B\langle\,\cdot\,|\psi\rangle_A|^2 = \frac{2}{3}, \tag{4.79}$$

Thus is a better fidelity than could have been achieved ($F = \frac{1}{2}$) if Bob had merely chosen a state at random, but it is not nearly as good as the fidelity that Bob requires.

But then Alice and Bob recall that they share some entangled pairs; why not use the entanglement as a *resource*? They carry out this protocol: Alice unites the unknown qubit $|\psi\rangle_C$ she wants to send to Bob with her member of a $|\phi^+\rangle_{AB}$ pair that she shares with Bob. On these two qubits she performs Bell measurement, projecting onto one of the four states $|\phi^\pm\rangle_{CA}, |\psi^\pm\rangle_{CA}$. She sends her measurement outcome (two bits of classical information) to Bob over the classical channel. Receiving this information, Bob performs one of four operations on his qubit $|\cdot\rangle_B$:

$$\begin{aligned} |\phi^+\rangle_{CA} &\rightarrow \mathbf{1}_B \\ |\psi^+\rangle_{CA} &\rightarrow \boldsymbol{\sigma}_1^{(B)} \\ |\psi^-\rangle_{CA} &\rightarrow \boldsymbol{\sigma}_2^{(B)} \\ |\phi^-\rangle_{CA} &\rightarrow \boldsymbol{\sigma}_3^{(B)}. \end{aligned} \tag{4.80}$$

This action transforms his qubit (his member of the $|\phi^+\rangle_{AB}$ pair that he initially shared with Alice) into a perfect copy of $|\psi\rangle_C$! This magic trick is called *quantum teleportation*.

It is a curious procedure. Initially, Bob's qubit $|\cdot\rangle_B$ is completely unentangled with the unknown qubit $|\psi\rangle_C$, but Alice's Bell measurement establishes a correlation between A and C. The measurement outcome is in fact completely random, as you'll see in a moment, so Alice (and Bob) actually acquire no information at all about $|\psi\rangle$ by making this measurement.

How then does the quantum state manage to travel from Alice to Bob? It is a bit puzzling. On the one hand, we can hardly say that the two classical bits that were transmitted carried this information – the bits were random. So we are tempted to say that the shared entangled pair made the teleportation possible. But remember that the entangled pair was actually prepared last year, long before Alice ever dreamed that she would be sending the qubit to Bob ...

We should also note that the teleportation procedure is fully consistent with the no-cloning theorem. True, a copy of the state $|\psi\rangle$ appeared in Bob's hands. But the original $|\psi\rangle_C$ had to be destroyed by Alice's measurement before the copy could be created.

How does it work? We merely note that for $|\psi\rangle = a|0\rangle + b|1\rangle$, we may write

$$\begin{aligned}
|\psi\rangle_C|\phi^+\rangle_{AB} &= (a|0\rangle_C + b|1\rangle_C)\frac{1}{\sqrt{2}}(|00\rangle_{AB} + |11\rangle_{AB})\\
&= \frac{1}{\sqrt{2}}(a|000\rangle_{CAB} + a|011\rangle_{CAB} + b|100\rangle_{CAB} + b|111\rangle_{CAB})\\
&= \frac{1}{2}a(|\phi^+\rangle_{CA} + |\phi^-\rangle_{CA})|0\rangle_B + \frac{1}{2}a(|\psi^+\rangle_{CA} + |\psi^-\rangle_{CA})|1\rangle_B\\
&+ \frac{1}{2}b(|\psi^+\rangle_{CA} - |\psi^-\rangle_{CA})|0\rangle_B + \frac{1}{2}b(|\phi^+\rangle_{CA} - |\phi^-\rangle_{CA})|1\rangle_B\\
&= \frac{1}{2}|\phi^+\rangle_{CA}(a|0\rangle_B + b|1\rangle_B)\\
&+ \frac{1}{2}|\psi^+\rangle_{CA}(a|1\rangle_B + b|0\rangle_B)\\
&+ \frac{1}{2}|\psi^-\rangle_{CA}(a|1\rangle_B - b|0\rangle_B)\\
&+ \frac{1}{2}|\phi^-\rangle_{CA}(a|0\rangle_B - b|1\rangle_B)\\
&= \frac{1}{2}|\phi^+\rangle_{CA}|\psi\rangle_B + \frac{1}{2}|\psi^+\rangle_{CA}\boldsymbol{\sigma}_1|\psi\rangle_B\\
&+ \frac{1}{2}|\psi^-\rangle_{CA}(-i\boldsymbol{\sigma}_2)|\psi\rangle_B + \frac{1}{2}|\phi^-\rangle_{CA}\boldsymbol{\sigma}_3|\psi\rangle_B. \qquad (4.81)
\end{aligned}$$

Thus we see that when we perform the Bell measurement on qubits C and

A, all four outcomes are equally likely, and that the actions prescribed in Eq. (4.80) will restore Bob's qubit to the initial state $|\psi\rangle$.

Chapter 5

Quantum Information Theory

Quantum information theory is a rich subject that could easily have occupied us all term. But because we are short of time (I'm anxious to move on to quantum computation), I won't be able to cover this subject in as much depth as I would have liked. We will settle for a brisk introduction to some of the main ideas and results. The lectures will perhaps be sketchier than in the first term, with more hand waving and more details to be filled in through homework exercises. Perhaps this chapter should have been called "quantum information theory for the impatient."

Quantum information theory deals with four main topics:

(1) Transmission of classical information over quantum channels (which we will discuss).

(2) The tradeoff between acquisition of information about a quantum state and disturbance of the state (briefly discussed in Chapter 4 in connection with quantum cryptography, but given short shrift here).

(3) Quantifying quantum entanglement (which we will touch on briefly).

(4) Transmission of quantum information over quantum channels. (We will discuss the case of a noiseless channel, but we will postpone discussion of the noisy channel until later, when we come to quantum error-correcting codes.)

These topics are united by a common recurring theme: the interpretation and applications of the Von Neumann entropy.

5.1 Shannon for Dummies

Before we can understand Von Neumann entropy and its relevance to quantum information, we must discuss Shannon entropy and its relevance to classical information.

Claude Shannon established the two core results of classical information theory in his landmark 1948 paper. The two central problems that he solved were:

(1) How much can a message be *compressed*; *i.e.*, how redundant is the information? (The "noiseless coding theorem.").

(2) At what *rate* can we communicate reliably over a noisy channel; *i.e.*, how much redundancy must be incorporated into a message to protect against errors? (The "noisy channel coding theorem.")

Both questions concern *redundancy* – how *unexpected* is the next letter of the message, on the average. One of Shannon's key insights was that *entropy* provides a suitable way to quantify redundancy.

I call this section "Shannon for Dummies" because I will try to explain Shannon's ideas quickly, with a minimum of ε's and δ's. That way, I can compress classical information theory to about 11 pages.

5.1.1 Shannon entropy and data compression

A message is a string of letters chosen from an alphabet of k letters

$$\{a_1, a_2, \ldots, a_k\}. \tag{5.1}$$

Let us suppose that the letters in the message are statistically independent, and that each letter a_x occurs with an *a priori* probability $p(a_x)$, where $\sum_{x=1}^{k} p(a_x) = 1$. For example, the simplest case is a binary alphabet, where 0 occurs with probability $1-p$ and 1 with probability p (where $0 \le p \le 1$).

Now consider long messages with n letters, $n \gg 1$. We ask: is it possible to compress the message to a shorter string of letters that conveys essentially the same information?

For n very large, the law of large numbers tells us that typical strings will contain (in the binary case) about $n(1-p)$ 0's and about np 1's. The number of distinct strings of this form is of order the binomial coefficient $\binom{n}{np}$, and from the Stirling approximation $logn! = n\log n - n + 0(\log n)$ we obtain

$$\log\binom{n}{np} = \log\left(\frac{n!}{(np)![n(1-p)]!}\right) \cong$$

$$n\log n - n - [np\log np - np + n(1-p)\log n(1-p) - n(1-p)]$$

$$= nH(p), \tag{5.2}$$

where

$$H(p) = -p\log p - (1-p)\log(1-p) \tag{5.3}$$

is the *entropy* function. Hence, the number of typical strings is of order $2^{nH(p)}$. (Logs are understood to have base 2 unless otherwise specified.)

To convey essentially all the information carried by a string of n bits, it suffices to choose a block code that assigns a positive integer to each of the typical strings. This block code has about $2^{nH(p)}$ letters (all occurring with equal *a priori* probability), so we may specify any one of the letters using a binary string of length $nH(p)$. Since $0 \le H(p) \le 1$ for $0 \le p \le 1$, and $H(p) = 1$ only for $p = \frac{1}{2}$, the block code shortens the message for any $p \ne \frac{1}{2}$ (whenever 0 and 1 are not equally probable). This is Shannon's result. The key idea is that we do not need a codeword for every sequence of letters, only for the *typical* sequences. The probability that the actual message is atypical becomes negligible asymptotically, *i.e.*, in the limit $n \to \infty$.

This reasoning generalizes easily to the case of k letters, where letter x occurs with probability $p(x)$.[1] In a string of n letters, x typically occurs about $np(x)$ times, and the number of typical strings is of order

$$\frac{n!}{\prod_x (np(x))!} \simeq 2^{-nH(X)}, \tag{5.4}$$

where we have again invoked the Stirling approximation and

$$H(X) = \sum_x -p(x)\log p(x). \tag{5.5}$$

[1]The ensemble in which each of n letters is drawn from the distribution X will be denoted X^n.

is the *Shannon* entropy (or simply entropy) of the ensemble $X = \{x, p(x)\}$. Adopting a block code that assigns integers to the typical sequences, the information in a string of n letters can be compressed to $H(X)$ bits. In this sense a letter x chosen from the ensemble carries, on the average, $H(X)$ bits of information.

It is useful to restate this reasoning in a slightly different language. A particular n-letter message

$$x_1x_2\cdots x_n, \tag{5.6}$$

occurs with *a priori* probability

$$P(x_1\cdots x_n) = p(x_1)p(x_2)\cdots p(x_n) \tag{5.7}$$

$$\log P(x_1\cdots x_n) = \sum_{i=1}^{n}\log p(x_i). \tag{5.8}$$

Applying the central limit theorem to this sum, we conclude that for "most sequences"

$$-\frac{1}{n}\log P(x_1, \cdots, x_n) \sim \langle -\log p(x)\rangle \equiv H(X), \tag{5.9}$$

where the brackets denote the mean value with respect to the probability distribution that governs the random variable x.

Of course, with ε's and δ's we can formulate these statements precisely. For any $\varepsilon, \delta > 0$ and for n sufficiently large, each "typical sequence" has a probability P satisfying

$$H(X) - \delta < -\frac{1}{n}\log P(x_1\cdots x_n) < H(X) + \delta, \tag{5.10}$$

and the total probability of all typical sequences exceeds $1-\varepsilon$. Or, in other words, sequences of letters occurring with a total probability greater than $1-\varepsilon$ ("typical sequences") each have probability P such that

$$2^{-n(H-\delta)} \ge P \ge 2^{-n(H+\delta)}, \tag{5.11}$$

and from eq. (5.11) we may infer upper and lower bounds on the *number* $N(\varepsilon, \delta)$ of typical sequences (since the sum of the probabilities of all typical sequences must lie between $1-\varepsilon$ and 1):

$$2^{n(H+\delta)} \ge N(\varepsilon, \delta) \ge (1-\varepsilon)2^{n(H-\delta)}. \tag{5.12}$$

With a block code of length $n(H+\delta)$ bits we can encode all typical sequences. Then no matter how the atypical sequences are encoded, the probability of error will still be less than ε.

Conversely, if we attempt to compress the message to less than $H-\delta'$ bits per letter, we will be unable to achieve a small error rate as $n \to \infty$, because we will be unable to assign unique codewords to all typical sequences. The probability P_{success} of successfully decoding the message will be bounded by

$$P_{\text{success}} \le 2^{n(H-\delta')}2^{-n(H-\delta)} + \varepsilon' = 2^{-n(\delta'-\delta)} + \varepsilon'; \tag{5.13}$$

we can correctly decode only $2^{n(H-\delta')}$ typical messages, each occurring with probability less than $2^{-n(H-\delta)}$ (the ε' is added to allow for the possibility that we manage to decode the atypical messages correctly). Since we may choose δ as small as we please, this success probability becomes small as $n \to \infty$.

We conclude that the optimal code compresses each letter to $H(X)$ bits asymptotically. This is Shannon's noiseless coding theorem.

5.1.2 Mutual information

The Shannon entropy $H(X)$ quantifies how much information is conveyed, on the average, by a letter drawn from the ensemble X, for it tells us how many bits are required (asymptotically as $n \to \infty$, where n is the number of letters drawn) to encode that information.

The mutual information $I(X;Y)$ quantifies how *correlated* two messages are. How much do we know about a message drawn from X^n when we have read a message drawn from Y^n?

For example, suppose we want to send a message from a transmitter to a receiver. But the communication channel is noisy, so that the message received (y) might differ from the message sent (x). The noisy channel can be characterized by the conditional probabilities $p(y|x)$ – the probability that y is received when x is sent. We suppose that the letter x is sent with *a priori* probability $p(x)$. We want to quantify how much we learn about x when we receive y; how much information do we gain?

As we have already seen, the entropy $H(X)$ quantifies my *a priori* ignorance per letter, before any message is received; that is, you would need to convey nH (noiseless) bits to me to completely specify (asymptotically) a particular message of n letters. But after I learn the value of y, I can use

Bayes' rule to update my probability distribution for x:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{5.14}$$

(I know $p(y|x)$ if I am familiar with the properties of the channel, and $p(x)$ if I know the *a priori* probabilities of the letters; thus I can compute $p(y) = \sum_x p(y|x)p(x)$.) Because of the new knowledge I have acquired, I am now less ignorant about x than before. Given the y's I have received, using an optimal code, you can specify a particular string of n letters by sending me

$$H(X|Y) = \langle -\log p(x|y)\rangle, \tag{5.15}$$

bits per letter. $H(X|Y)$ is called the "conditional entropy." From $p(x|y) = p(x,y)/p(y)$, we see that

$$\begin{aligned} H(X|Y) &= \langle -\log p(x,y) + \log p(y)\rangle \\ &= H(X,Y) - H(Y), \end{aligned} \tag{5.16}$$

and similarly

$$\begin{aligned} H(Y|X) &\equiv \langle -\log p(y|x)\rangle \\ &= \langle -\log \left(\frac{p(x,y)}{p(x)}\right)\rangle = H(X,Y) - H(X). \end{aligned} \tag{5.17}$$

We may interpret $H(X|Y)$, then, as the number of *additional* bits per letter needed to specify *both* x and y once y is known. Obviously, then, this quantity cannot be negative.

The information about X that I *gain* when I learn Y is quantified by how much the number of bits per letter needed to specify X is *reduced* when Y is known. Thus is

$$\begin{aligned} I(X;Y) &\equiv H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{5.18}$$

$I(X;Y)$ is called the mutual information. It is obviously symmetric under interchange of X and Y; I find out as much about X by learning Y as about Y

by learning X. Learning Y can never *reduce* my knowledge of X, so $I(X;Y)$ is obviously nonnegative. (The inequalities $H(X) \geq H(X|Y) \geq 0$ are easily proved using the convexity of the log function; see for example *Elements of Information Theory* by T. Cover and J. Thomas.)

Of course, if X and Y are completely uncorrelated, we have $p(x,y) = p(x)p(y)$, and

$$I(X;Y) \equiv \langle \log \frac{p(x,y)}{p(x)p(y)} \rangle = 0; \tag{5.19}$$

naturally, we can't find out about X by learning Y if there is no correlation!

5.1.3 The noisy channel coding theorem

If we want to communicate over a noisy channel, it is obvious that we can improve the reliability of transmission through redundancy. For example, I might send each bit many times, and the receiver could use majority voting to decode the bit.

But given a channel, is it always possible to find a code that can ensure arbitrarily good reliability (as $n \to \infty$)? And what can be said about the *rate* of such codes; *i.e.*, how many bits are required per letter of the message?

In fact, Shannon showed that any channel can be used for arbitrarily reliable communication at a finite (nonzero) rate, as long as there is *some* correlation between input and output. Furthermore, he found a useful expression for the optimal rate that can be attained. These results are the content of the "noisy channel coding theorem."

Suppose, to be concrete, that we are using a binary alphabet, 0 and 1 each occurring with *a priori* probability $\frac{1}{2}$. And suppose that the channel is the "binary symmetric channel" – it acts on each bit independently, flipping its value with probability p, and leaving it intact with probability $1-p$. That is, the conditional probabilities are

$$\begin{array}{ll} p(0|0) = 1-p, & p(0|1) = p, \\ p(1|0) = p, & p(1|1) = 1-p. \end{array} \tag{5.20}$$

We want to construct a family of codes of increasing block size n, such that the probability of a decoding error goes to zero as $n \to \infty$. If the number of bits encoded in the block is k, then the code consists of a choice of

2^k "codewords" among the 2^n possible strings of n bits. We define the rate R of the code (the number of data bits carried per bit transmitted) as

$$R = \frac{k}{n}. \tag{5.21}$$

We should design our code so that the code strings are as "far apart" as possible. That is for a given rate R, we want to maximize the number of bits that must be flipped to change one codeword to another (this number is called the "Hamming distance" between the two codewords).

For any input string of length n bits, errors will typically cause about np of the bits to flip – hence the input typically diffuses to one of about $2^{nH(p)}$ typical output strings (occupying a "sphere" of "Hamming radius" np about the input string). To decode reliably, we will want to choose our input codewords so that the error spheres of two different codewords are unlikely to overlap. Otherwise, two different inputs will sometimes yield the same output, and decoding errors will inevitably occur. If we are to avoid such decoding ambiguities, the total number of strings contained in all 2^{nR} error spheres must not exceed the total number 2^n of bits in the output message; we require

$$2^{nH(p)}2^{nR} \leq 2^n \tag{5.22}$$

or

$$R \leq 1 - H(p) \equiv C(p). \tag{5.23}$$

If transmission is highly reliable, we cannot expect the rate of the code to exceed $C(p)$. But is the rate $R = C(p)$ actually *attainable* (asymptotically)?

In fact transmission with R arbitrarily close to C and arbitrarily small error probability is possible. Perhaps the most ingenious of Shannon's ideas was to demonstrate that C can be attained by considering an average over "random codes." (Obviously, choosing a code at random is not the most clever possible procedure, but, perhaps surprisingly, it turns out that random coding achieves as high a rate (asymptotically for large n) as any other coding scheme.) Since C is the optimal rate for reliable transmission of data over the noisy channel it is called the *channel capacity*.

Suppose that 2^{nR} codewords are chosen at random by sampling the ensemble X^n. A message (one of the codewords) is sent. To decode the message, we draw a "Hamming sphere" around the message received that contains

$$2^{n(H(p)+\delta)}, \tag{5.24}$$

strings. The message is decoded as the codeword contained in this sphere, assuming such a codeword exists and is unique. If no such codeword exists, or the codeword is not unique, then we will assume that a decoding error occurs.

How likely is a decoding error? We have chosen the decoding sphere large enough so that failure of a valid codeword to appear in the sphere is atypical, so we need only worry about more than one valid codeword occupying the sphere. Since there are altogether 2^n possible strings, the Hamming sphere around the output contains a fraction

$$\frac{2^{n(H(p)+\delta)}}{2^n} = 2^{-n(C(p)-\delta)}, \tag{5.25}$$

of all strings. Thus, the probability that one of the 2^{nR} randomly chosen codewords occupies this sphere "by accident" is

$$2^{-n(C(p)-R-\delta)}, \tag{5.26}$$

Since we may choose δ as small as we please, R can be chosen as close to C as we please (but below C), and this error probability will still become exponentially small as $n \to \infty$.

So far we have shown that, the *average* probability of error is small, where we average over the choice of random code, and for each specified code, we also average over all codewords. Thus there must exist one particular code with average probability of error (averaged over codewords) less than ε. But we would like a stronger result – that the probability of error is small for *every* codeword.

To establish the stronger result, let P_i denote the probability of a decoding error when codeword i is sent. We have demonstrated the existence of a code such that

$$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} P_i < \varepsilon. \tag{5.27}$$

Let $N_{2\varepsilon}$ denote the number of codewords with $P_i > 2\varepsilon$. Then we infer that

$$\frac{1}{2^{nR}}(N_{2\varepsilon})2\varepsilon < \varepsilon \text{ or } N_{2\varepsilon} < 2^{nR-1}, \tag{5.28}$$

we see that we can throw away at most half of the codewords, to achieve $P_i < 2\varepsilon$ for *every* codeword. The new code we have constructed has

$$\text{Rate} = R - \frac{1}{n}, \tag{5.29}$$

which approaches R as $n \to \infty$

We have seen, then, that $C(p) = 1 - H(p)$ is the maximum rate that can be attained asymptotically with an arbitrarily small probability of error.

Consider now how these arguments generalize to more general alphabets and channels. We are given a channel specified by the $p(y|x)$'s, and let us specify a probability distribution $X = \{x, p(x)\}$ for the input letters. We will send strings of n letters, and we will assume that the channel acts on each letter independently. (A channel acting this way is said to be "memoryless.") Of course, once $p(y|x)$ and X are specified, $p(x|y)$ and $Y = \{y, p(y)\}$ are determined.

To establish an attainable rate, we again consider averaging over random codes, where codewords are chosen with *a priori* probability governed by X^n. Thus with high probability, these codewords will be chosen from a typical set of strings of letters, where there are about $2^{nH(X)}$ such typical strings.

For a typical received message in Y^n, there are about $2^{nH(X|Y)}$ messages that could have been sent. We may decode by associating with the received message a "sphere" containing $2^{n(H(X|Y)+\delta)}$ possible inputs. If there exists a unique codeword in this sphere, we decode the message as that codeword.

As before, it is unlikely that no codeword will be in the sphere, but we must exclude the possibility that there are more than one. Each decoding sphere contains a fraction

$$\begin{aligned} \frac{2^{n(H(X|Y)+\delta)}}{2^{nH(X)}} &= 2^{-n(H(X)-H(X|Y)-\delta)} \\ &= 2^{-n(I(X;Y)-\delta)}, \end{aligned} \tag{5.30}$$

of the typical inputs. If there are 2^{nR} codewords, the probability that any one falls in the decoding sphere by accident is

$$2^{nR}2^{-n(I(X;Y)-\delta)} = 2^{-n(I(X;Y)-R-\delta)}. \tag{5.31}$$

Since δ can be chosen arbitrarily small, we can choose R as close to I as we please (but less than I), and still have the probability of a decoding error become exponentially small as $n \to \infty$.

This argument shows that when we average over random codes and over codewords, the probability of an error becomes small for any rate $R < I$. The same reasoning as before then demonstrates the existence of a particular code with error probability $< \varepsilon$ for every codeword. This is a satisfying result, as it is consistent with our interpretation of I as the information that we gain about the input X when the signal Y is received – that is, I is the information per letter that we can send over the channel.

The mutual information $I(X;Y)$ depends not only on the channel conditional probabilities $p(y|x)$ but also on the priori probabilities $p(x)$ of the letters. The above random coding argument applies for any choice of the $p(x)$'s, so we have demonstrated that errorless transmission is possible for any rate R less than

$$C \equiv \underset{\{p(x)\}}{\mathrm{Max}} I(X;Y). \tag{5.32}$$

C is called the *channel capacity* and depends only on the conditional probabilities $p(y|x)$ that define the channel.

We have now shown that any rate $R < C$ is attainable, but is it possible for R to exceed C (with the error probability still approaching 0 for large n)? To show that C is an upper bound on the rate may seem more subtle in the general case than for the binary symmetric channel – the probability of error is different for different letters, and we are free to exploit this in the design of our code. However, we may reason as follows:

Suppose we have chosen 2^{nR} strings of n letters as our codewords. Consider a probability distribution (denoted $\tilde{X}^n$) in which each codeword occurs with equal probability ($= 2^{-nR}$). Evidently, then,

$$H(\tilde{X}^n) = nR. \tag{5.33}$$

Sending the codewords through the channel we obtain a probability distribution $\tilde{Y}^n$ of output states.

Because we assume that the channel acts on each letter independently, the conditional probability for a string of n letters factorizes:

$$p(y_1y_2\cdots y_n|x_1x_2\cdots x_n) = p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n), \tag{5.34}$$

and it follows that the conditional entropy satisfies

$$\begin{aligned} H(\tilde{Y}^n|\tilde{X}^n) &= \langle -\log p(y^n|x^n)\rangle = \sum_i \langle -\log p(y_i|x_i)\rangle \\ &= \sum_i H(\tilde{Y}_i|\tilde{X}_i), \end{aligned} \tag{5.35}$$

where $\tilde{X}_i$ and $\tilde{Y}_i$ are the marginal probability distributions for the ith letter determined by our distribution on the codewords. Recall that we also know that $H(X,Y) \leq H(X) + H(Y)$, or

$$H(\tilde{Y}^n) \leq \sum_i H(\tilde{Y}_i). \tag{5.36}$$

It follows that

$$\begin{aligned} I(\tilde{Y}^n;\tilde{X}^n) &= H(\tilde{Y}^n) - H(\tilde{Y}^n|\tilde{X}^n) \\ &\leq \sum_i (H(\tilde{Y}_i) - H(\tilde{Y}_i|\tilde{X}_i)) \\ &= \sum_i I(\tilde{Y}_i;\tilde{X}_i) \leq nC; \end{aligned} \tag{5.37}$$

the mutual information of the messages sent and received is bounded above by the sum of the mutual information per letter, and the mutual information for each letter is bounded above by the capacity (because C is defined as the maximum of $I(X;Y)$).

Recalling the symmetry of mutual information, we have

$$\begin{aligned} I(\tilde{X}^n;\tilde{Y}^n) &= H(\tilde{X}^n) - H(\tilde{X}^n|\tilde{Y}^n) \\ &= nR - H(\tilde{X}^n|\tilde{Y}^n) \leq nC. \end{aligned} \tag{5.38}$$

Now, if we can decode reliably as $n \to \infty$, this means that the input codeword is completely determined by the signal received, or that the conditional entropy of the input (per letter) must get small

$$\frac{1}{n} H(\tilde{X}^n|\tilde{Y}^n) \to 0. \tag{5.39}$$

If errorless transmission is possible, then, eq. (5.38) becomes

$$R \leq C, \tag{5.40}$$

in the limit $n \to \infty$. The rate cannot exceed the capacity. (Remember that the conditional entropy, unlike the mutual information, is *not* symmetric. Indeed $(1/n)H(\tilde{Y}^n|\tilde{X}^n)$ does *not* become small, because the channel introduces uncertainty about what message will be received. But if we can decode accurately, there is no uncertainty about what codeword was sent, once the signal has been received.)

We have now shown that the capacity C is the highest rate of communication through the noisy channel that can be attained, where the probability of error goes to zero as the number of letters in the message goes to infinity. This is Shannon's noisy channel coding theorem.

Of course the method we have used to show that $R = C$ is asymptotically attainable (averaging over random codes) is not very constructive. Since a random code has no structure or pattern, encoding and decoding would be quite unwieldy (we require an exponentially large code book). Nevertheless, the theorem is important and useful, because it tells us what is in principle attainable, and furthermore, what is not attainable, even in principle. Also, since $I(X;Y)$ is a concave function of $X = \{x, p(x)\}$ (with $\{p(y|x)\}$ fixed), it has a unique local maximum, and C can often be computed (at least numerically) for channels of interest.

5.2 Von Neumann Entropy

In classical information theory, we often consider a source that prepares messages of n letters ($n \gg 1$), where each letter is drawn independently from an ensemble $X = \{x, p(x)\}$. We have seen that the Shannon information $H(X)$ is the number of incompressible bits of information carried per letter (asymptotically as $n \to \infty$).

We may also be interested in correlations between messages. The correlations between two ensembles of letters X and Y are characterized by conditional probabilities $p(y|x)$. We have seen that the mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{5.41}$$

is the number of bits of information per letter about X that we can acquire by reading Y (or vice versa). If the $p(y|x)$'s characterize a noisy channel, then, $I(X;Y)$ is the amount of information per letter than can be transmitted through the channel (given the *a priori* distribution for the X's).

We would like to generalize these considerations to *quantum* information. So let us imagine a source that prepares messages of n letters, but where each letter is chosen from an ensemble of quantum states. The signal alphabet consists of a set of quantum states $\boldsymbol{\rho}_x$, each occurring with a specified *a priori* probability p_x.

As we have already discussed at length, the probability of any outcome of any measurement of a letter chosen from this ensemble, if the observer has no

knowledge about which letter was prepared, can be completely characterized by the density matrix

$$\boldsymbol{\rho} = \sum_x p_x \boldsymbol{\rho}_x; \tag{5.42}$$

for the POVM $\{\boldsymbol{F}_a\}$, we have

$$\mathrm{Prob}(a) = tr(\boldsymbol{F}_a \boldsymbol{\rho}). \tag{5.43}$$

For this (or any) density matrix, we may define the Von Neumann entropy

$$S(\boldsymbol{\rho}) = -tr(\boldsymbol{\rho} \log \boldsymbol{\rho}). \tag{5.44}$$

Of course, if we choose an orthonormal basis $\{|a\rangle\}$ that diagonalizes $\boldsymbol{\rho}$,

$$\boldsymbol{\rho} = \sum_a \lambda_a |a\rangle\langle a|, \tag{5.45}$$

then

$$S(\boldsymbol{\rho}) = H(A), \tag{5.46}$$

where $H(A)$ is the Shannon entropy of the ensemble $A = \{a, \lambda_a\}$.

In the case where the signal alphabet consists of mutually orthogonal pure states, the quantum source reduces to a classical one; all of the signal states can be perfectly distinguished, and $S(\boldsymbol{\rho}) = H(X)$. The quantum source is more interesting when the signal states $\boldsymbol{\rho}$ are not mutually commuting. We will argue that the Von Neumann entropy quantifies the incompressible information content of the quantum source (in the case where the signal states are pure) much as the Shannon entropy quantifies the information content of a classical source.

Indeed, we will find that Von Neumann entropy plays a dual role. It quantifies not only the *quantum* information content per letter of the ensemble (the minimum number of qubits per letter needed to reliably encode the information) but also its *classical* information content (the maximum amount of information per letter—in bits, not qubits—that we can gain about the preparation by making the best possible measurement). And, we will see that Von Neumann information enters quantum information in yet a third way: quantifying the entanglement of a bipartite pure state. Thus quantum information theory is largely concerned with the interpretation and uses of Von

Neumann entropy, much as classical information theory is largely concerned with the interpretation and uses of Shannon entropy.

In fact, the mathematical machinery we need to develop quantum information theory is very similar to Shannon's mathematics (typical sequences, random coding, ...); so similar as to sometimes obscure that the conceptional context is really quite different. The central issue in quantum information theory is that nonorthogonal pure quantum states cannot be perfectly distinguished, a feature with no classical analog.

5.2.1 Mathematical properties of $S(\boldsymbol{\rho})$

There are a handful of properties of $S(\boldsymbol{\rho})$ that are frequently useful (many of which are closely analogous to properties of $H(X)$). I list some of these properties below. Most of the proofs are not difficult (a notable exception is the proof of strong subadditivity), and are included in the exercises at the end of the chapter. Some proofs can also be found in A. Wehrl, "General Properties of Entropy," Rev. Mod. Phys. **50** (1978) 221, or in Chapter 9 of A. Peres, *Quantum Theory: Concepts and Methods.*

(1) Purity. A pure state $\boldsymbol{\rho} = |\varphi\rangle\langle\varphi|$ has $S(\boldsymbol{\rho}) = 0$.

(2) Invariance. The entropy is unchanged by a unitary change of basis:

$$S(\boldsymbol{U}\boldsymbol{\rho}\boldsymbol{U}^{-1}) = S(\boldsymbol{\rho}). \tag{5.47}$$

This is obvious, since $S(\boldsymbol{\rho})$ depends only on the eigenvalues of $\boldsymbol{\rho}$.

(3) Maximum. If $\boldsymbol{\rho}$ has D nonvanishing eigenvalues, then

$$S(\boldsymbol{\rho}) \leq \log D, \tag{5.48}$$

with equality when all the nonzero eigenvalues are equal. (The entropy is maximized when the quantum state is chosen *randomly.*)

(4) Concavity. For $\lambda_1, \lambda_2, \cdots, \lambda_n \geq 0$ and $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$

$$S(\lambda_1\boldsymbol{\rho}_1 + \cdots + \lambda_n\boldsymbol{\rho}_n) \geq \lambda_1 S(\boldsymbol{\rho}_1) + \cdots + \lambda_n S(\boldsymbol{\rho}_n). \tag{5.49}$$

That is, the Von Neumann entropy is larger if we are *more ignorant* about how the state was prepared. This property is a consequence of the convexity of the log function.

(5) Entropy of measurement. Suppose that, in a state $\boldsymbol{\rho}$, we measure the observable

$$\boldsymbol{A} = \sum_y |a_y\rangle a_y \langle a_y|, \tag{5.50}$$

so that the outcome a_y occurs with probability

$$p(a_y) = \langle a_y|\boldsymbol{\rho}|a_y\rangle. \tag{5.51}$$

Then the Shannon entropy of the ensemble of measurement outcomes $Y = \{a_y, p(a_y)\}$ satisfies

$$H(Y) \geq S(\boldsymbol{\rho}), \tag{5.52}$$

with equality when $\boldsymbol{A}$ and $\boldsymbol{\rho}$ commute. Mathematically, this is the statement that $S(\boldsymbol{\rho})$ increases if we replace all off-diagonal matrix elements of $\boldsymbol{\rho}$ by zero, in any basis. Physically, it says that the randomness of the measurement outcome is minimized if we choose to measure an observable that commutes with the density matrix. But if we measure a "bad" observable, the result will be less predictable.

(6) Entropy of preparation. If a pure state is drawn randomly from the ensemble $\{|\varphi_x\rangle, p_x\}$, so that the density matrix is

$$\boldsymbol{\rho} = \sum_x p_x |\varphi_x\rangle\langle\varphi_x|, \tag{5.53}$$

then

$$H(X) \geq S(\boldsymbol{\rho}), \tag{5.54}$$

with equality if the signal states $|\varphi_x\rangle$ are mutually orthogonal. This statement indicates that *distinguishability is lost* when we mix nonorthogonal pure states. (We can't fully recover the information about which state was prepared, because, as we'll discuss later on, the information gain attained by performing a measurement cannot exceed $S(\boldsymbol{\rho})$.)

(7) Subadditivity. Consider a bipartite system AB in the state $\boldsymbol{\rho}_{AB}$. Then

$$S(\boldsymbol{\rho}_{AB}) \leq S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_B), \tag{5.55}$$

(where $\boldsymbol{\rho}_A = \mathrm{tr}_B \boldsymbol{\rho}_{AB}$ and $\boldsymbol{\rho}_B = \mathrm{tr}_A \boldsymbol{\rho}_{AB}$), with equality for $\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$. Thus, entropy is *additive* for uncorrelated systems, but otherwise the entropy of the whole is less than the sum of the entropy of the parts. This property is analogous to the property

$$H(X,Y) \leq H(X) + H(Y), \tag{5.56}$$

(or $I(X;Y) \geq 0$) of Shannon entropy; it holds because some of the information in XY (or AB) is encoded in the correlations between X and Y (A and B).

(8) Strong subadditivity. For any state $\boldsymbol{\rho}_{ABC}$ of a tripartite system,

$$S(\boldsymbol{\rho}_{ABC}) + S(\boldsymbol{\rho}_B) \leq S(\boldsymbol{\rho}_{AB}) + S(\boldsymbol{\rho}_{BC}). \tag{5.57}$$

This property is called "strong" subadditivity in that it reduces to subadditivity in the event that B is one-dimensional. The proof of the corresponding property of Shannon entropy is quite simple, but the proof for Von Neumann entropy turns out to be surprisingly difficult (it is sketched in Wehrl). You may find the strong subadditivity property easier to remember by thinking about it this way: AB and BC can be regarded as two *overlapping* subsystems. The entropy of their union (ABC) plus the entropy of their intersection (B) does not exceed the sum of the entropies of the subsystems (AB and BC). We will see that strong subadditivity has deep and important consequences.

(9) Triangle inequality (Araki-Lieb inequality): For a bipartite system,

$$S(\boldsymbol{\rho}_{AB}) \geq |S(\boldsymbol{\rho}_A) - S(\boldsymbol{\rho}_B)|. \tag{5.58}$$

The triangle inequality contrasts sharply with the analogous property of Shannon entropy

$$H(X,Y) \geq H(X), H(Y), \tag{5.59}$$

or

$$H(X|Y), H(Y|X) \geq 0. \tag{5.60}$$

The Shannon entropy of a classical bipartite system exceeds the Shannon entropy of either part – there is more information in the whole system than in part of it! Not so for the Von Neumann entropy. In the extreme case of a bipartite pure quantum state, we have $S(\boldsymbol{\rho}_A) = S(\boldsymbol{\rho}_B)$ (and nonzero if the state is entangled) while $S(\boldsymbol{\rho}_{AB}) = 0$. The bipartite state has a definite preparation, but if we measure observables of the subsystems, the measurement outcomes are inevitably random and unpredictable. We cannot discern how the state was prepared by observing the two subsystems separately, rather, information is encoded in the nonlocal quantum correlations. The juxtaposition of the positivity of conditional Shannon entropy (in the classical case) with the triangle inequality (in the quantum case) nicely characterizes a key distinction between quantum and classical information.

5.2.2 Entropy and thermodynamics

Of course, the concept of entropy first entered science through the study of thermodynamics. I will digress briefly here on some thermodynamic implications of the mathematic properties of $S(\boldsymbol{\rho})$.

There are two distinct (but related) possible approaches to the foundations of quantum statistical physics. In the first, we consider the evolution of an isolated (closed) quantum system, but we perform some *coarse graining* to define our thermodynamic variables. In the second approach, which is perhaps better motivated physically, we consider an *open* system, a quantum system in contact with its environment, and we track the evolution of the open system without monitoring the environment.

For an open system, the crucial mathematical property of the Von Neumann entropy is *subadditivity*. If the system (A) and environment (E) are initially uncorrelated with one another

$$\boldsymbol{\rho}_{AE} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_E, \tag{5.61}$$

then entropy is additive:

$$S(\boldsymbol{\rho}_{AE}) = S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_E). \tag{5.62}$$

Now suppose that the open system evolves for a while. The evolution is described by a unitary operator $\boldsymbol{U}_{AE}$ that acts on the combined system A plus E:

$$\boldsymbol{\rho}_{AE} \to \boldsymbol{\rho}'_{AE} = \boldsymbol{U}_{AE}\boldsymbol{\rho}_{AE}\boldsymbol{U}^{-1}_{AE}, \tag{5.63}$$

and since unitary evolution preserves S, we have

$$S(\boldsymbol{\rho}'_{AE}) = S(\boldsymbol{\rho}_{AE}). \tag{5.64}$$

Finally, we apply subadditivity to the state $\boldsymbol{\rho}'_{AE}$ to infer that

$$S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_E) = S(\boldsymbol{\rho}'_{AE}) \leq S(\boldsymbol{\rho}'_A) + S(\boldsymbol{\rho}'_E), \tag{5.65}$$

(with equality in the event that A and E remain uncorrelated). If we define the "total" entropy of the world as the sum of the entropy of the system and the entropy of the environment, we conclude that *the entropy of the world cannot decrease*. This is one form of the second law of thermodynamics. But note that we assumed that system and environment were initially uncorrelated to derive this "law."

Typically, the interaction of system and environment *will* induce correlations so that (assuming no initial correlations) the entropy will actually *increase*. From our discussion of the master equation, in §3.5 you'll recall that the environment typically "forgets" quickly, so that if our time resolution is coarse enough, we can regard the system and environment as "initially" uncorrelated (in effect) at each instant of time (the Markovian approximation). Under this assumption, the "total" entropy will increase monotonically, asymptotically approaching its theoretical maximum, the largest value it can attain consistent with all relevant conservation laws (energy, charge, baryon number, etc.)

Indeed, the usual assumption underlying quantum statistical physics is that system and environment are in the "most probable configuration," that which maximizes $S(\boldsymbol{\rho}_A)+S(\boldsymbol{\rho}_E)$. In this configuration, all "accessible" states are equally likely.

From a microscopic point of view, information initially encoded in the system (our ability to distinguish one initial state from another, initially orthogonal, state) is lost; it winds up encoded in quantum entanglement between system and environment. In principle that information could be recovered, but in practice it is totally inaccessible to localized observers. Hence thermodynamic irreversibility.

Of course, we can adapt this reasoning to apply to a large closed system (the whole universe?). We may divide the system into a small part of the whole and the rest (the environment of the small part). Then the sum of the entropies of the parts will be nondecreasing. This is a particular type of coarse graining. That part of a closed system behaves like an open system

is why the microcanonical and canonical ensembles of statistical mechanics yield the same predictions for large systems.

5.3 Quantum Data Compression

What is the quantum analog of the noiseless coding theorem?

We consider a long message consisting of n letters, where each letter is chosen at random from the ensemble of pure states

$$\{|\varphi_x\rangle, p_x\}, \tag{5.66}$$

and the $|\varphi_x\rangle$'s are not necessarily mutually orthogonal. (For example, each $|\varphi_x\rangle$ might be the polarization state of a single photon.) Thus, each letter is described by the density matrix

$$\boldsymbol{\rho} = \sum_x p_x |\varphi_x\rangle\langle\varphi_x|, \tag{5.67}$$

and the entire message has the density matrix

$$\boldsymbol{\rho}^n = \boldsymbol{\rho} \otimes \cdots \otimes \boldsymbol{\rho}. \tag{5.68}$$

Now we ask, how *redundant* is this quantum information? We would like to devise a *quantum code* that enables us to compress the message to a smaller Hilbert space, but without compromising the fidelity of the message. For example, perhaps we have a quantum memory device (the hard disk of a quantum computer?), and we know the *statistical* properties of the recorded data (*i.e.*, we know $\boldsymbol{\rho}$). We want to conserve space on the device by compressing the data.

The optimal compression that can be attained was found by Ben Schumacher. Can you guess the answer? The best possible compression compatible with arbitrarily good fidelity as $n \to \infty$ is compression to a Hilbert space $\mathcal{H}$ with

$$\log(\dim \mathcal{H}) = nS(\boldsymbol{\rho}). \tag{5.69}$$

In this sense, the Von Neumann entropy is the number of *qubits* of quantum information carried per letter of the message. For example, if the message consists of n photon polarization states, we can compress the message to

$m = nS(\boldsymbol{\rho})$ photons – compression is always possible unless $\boldsymbol{\rho} = \frac{1}{2}\mathbf{1}$. (We can't compress random qubits just as we can't compress random bits.)

Once Shannon's results are known and understood, the proof of Schumacher's theorem is not difficult. Schumacher's important contribution was to ask the right question, and so to establish for the first time a precise (quantum) information theoretic interpretation of Von Neumann entropy.[2]

5.3.1 Quantum data compression: an example

Before discussing Schumacher's quantum data compression protocol in full generality, it is helpful to consider a simple example. So suppose that our letters are single qubits drawn from the ensemble

$$\begin{array}{ll} |\uparrow_z\rangle = \binom{1}{0} & p = \frac{1}{2}, \\ |\uparrow_x\rangle = \binom{1/\sqrt{2}}{1/\sqrt{2}} & p = \frac{1}{2}, \end{array} \tag{5.70}$$

so that the density matrix of each letter is

$$\begin{aligned} \boldsymbol{\rho} &= \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\uparrow_x\rangle\langle\uparrow_x| \\ &= \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \end{aligned} \tag{5.71}$$

As is obvious from symmetry, the eigenstates of $\boldsymbol{\rho}$ are qubits oriented up and down along the axis $\hat{n} = \frac{1}{\sqrt{2}}(\hat{x} + \hat{z})$,

$$\begin{aligned} |0'\rangle &\equiv |\uparrow_{\hat{n}}\rangle = \begin{pmatrix} \cos\frac{\pi}{8} \\ \sin\frac{\pi}{8} \end{pmatrix}, \\ |1'\rangle &\equiv |\downarrow_{\hat{n}}\rangle = \begin{pmatrix} \sin\frac{\pi}{8} \\ -\cos\frac{\pi}{8} \end{pmatrix}; \end{aligned} \tag{5.72}$$

the eigenvalues are

$$\begin{aligned} \lambda(0') &= \frac{1}{2} + \frac{1}{2\sqrt{2}} = \cos^2\frac{\pi}{8}, \\ \lambda(1') &= \frac{1}{2} - \frac{1}{2\sqrt{2}} = \sin^2\frac{\pi}{8}; \end{aligned} \tag{5.73}$$

[2] An interpretation of $S(\boldsymbol{\rho})$ in terms of *classical* information encoded in quantum states was actually known earlier, as we'll soon discuss.

(evidently $\lambda(0') + \lambda(1') = 1$ and $\lambda(0')\lambda(1') = \frac{1}{8} = \det\boldsymbol{\rho}$). The eigenstate $|0'\rangle$ has equal (and relatively large) overlap with both signal states

$$|\langle 0'|\uparrow_z\rangle|^2 = |\langle 0'|\uparrow_x\rangle|^2 = \cos^2\frac{\pi}{8} = .8535, \tag{5.74}$$

while $|1'\rangle$ has equal (and relatively small) overlap with both

$$|\langle 1'|\uparrow_z\rangle|^2 = |\langle 1'|\uparrow_x\rangle|^2 = \sin^2\frac{\pi}{8} = .1465. \tag{5.75}$$

Thus if we don't know whether $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$ was sent, the best guess we can make is $|\psi\rangle = |0'\rangle$. This guess has the maximal *fidelity*

$$F = \frac{1}{2}|\langle\uparrow_z|\psi\rangle|^2 + \frac{1}{2}|\langle\uparrow_x|\psi\rangle|^2, \tag{5.76}$$

among all possible qubit states $|\psi\rangle$ ($F = .8535$).

Now imagine that Alice needs to send three letters to Bob. But she can afford to send only two qubits (quantum channels are very expensive!). Still she wants Bob to reconstruct her state with the highest possible fidelity.

She could send Bob two of her three letters, and ask Bob to guess $|0'\rangle$ for the third. Then Bob receives the two letters with $F = 1$, and he has $F = .8535$ for the third; hence $F = .8535$ overall. But is there a more clever procedure that achieves higher fidelity?

There *is* a better procedure. By diagonalizing $\boldsymbol{\rho}$, we decomposed the Hilbert space of a single qubit into a "likely" one-dimensional subspace (spanned by $|0'\rangle$) and an "unlikely" one-dimensional subspace (spanned by $|1'\rangle$). In a similar way we can decompose the Hilbert space of three qubits into likely and unlikely subspaces. If $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle|\psi_3\rangle$ is any signal state (with each of three qubits in either the $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$ state), we have

$$\begin{aligned} &|\langle 0'0'0'|\psi\rangle|^2 = \cos^6\left(\frac{\pi}{8}\right) = .6219, \\ &|\langle 0'0'1'|\psi\rangle|^2 = |\langle 0'1'0'|\psi\rangle|^2 = |\langle 1'0'0'|\psi\rangle|^2 = \cos^4\left(\frac{\pi}{8}\right)\sin^2\left(\frac{\pi}{8}\right) = .1067, \\ &|\langle 0'1'1'|\psi\rangle|^2 = |\langle 1'0'1'|\psi\rangle|^2 = |\langle 1'1'0'|\psi\rangle|^2 = \cos^2\left(\frac{\pi}{8}\right)\sin^4\left(\frac{\pi}{8}\right) = .0183, \\ &|\langle 1'1'1'|\psi\rangle|^2 = \sin^6\left(\frac{\pi}{8}\right) = .0031. \end{aligned} \tag{5.77}$$

Thus, we may decompose the space into the likely subspace Λ spanned by $\{|0'0'0'\rangle, |0'0'1'\rangle, |0'1'0'\rangle, |1'0'0'\rangle\}$, and its orthogonal complement $\Lambda^\perp$. If we make a ("fuzzy") measurement that projects a signal state onto Λ or $\Lambda^\perp$, the probability of projecting onto the likely subspace is

$$P_{likely} = .6219 + 3(.1067) = .9419, \tag{5.78}$$

while the probability of projecting onto the unlikely subspace is

$$P_{unlikely} = 3(.0183) + .0031 = .0581. \tag{5.79}$$

To perform this fuzzy measurement, Alice could, for example, first apply a unitary transformation $\boldsymbol{U}$ that rotates the four high-probability basis states to

$$|\cdot\rangle|\cdot\rangle|0\rangle, \tag{5.80}$$

and the four low-probability basis states to

$$|\cdot\rangle|\cdot\rangle|1\rangle; \tag{5.81}$$

then Alice measures the third qubit to complete the fuzzy measurement. If the outcome is $|0\rangle$, then Alice's input state has been projected (in effect) onto Λ. She sends the remaining two (unmeasured) qubits to Bob. When Bob receives this (compressed) two-qubit state $|\psi_{\text{comp}}\rangle$, he decompresses it by appending $|0\rangle$ and applying $\boldsymbol{U}^{-1}$, obtaining

$$|\psi'\rangle = \boldsymbol{U}^{-1}(|\psi_{\text{comp}}\rangle|0\rangle). \tag{5.82}$$

If Alice's measurement of the third qubit yields $|1\rangle$, she has projected her input state onto the low-probability subspace $\Lambda^\perp$. In this event, the best thing she can do is send the state that Bob will decompress to the most likely state $|0'0'0'\rangle$ – that is, she sends the state $|\psi_{\text{comp}}\rangle$ such that

$$|\psi'\rangle = \boldsymbol{U}^{-1}(|\psi_{\text{comp}}\rangle|0\rangle) = |0'0'0'\rangle. \tag{5.83}$$

Thus, if Alice encodes the three-qubit signal state $|\psi\rangle$, sends two qubits to Bob, and Bob decodes as just described, then Bob obtains the state $\boldsymbol{\rho}'$

$$|\psi\rangle\langle\psi| \to \boldsymbol{\rho}' = \boldsymbol{E}|\psi\rangle\langle\psi|\boldsymbol{E} + |0'0'0'\rangle\langle\psi|(1-\boldsymbol{E})|\psi\rangle\langle 0'0'0'|, \tag{5.84}$$

where $\boldsymbol{E}$ is the projection onto Λ. The fidelity achieved by this procedure is

$$\begin{aligned} F = \langle\psi|\boldsymbol{\rho}'|\psi\rangle &= (\langle\psi|\boldsymbol{E}|\psi\rangle)^2 + (\langle\psi|(1-\boldsymbol{E})|\psi\rangle)(\langle\psi|0'0'0'\rangle)^2 \\ &= (.9419)^2 + (.0581)(.6219) = .9234. \end{aligned} \tag{5.85}$$

This is indeed better than the naive procedure of sending two of the three qubits each with perfect fidelity.

As we consider longer messages with more letters, the fidelity of the compression improves. The Von-Neumann entropy of the one-qubit ensemble is

$$S(\boldsymbol{\rho}) = H\left(\cos^2\frac{\pi}{8}\right) = .60088\ldots \tag{5.86}$$

Therefore, according to Schumacher's theorem, we can shorten a long message by the factor (say) .6009, and still achieve very good fidelity.

5.3.2 Schumacher encoding in general

The key to Shannon's noiseless coding theorem is that we can code the typical sequences and ignore the rest, without much loss of fidelity. To quantify the compressibility of quantum information, we promote the notion of a typical *sequence* to that of a typical *subspace*. The key to Schumacher's noiseless quantum coding theorem is that we can code the typical subspace and ignore its orthogonal complement, without much loss of fidelity.

We consider a message of n letters where each letter is a pure quantum state drawn from the ensemble $\{|\varphi_x\rangle, p_x\}$, so that the density matrix of a single letter is

$$\boldsymbol{\rho} = \sum_x p_x|\varphi_x\rangle\langle\varphi_x|. \tag{5.87}$$

Furthermore, the letters are drawn independently, so that the density matrix of the entire message is

$$\boldsymbol{\rho}^n \equiv \boldsymbol{\rho}\otimes\cdots\otimes\boldsymbol{\rho}. \tag{5.88}$$

We wish to argue that, for n large, this density matrix has nearly all of its support on a subspace of the full Hilbert space of the messages, where the dimension of this subspace asymptotically approaches $2^{nS(\boldsymbol{\rho})}$.

This conclusion follows directly from the corresponding classical statement, if we consider the orthonormal basis in which $\boldsymbol{\rho}$ is diagonal. Working in this basis, we may regard our quantum information source as an effectively classical source, producing messages that are strings of $\boldsymbol{\rho}$ eigenstates, each with a probability given by the product of the corresponding eigenvalues.

For a specified n and δ, define the typical subspace Λ as the space spanned by the eigenvectors of $\boldsymbol{\rho}^n$ with eigenvalues λ satisfying

$$2^{-n(S-\delta)} \geq \lambda \geq e^{-n(S+\delta)}. \tag{5.89}$$

Borrowing directly from Shannon, we conclude that for any $\delta, \varepsilon > 0$ and n sufficiently large, the sum of the eigenvalues of $\boldsymbol{\rho}^n$ that obey this condition satisfies

$$\mathrm{tr}(\boldsymbol{\rho}^n \boldsymbol{E}) > 1 - \varepsilon, \tag{5.90}$$

(where $\boldsymbol{E}$ denotes the projection onto the typical subspace) and the number $\dim(\Lambda)$ of such eigenvalues satisfies

$$2^{n(S+\delta)} \geq \dim(\Lambda) \geq (1-\varepsilon)2^{n(S-\delta)}. \tag{5.91}$$

Our coding strategy is to send states in the typical subspace faithfully. For example, we can make a fuzzy measurement that projects the input message onto either Λ or $\Lambda^\perp$; the outcome will be Λ with probability $P_\Lambda = \mathrm{tr}(\boldsymbol{\rho}^n \boldsymbol{E}) > 1 - \varepsilon$. In that event, the projected state is coded and sent. Asymptotically, the probability of the other outcome becomes negligible, so it matters little what we do in that case.

The coding of the projected state merely packages it so it can be carried by a minimal number of qubits. For example, we apply a unitary change of basis $\boldsymbol{U}$ that takes each state $|\psi_{\mathrm{typ}}\rangle$ in Λ to a state of the form

$$\boldsymbol{U}|\psi_{\mathrm{typ}}\rangle = |\psi_{\mathrm{comp}}\rangle|0_{\mathrm{rest}}\rangle, \tag{5.92}$$

where $|\psi_{\mathrm{comp}}\rangle$ is a state of $n(S+\delta)$ qubits, and $|0_{\mathrm{rest}}\rangle$ denotes the state $|0\rangle \otimes \ldots \otimes |0\rangle$ of the remaining qubits. Alice sends $|\psi_{\mathrm{comp}}\rangle$ to Bob, who decodes by appending $|0_{\mathrm{rest}}\rangle$ and applying $\boldsymbol{U}^{-1}$.

Suppose that

$$|\varphi_i\rangle = |\varphi_{x_1(i)}\rangle \ldots |\varphi_{x_n(i)}\rangle, \tag{5.93}$$

denotes any one of the n-letter pure state messages that might be sent. After coding, transmission, and decoding are carried out as just described, Bob has reconstructed a state

$$\begin{aligned} |\varphi_i\rangle\langle\varphi_i| \to \boldsymbol{\rho}_i' = \ & \boldsymbol{E}|\varphi_i\rangle\langle\varphi_i|\boldsymbol{E} \\ & + \boldsymbol{\rho}_{i,\mathrm{Junk}}\langle\varphi_i|(\boldsymbol{1}-\boldsymbol{E})|\varphi_i\rangle, \end{aligned} \tag{5.94}$$

where $\boldsymbol{\rho}_{i,\mathrm{Junk}}$ is the state we choose to send if the fuzzy measurement yields the outcome $\Lambda^\perp$. What can we say about the fidelity of this procedure?

The fidelity varies from message to message (in contrast to the example discussed above), so we consider the fidelity averaged over the ensemble of possible messages:

$$\begin{aligned} F &= \sum_i p_i \langle\varphi_i|\boldsymbol{\rho}_i'|\varphi_i\rangle \\ &= \sum_i p_i \langle\varphi_i|\boldsymbol{E}|\varphi_i\rangle\langle\varphi_i|\boldsymbol{E}|\varphi_i\rangle + \sum_i p_i \langle\varphi_i|\boldsymbol{\rho}_{i,\mathrm{Junk}}|\varphi_i\rangle\langle\varphi_i|\boldsymbol{1}-\boldsymbol{E}|\varphi_i\rangle \\ &\geq \sum_i p_i \parallel \boldsymbol{E}|\varphi_i\rangle \parallel^4, \end{aligned} \tag{5.95}$$

where the last inequality holds because the "junk" term is nonnegative. Since any real number satisfies

$$(x-1)^2 \geq 0, \text{ or } x^2 \geq 2x - 1, \tag{5.96}$$

we have (setting $x = \parallel \boldsymbol{E}|\varphi_i\rangle \parallel^2$)

$$\parallel \boldsymbol{E}|\varphi_i\rangle \parallel^4 \geq 2 \parallel \boldsymbol{E}|\varphi_i\rangle \parallel^2 - 1 = 2\langle\varphi_i|\boldsymbol{E}|\varphi_i\rangle - 1, \tag{5.97}$$

and hence

$$\begin{aligned} F &\geq \sum_i p_i (2\langle\varphi_i|\boldsymbol{E}|\varphi_i\rangle - 1) \\ &= 2\ \mathrm{tr}(\boldsymbol{\rho}^n\boldsymbol{E}) - 1 > 2(1-\varepsilon) - 1 = 1 - 2\varepsilon. \end{aligned} \tag{5.98}$$

We have shown, then, that it is possible to compress the message to fewer than $n(S+\delta)$ qubits, while achieving an average fidelity that becomes arbitrarily good a n gets large.

So we have established that the message may be compressed, with insignificant loss of fidelity, to $S+\delta$ qubits per letter. Is further compression possible?

Let us suppose that Bob will decode the message $\boldsymbol{\rho}_{\mathrm{comp},i}$ that he receives by appending qubits and applying a unitary transformation $\boldsymbol{U}^{-1}$, obtaining

$$\boldsymbol{\rho}_i' = \boldsymbol{U}^{-1}(\boldsymbol{\rho}_{\mathrm{comp},i} \otimes |0\rangle\langle 0|)\boldsymbol{U} \tag{5.99}$$

("unitary decoding"). Suppose that $\boldsymbol{\rho}_{\mathrm{comp}}$ has been compressed to $n(S-\delta)$ qubits. Then, *no matter how the input message have been encoded*, the decoded messages are all contained in a subspace Λ' of Bob's Hilbert space of dimension $2^{n(S-\delta)}$. (We are *not* assuming now that Λ' has anything to do with the typical subspace.)

If the input message is $|\varphi_i\rangle$, then the message reconstructed by Bob is $\boldsymbol{\rho}'_i$ which can be diagonalized as

$$\boldsymbol{\rho}'_i = \sum_{a_i} |a_i\rangle \lambda_{a_i} \langle a_i|, \tag{5.100}$$

where the $|a_i\rangle$'s are mutually orthogonal states in Λ'. The fidelity of the reconstructed message is

$$\begin{aligned} F_i &= \langle\varphi_i|\boldsymbol{\rho}'_i|\varphi_i\rangle \\ &= \sum_{a_i} \lambda_{a_i} \langle\varphi_i|a_i\rangle\langle a_i|\varphi_i\rangle \\ &\le \sum_{a_i} \langle\varphi_i|a_i\rangle\langle a_i|\varphi_i\rangle \le \langle\varphi_i|\boldsymbol{E}'|\varphi_i\rangle, \end{aligned} \tag{5.101}$$

where $\boldsymbol{E}'$ denotes the orthogonal projection onto the subspace Λ'. The average fidelity therefore obeys

$$F = \sum_i p_i F_i \le \sum_i p_i \langle\varphi_i|\boldsymbol{E}'|\varphi_i\rangle = \mathrm{tr}(\boldsymbol{\rho}^n \boldsymbol{E}'). \tag{5.102}$$

But since $\boldsymbol{E}'$ projects onto a space of dimension $2^{n(S-\delta)}$, $\mathrm{tr}(\boldsymbol{\rho}^n\boldsymbol{E}')$ can be no larger than the sum of the $2^{n(S-\delta)}$ largest eigenvalues of $\boldsymbol{\rho}^n$. It follows from the properties of typical subspaces that this sum becomes as small as we please; for n large enough

$$F \le \mathrm{tr}(\boldsymbol{\rho}^n\boldsymbol{E}') < \varepsilon. \tag{5.103}$$

Thus we have shown that, if we attempt to compress to $S-\delta$ qubits per letter, then the fidelity inevitably becomes poor for n sufficiently large. We conclude then, that $S(\boldsymbol{\rho})$ qubits per letter is the optimal compression of the quantum information that can be attained if we are to obtain good fidelity as n goes to infinity. This is Schumacher's noiseless quantum coding theorem.

The above argument applies to any conceivable encoding scheme, but only to a restricted class of decoding schemes (unitary decodings). A more general decoding scheme can certainly be contemplated, described by a *superoperator*. More technology is then required to prove that better compression than S

qubits per letter is not possible. But the conclusion is the same. The point is that $n(S-\delta)$ qubits are not sufficient to distinguish all of the typical states.

To summarize, there is a close analogy between Shannon's noiseless coding theorem and Schumacher's noiseless quantum coding theorem. In the classical case, nearly all long messages are typical sequences, so we can code only these and still have a small probability of error. In the quantum case, nearly all long messages have nearly unit overlap with the typical subspace, so we can code only the typical subspace and still achieve good fidelity.

In fact, Alice could send effectively classical information to Bob—the string $x_1x_2\cdots x_n$ encoded in mutually orthogonal quantum states—and Bob could then follow these classical instructions to reconstruct Alice's state. By this means, they could achieve high-fidelity compression to $H(X)$ bits—or qubits—per letter. But if the letters are drawn from an ensemble of *nonorthogonal* pure states, this amount of compression is not optimal; some of the classical information about the preparation of the state has become redundant, because the nonorthogonal states cannot be perfectly distinguished. Thus Schumacher coding can go further, achieving optimal compression to $S(\boldsymbol{\rho})$ qubits per letter. The information has been packaged more efficiently, but at a price—Bob has received what Alice intended, but Bob can't know what he has. In contrast to the classical case, Bob can't make any measurement that is certain to decipher Alice's message correctly. An attempt to read the message will unavoidably disturb it.

5.3.3 Mixed-state coding: Holevo information

The Schumacher theorem characterizes the compressibility of an ensemble of pure states. But what if the letters are drawn from an ensemble of *mixed* states? The compressibility in that case is not firmly established, and is the subject of current research.[3]

It is easy to see that $S(\boldsymbol{\rho})$ won't be the answer for mixed states. To give a trivial example, suppose that a particular mixed state $\boldsymbol{\rho}_0$ with $S(\boldsymbol{\rho}_0) \neq 0$ is chosen with probability $p_0 = 1$. Then the message is always $\boldsymbol{\rho}_0 \otimes \boldsymbol{\rho}_0 \otimes \cdots \otimes \boldsymbol{\rho}_0$ and it carries no information; Bob can reconstruct the message perfectly without receiving *anything* from Alice. Therefore, the message can be compressed to zero qubits per letters, which is less than $S(\boldsymbol{\rho}) > 0$.

To construct a slightly less trivial example, recall that for an ensemble of

[3]See M. Horodecki, `quant-ph/9712035`.

mutually orthogonal pure states, the Shannon entropy of the ensemble equals the Von Neumann entropy

$$H(X) = S(\boldsymbol{\rho}), \tag{5.104}$$

so that the classical and quantum compressibility coincide. This makes sense, since the orthogonal states are perfectly distinguishable. In fact, if Alice wants to send the message

$$|\varphi_{x_1}\rangle\varphi_{x_2}\rangle \cdots |\varphi_{x_n}\rangle \tag{5.105}$$

to Bob, she can send the classical message $x_1 \ldots x_n$ to Bob, who can reconstruct the state with perfect fidelity.

But now suppose that the letters are drawn from an ensemble of mutually orthogonal *mixed* states $\{\boldsymbol{\rho}_x, p_x\}$,

$$tr\boldsymbol{\rho}_x\boldsymbol{\rho}_y = 0 \text{ for } x \neq y; \tag{5.106}$$

that is, $\boldsymbol{\rho}_x$ and $\boldsymbol{\rho}_y$ have support on mutually orthogonal subspaces of the Hilbert space. These mixed states are also perfectly distinguishable, so again the messages are essentially classical, and therefore can be compressed to $H(X)$ qubits per letter. For example, we can extend the Hilbert space $\mathcal{H}_A$ of our letters to the larger space $\mathcal{H}_A \otimes \mathcal{H}_B$, and choose a purification of each $\boldsymbol{\rho}_x$, a pure state $|\varphi_x\rangle_{AB} \in \mathcal{H}_A \otimes \mathcal{H}_B$ such that

$$tr_B(|\varphi_x\rangle_{AB}\ {}_{AB}\langle\varphi_x|) = (\boldsymbol{\rho}_x)_A. \tag{5.107}$$

These pure states are mutually orthogonal, and the ensemble $\{|\varphi_x\rangle_{AB}, p_x\}$ has Von Neumann entropy $H(X)$; hence we may Schumacher compress a message

$$|\varphi_{x_1}\rangle_{AB} \cdots |\varphi_{x_n}\rangle_{AB}, \tag{5.108}$$

to $H(X)$ qubits per letter (asymptotically). Upon decompressing this state, Bob can perform the partial trace by "throwing away" subsystem B, and so reconstruct Alice's message.

To make a reasonable guess about what expression characterizes the compressibility of a message constructed from a mixed state alphabet, we might seek a formula that reduces to $S(\boldsymbol{\rho})$ for an ensemble of pure states, and to

$H(X)$ for an ensemble of mutually orthogonal mixed states. Choosing a basis in which

$$\boldsymbol{\rho} = \sum_x p_x \boldsymbol{\rho}_x, \tag{5.109}$$

is block diagonalized, we see that

$$\begin{aligned} S(\boldsymbol{\rho}) &= -\mathrm{tr}\boldsymbol{\rho}\log\boldsymbol{\rho} = -\sum_x \mathrm{tr}(p_x\boldsymbol{\rho}_x)\log(p_x\boldsymbol{\rho}_x) \\ &= -\sum_x p_x \log p_x - \sum_x p_x \mathrm{tr}\boldsymbol{\rho}_x \log\boldsymbol{\rho}_x \\ &= H(X) + \sum_x p_x S(\boldsymbol{\rho}_x), \end{aligned} \tag{5.110}$$

(recalling that $\mathrm{tr}\boldsymbol{\rho}_x = 1$ for each x). Therefore we may write the Shannon entropy as

$$H(X) = S(\boldsymbol{\rho}) - \sum_x p_x S(\boldsymbol{\rho}_x) \equiv \chi(\mathcal{E}). \tag{5.111}$$

The quantity $\chi(\mathcal{E})$ is called the *Holevo information* of the ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$. Evidently, it depends not just on the density matrix $\boldsymbol{\rho}$, but also on the particular way that $\boldsymbol{\rho}$ is realized as an ensemble of mixed states. We have found that, for either an ensemble of pure states, or for an ensemble of *mutually orthogonal* mixed states, the Holevo information $\chi(\mathcal{E})$ is the optimal number of qubits per letter that can be attained if we are to compress the messages while retaining good fidelity for large n.

The Holevo information can be regarded as a generalization of Von Neumann entropy, reducing to $S(\boldsymbol{\rho})$ for an ensemble of pure states. It also bears a close resemblance to the mutual information of classical information theory:

$$I(Y;X) = H(Y) - H(Y|X) \tag{5.112}$$

tells us how much, on the average, the Shannon entropy of Y is reduced once we learn the value of X; similarly,

$$\chi(\mathcal{E}) = S(\boldsymbol{\rho}) - \sum_x p_x S(\boldsymbol{\rho}_x) \tag{5.113}$$

tells us how much, on the average, the Von Neumann entropy of an ensemble is reduced when we know which preparation was chosen. Like the classical mutual information, the Holevo information is always nonnegative, as follows from the concavity property of $S(\boldsymbol{\rho})$,

$$S(\sum p_x \boldsymbol{\rho}_x) \geq \sum_x p_x S(\boldsymbol{\rho}_x). \tag{5.114}$$

Now we wish to explore the connection between the Holevo information and the compressibility of messages constructed from an alphabet of *nonorthogonal* mixed states. In fact, it can be shown that, in general, high-fidelity compression to less than χ qubits per letter is not possible.

To establish this result we use a "monotonicity" property of χ that was proved by Lindblad and by Uhlmann: A superoperator cannot increase the Holevo information. That is, if \$ is any superoperator, let it act on an ensemble of mixed states according to

$$\$: \mathcal{E} = \{\boldsymbol{\rho}_x, p_x\} \to \mathcal{E}' = \{\$(\boldsymbol{\rho}_x), p_x\}; \tag{5.115}$$

then

$$\chi(\mathcal{E}') \leq \chi(\mathcal{E}). \tag{5.116}$$

Lindblad–Uhlmann monotonicity is closely related to the strong subadditivity of the Von Neumann entropy, as you will show in a homework exercise.

The monotonicity of χ provides a further indication that χ quantifies an amount of information encoded in a quantum system. The decoherence described by a superoperator can only retain or reduce this quantity of information – it can never increase it. Note that, in contrast, the Von Neumann entropy is not monotonic. A superoperator might take an initial pure state to a mixed state, increasing $S(\boldsymbol{\rho})$. But another superoperator takes every mixed state to the "ground state" $|0\rangle\langle 0|$, and so reduces the entropy of an initial mixed state to zero. It would be misleading to interpret this reduction of S as an "information gain," in that our ability to distinguish the different possible preparations has been completely destroyed. Correspondingly, decay to the ground state reduces the Holevo information to zero, reflecting that we have lost the ability to reconstruct the initial state.

We now consider messages of n letters, each drawn independently from the ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$; the ensemble of all such input messages is denoted $\mathcal{E}^{(n)}$. A code is constructed that compresses the messages so that they all occupy a Hilbert space $\tilde{\mathcal{H}}^{(n)}$; the ensemble of compressed messages is denoted $\tilde{\mathcal{E}}^{(n)}$. Then decompression is performed with a superoperator \$,

$$\$: \tilde{\mathcal{E}}^{(n)} \to \mathcal{E}'^{(n)}, \tag{5.117}$$

to obtain an ensemble $\mathcal{E}'^{(n)}$ of output messages.

Now suppose that this coding scheme has high fidelity. To minimize technicalities, let us not specify in detail how the fidelity of $\mathcal{E}'^{(n)}$ relative to $\mathcal{E}^{(n)}$ should be quantified. Let us just accept that if $\mathcal{E}'^{(n)}$ has high fidelity, then for any δ and n sufficiently large

$$\frac{1}{n}\chi(\mathcal{E}^{(n)}) - \delta \leq \frac{1}{n}\chi(\mathcal{E}'^{(n)}) \leq \frac{1}{n}\chi(\mathcal{E}^{(n)}) + \delta; \tag{5.118}$$

the Holevo information per letter of the output approaches that of the input. Since the input messages are product states, it follows from the additivity of $S(\boldsymbol{\rho})$ that

$$\chi(\mathcal{E}^{(n)}) = n\chi(\mathcal{E}), \tag{5.119}$$

and we also know from Lindblad–Uhlmann monotonicity that

$$\chi(\mathcal{E}'^{(n)}) \leq \chi(\tilde{\mathcal{E}}^{(n)}). \tag{5.120}$$

By combining eqs. (5.118)-(5.120), we find that

$$\frac{1}{n}\chi(\tilde{\mathcal{E}}^{(n)}) \geq \chi(\mathcal{E}) - \delta. \tag{5.121}$$

Finally, $\chi(\tilde{\mathcal{E}}^{(n)})$ is bounded above by $S(\tilde{\boldsymbol{\rho}}^{(n)})$, which is in turn bounded above by $\log \dim \tilde{\mathcal{H}}^{(n)}$. Since δ may be as small as we please, we conclude that, asymptotically as $n \to \infty$,

$$\frac{1}{n}\log(\dim \tilde{\mathcal{H}}^{(n)}) \geq \chi(\mathcal{E}); \tag{5.122}$$

high-fidelity compression to fewer than $\chi(\mathcal{E})$ qubits per letter is not possible.

One is sorely tempted to conjecture that compression to $\chi(\mathcal{E})$ qubits per letter is asymptotically attainable. As of mid-January, 1998, this conjecture still awaits proof or refutation.

5.4 Accessible Information

The close analogy between the Holevo information $\chi(\mathcal{E})$ and the classical mutual information $I(X;Y)$, as well as the monotonicity of χ, suggest that χ is related to the amount of *classical* information that can be stored in and recovered from a quantum system. In this section, we will make this connection precise.

The previous section was devoted to quantifying the *quantum* information content – measured in *qubits* – of messages constructed from an alphabet of quantum states. But now we will turn to a quite different topic. We want to quantify the *classical* information content – measured in bits – that can be extracted from such messages, particularly in the case where the alphabet includes letters that are not mutually orthogonal.

Now, why would we be so foolish as to store classical information in nonorthogonal quantum states that cannot be perfectly distinguished? Storing information this way should surely be avoided as it will degrade the classical signal. But perhaps we can't help it. For example, maybe I am a communications engineer, and I am interested in the intrinsic physical limitations on the classical capacity of a high bandwidth optical fiber. Clearly, to achieve a higher throughout of classical information per unit power, we should choose to encode information in single photons, and to attain a high rate, we should increase the number of photons transmitted per second. But if we squeeze photon wavepackets together tightly, the wavepackets will overlap, and so will not be perfectly distinguishable. How do we maximize the classical information transmitted in that case? As another important example, maybe I am an experimental physicist, and I want to use a delicate quantum system to construct a very sensitive instrument that measures a classical force acting on the system. We can model the force as a free parameter x in the system's Hamiltonian $\boldsymbol{H}(x)$. Depending on the value of x, the state of the system will evolve to various possible final (nonorthogonal) states $\boldsymbol{\rho}_x$. How much information about x can our apparatus acquire?

While physically this is a much different issue than the compressibility of quantum information, mathematically the two questions are related. We will find that the Von Neumann entropy and its generalization the Holevo information will play a central role in the discussion.

Suppose, for example, that Alice prepares a pure quantum state drawn from the ensemble $\mathcal{E} = \{|\varphi_x\rangle, p_x\}$. Bob knows the ensemble, but not the particular state that Alice chose. He wants to acquire as much information as possible about x.

Bob collects his information by performing a generalized measurement, the POVM $\{\boldsymbol{F}_y\}$. If Alice chose preparation x, Bob will obtain the measurement outcome y with conditional probability

$$p(y|x) = \langle\varphi_x|\boldsymbol{F}_y|\varphi_x\rangle. \tag{5.123}$$

These conditional probabilities, together with the ensemble X, determine the amount of information that Bob gains on the average, the mutual information $I(X;Y)$ of preparation and measurement outcome.

Bob is free to perform the measurement of his choice. The "best" possible measurement, that which maximizes his information gain, is called the *optimal measurement* determined by the ensemble. The maximal information gain is

$$\mathrm{Acc}(\mathcal{E}) = \underset{\{\boldsymbol{F}_y\}}{\mathrm{Max}}\, I(X;Y), \tag{5.124}$$

where the Max is over all POVM's. This quantity is called the *accessible information* of the ensemble $\mathcal{E}$.

Of course, if the states $|\varphi_x\rangle$ are mutually orthogonal, then they are perfectly distinguishable. The orthogonal measurement

$$\boldsymbol{E}_y = |\varphi_y\rangle\langle\varphi_y|, \tag{5.125}$$

has conditional probability

$$p(y|x) = \delta_{y,x}, \tag{5.126}$$

so that $H(X|Y) = 0$ and $I(X;Y) = H(X)$. This measurement is clearly optimal – the preparation is completely determined – so that

$$\mathrm{Acc}(\mathcal{E}) = H(X), \tag{5.127}$$

for an ensemble of mutually orthogonal (pure *or* mixed) states.

But the problem is much more interesting when the signal states are nonorthogonal pure states. In this case, no useful general formula for $\mathrm{Acc}(\mathcal{E})$ is known, but there is an upper bound

$$\mathrm{Acc}(\mathcal{E}) \le S(\boldsymbol{\rho}). \tag{5.128}$$

We have seen that this bound is saturated in the case of orthogonal signal states, where $S(\boldsymbol{\rho}) = H(X)$. In general, we know from classical information theory that $I(X;Y) \le H(X)$; but for nonorthogonal states we have $S(\boldsymbol{\rho}) <$

$H(X)$, so that eq. (5.128) is a better bound. Even so, this bound is not tight; in many cases $\mathrm{Acc}(\mathcal{E})$ is strictly less than $S(\boldsymbol{\rho})$.

We obtain a sharper relation between $\mathrm{Acc}(\mathcal{E})$ and $S(\boldsymbol{\rho})$ if we consider the accessible information per letter in a message containing n letters. Now Bob has more flexibility – he can choose to perform a collective measurement on all n letters, and thereby collect more information than if he were restricted to measuring only one letter at a time. Furthermore, Alice can choose to prepare, rather than arbitrary messages with each letter drawn from the ensemble $\mathcal{E}$, an ensemble of special messages (a code) designed to be maximally distinguishable.

We will then see that Alice and Bob can find a code such that the marginal ensemble for each letter is $\mathcal{E}$, and the accessible information per letter asymptotically approaches $S(\boldsymbol{\rho})$ as $n \to \infty$. In this sense, $S(\boldsymbol{\rho})$ characterizes the accessible information of an ensemble of *pure* quantum states.

Furthermore, these results generalize to ensembles of mixed quantum states, with the Holevo information replacing the Von Neumann entropy. The accessible information of an ensemble of mixed states $\{\boldsymbol{\rho}_x, p_x\}$ satisfies

$$\mathrm{Acc}(\mathcal{E}) \leq \chi(\mathcal{E}), \tag{5.129}$$

a result known as the *Holevo bound*. This bound is not tight in general (though it is saturated for ensembles of mutually orthogonal mixed states). However, if Alice and Bob choose an n-letter code, where the marginal ensemble for each letter is $\mathcal{E}$, and Bob performs an optimal POVM on all n letters collectively, then the best attainable accessible information per letter is $\chi(\mathcal{E})$ – *if* all code words are required to be *product* states. In this sense, $\chi(\mathcal{E})$ characterizes the accessible information of an ensemble of *mixed* quantum states.

One way that an alphabet of mixed quantum states might arise is that Alice might try to send pure quantum states to Bob through a noisy quantum channel. Due to decoherence in the channel, Bob receives mixed states that he must decode. In this case, then, $\chi(\mathcal{E})$ characterizes the maximal amount of classical information that can be transmitted to Bob through the noisy quantum channel.

For example, Alice might send to Bob n photons in certain polarization states. If we suppose that the noise acts on each photon independently, and that Alice sends unentangled states of the photons, then $\chi(\mathcal{E})$ is the maximal

amount of information that Bob can acquire per photon. Since

$$\chi(\mathcal{E}) \leq S(\boldsymbol{\rho}) \leq 1, \tag{5.130}$$

it follows in particular that a single (unentangled) photon can carry at most one bit of classical information.

5.4.1 The Holevo Bound

The Holevo bound on the accessible information is not an easy theorem, but like many good things in quantum information theory, it follows easily once the strong subadditivity of Von Neumann entropy is established. Here we will assume strong subadditivity and show that the Holevo bound follows.

Recall the setting: Alice prepares a quantum state drawn from the ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$, and then Bob performs the POVM $\{\boldsymbol{F}_y\}$. The joint probability distribution governing Alice's preparation x and Bob's outcome y is

$$p(x, y) = p_x tr\{\boldsymbol{F}_y \boldsymbol{\rho}_x\}. \tag{5.131}$$

We want to show that

$$I(X;Y) \leq \chi(\mathcal{E}). \tag{5.132}$$

Since strong subadditivity is a property of three subsystems, we will need to identify three systems to apply it to. Our strategy will be to prepare an input system X that stores a classical record of what preparation was chosen and an output system Y whose classical correlations with x are governed by the joint probability distribution $p(x, y)$. Then applying strong subadditivity to X, Y, and our quantum system Q, we will be able to relate $I(X;Y)$ to $\chi(\mathcal{E})$.

Suppose that the initial state of the system XQY is

$$\boldsymbol{\rho}_{XQY} = \sum_x p_x |x\rangle\langle x| \otimes \boldsymbol{\rho}_x \otimes |0\rangle\langle 0|, \tag{5.133}$$

where the $|x\rangle$'s are mutually orthogonal pure states of the input system X, and $|0\rangle$ is a particular pure state of the output system Y. By performing partial traces, we see that

$$\begin{aligned} \boldsymbol{\rho}_X &= \sum_x p_x |x\rangle\langle x| \to S(\boldsymbol{\rho}_X) = H(X) \\ \boldsymbol{\rho}_Q &= \sum_x p_x \boldsymbol{\rho}_x \equiv \boldsymbol{\rho} \to S(\boldsymbol{\rho}_{QY}) = S(\boldsymbol{\rho}_Q) = S(\boldsymbol{\rho}). \end{aligned} \tag{5.134}$$

and since the $|x\rangle$'s are mutually orthogonal, we also have

$$\begin{aligned} S(\boldsymbol{\rho}_{XQY}) = S(\boldsymbol{\rho}_{XQ}) &= \sum_x -\mathrm{tr}(p_x\boldsymbol{\rho}_x \log p_x\boldsymbol{\rho}_x) \\ &= H(X) + \sum_x p_x S(\boldsymbol{\rho}_x). \end{aligned} \quad (5.135)$$

Now we will perform a unitary transformation that "imprints" Bob's measurement result in the output system Y. Let us suppose, for now, that Bob performs an orthogonal measurement $\{\boldsymbol{E}_y\}$, where

$$\boldsymbol{E}_y\boldsymbol{E}_{y'} = \delta_{y,y'}\boldsymbol{E}_y, \quad (5.136)$$

(we'll consider more general POVM's shortly). Our unitary transformation $\boldsymbol{U}_{QY}$ acts on QY according to

$$\boldsymbol{U}_{QY} : |\varphi\rangle_Q \otimes |0\rangle_Y = \sum_y \boldsymbol{E}_y|\varphi\rangle_Q \otimes |y\rangle_Y, \quad (5.137)$$

(where the $|y\rangle$'s are mutually orthogonal), and so transforms $\boldsymbol{\rho}_{XQY}$ as

$$\boldsymbol{U}_{QY} : \boldsymbol{\rho}_{XQY} \to \boldsymbol{\rho}'_{XQY} = \sum_{x,y,y'} p_x|x\rangle\langle x| \otimes \boldsymbol{E}_y\boldsymbol{\rho}_x\boldsymbol{E}_{y'} \otimes |y\rangle\langle y'|. \quad (5.138)$$

Since Von Neumann entropy is invariant under a unitary change of basis, we have

$$\begin{aligned} S(\boldsymbol{\rho}'_{XQY}) &= S(\boldsymbol{\rho}_{XQY}) = H(x) + \sum_x p_x S(\boldsymbol{\rho}_x), \\ S(\boldsymbol{\rho}'_{QY}) &= S(\boldsymbol{\rho}_{QY}) = S(\boldsymbol{\rho}), \end{aligned} \quad (5.139)$$

and taking a partial trace of eq. (5.138) we find

$$\begin{aligned} \boldsymbol{\rho}'_{XY} &= \sum_{x,y} p_x \mathrm{tr}(\boldsymbol{E}_y\boldsymbol{\rho}_x)|x\rangle\langle x| \otimes |y\rangle\langle y| \\ &= \sum_{x,y} p(x,y)|x,y\rangle\langle x,y| \to S(\boldsymbol{\rho}'_{XY}) = H(X,Y), \end{aligned} \quad (5.140)$$

(using eq. (5.136). Evidently it follows that

$$\boldsymbol{\rho}'_Y = \sum_y p(y)|y\rangle\langle y| \to S(\boldsymbol{\rho}'_Y) = H(Y). \quad (5.141)$$

Now we invoke strong subadditivity in the form

$$S(\boldsymbol{\rho}'_{XQY}) + S(\boldsymbol{\rho}'_Y) \le S(\boldsymbol{\rho}'_{XY}) + S(\boldsymbol{\rho}'_{QY}), \quad (5.142)$$

which becomes

$$H(X) + \sum_x p_x S(\boldsymbol{\rho}_x) + H(Y) \le H(X,Y) + S(\boldsymbol{\rho}), \quad (5.143)$$

or

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \le S(\boldsymbol{\rho}) - \sum_x p_x S(\boldsymbol{\rho}_x) = \chi(\mathcal{E}). \quad (5.144)$$

This is the Holevo bound.

One way to treat more general POVM's is to enlarge the system by appending one more subsystem Z. We then construct a unitary $\boldsymbol{U}_{QYZ}$ acting as

$$\boldsymbol{U}_{QYZ} : |\varphi\rangle_Q \otimes |0\rangle_Y \otimes |0\rangle_Z = \sum_y \sqrt{\boldsymbol{F}_y}|\varphi\rangle_A \otimes |y\rangle_Y \otimes |y\rangle_Z, \quad (5.145)$$

so that

$$\boldsymbol{\rho}'_{XQYZ} = \sum_{x,y,y'} p_x|x\rangle\langle x| \otimes \sqrt{\boldsymbol{F}_y}\boldsymbol{\rho}_x\sqrt{\boldsymbol{F}_{y'}} \otimes |y\rangle\langle y'| \otimes |y\rangle\langle y'|. \quad (5.146)$$

Then the partial trace over Z yields

$$\boldsymbol{\rho}'_{XQY} = \sum_{x,y} p_x|x\rangle\langle x| \otimes \sqrt{\boldsymbol{F}_y}\boldsymbol{\rho}_x\sqrt{\boldsymbol{F}_y} \otimes |y\rangle\langle y|, \quad (5.147)$$

and

$$\begin{aligned} \boldsymbol{\rho}'_{XY} &= \sum_{x,y} p_x tr(\boldsymbol{F}_y\boldsymbol{\rho}_x)|x\rangle\langle x| \otimes |y\rangle\langle y| \\ &= \sum_{x,y} p(x,y)|x,y\rangle\langle x,y| \\ &\to S(\boldsymbol{\rho}'_{XY}) = H(X,Y). \end{aligned} \quad (5.148)$$

The rest of the argument then runs as before.

5.4.2 Improving distinguishability: the Peres–Wootters method

To better acquaint ourselves with the concept of accessible information, let's consider a single-qubit example. Alice prepares one of the three possible pure states

$$|\varphi_1\rangle = |\uparrow_{\hat{n}_1}\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$
$$|\varphi_2\rangle = |\uparrow_{\hat{n}_2}\rangle = \begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix},$$
$$|\varphi_3\rangle = |\uparrow_{\hat{n}_3}\rangle = \begin{pmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix}; \tag{5.149}$$

a spin-$\frac{1}{2}$ object points in one of three directions that are symmetrically distributed in the xz-plane. Each state has *a priori* probability $\frac{1}{3}$. Evidently, Alice's "signal states" are nonorthogonal:

$$\langle\varphi_1|\varphi_2\rangle = \langle\varphi_1|\varphi_3\rangle = \langle\varphi_2|\varphi_3\rangle = -\frac{1}{2}. \tag{5.150}$$

Bob's task is to find out as much as he can about what Alice prepared by making a suitable measurement. The density matrix of Alice's ensemble is

$$\boldsymbol{\rho} = \frac{1}{3}(|\varphi_1\rangle\langle\varphi_1| + |\varphi_2\rangle\langle\varphi_3| + |\varphi_3\rangle\langle\varphi_3|) = \frac{1}{2}\mathbf{1}, \tag{5.151}$$

which has $S(\boldsymbol{\rho}) = 1$. Therefore, the Holevo bound tells us that the mutual information of Alice's preparation and Bob's measurement outcome cannot exceed 1 bit.

In fact, though, the accessible information is considerably less than the one bit allowed by the Holevo bound. In this case, Alice's ensemble has enough symmetry that it is not hard to guess the optimal measurement. Bob may choose a POVM with three outcomes, where

$$\boldsymbol{F}_{\bar{a}} = \frac{2}{3}(\mathbf{1} - |\varphi_a\rangle\langle\varphi_a|), \quad a = 1,2,3; \tag{5.152}$$

we see that

$$p(a|b) = \langle\varphi_b|\boldsymbol{F}_{\bar{a}}|\varphi_b\rangle = \begin{cases} 0 & a = b, \\ \frac{1}{2} & a \neq b. \end{cases} \tag{5.153}$$

Therefore, the measurement outcome a *excludes* the possibility that Alice prepared a, but leaves equal *a posteriori* probabilities $\left(p = \frac{1}{2}\right)$ for the other two states. Bob's information gain is

$$I = H(X) - H(X|Y) = \log_2 3 - 1 = .58496. \tag{5.154}$$

To show that this measurement is really optimal, we may appeal to a variation on a theorem of Davies, which assures us that an optimal POVM can be chosen with three $\boldsymbol{F}_a$'s that share the same three-fold symmetry as the three states in the input ensemble. This result restricts the possible POVM's enough so that we can check that eq. (5.152) is optimal with an explicit calculation. Hence we have found that the ensemble $\mathcal{E} = \{|\varphi_a\rangle, p_a = \frac{1}{3}\}$ has accessible information.

$$\mathrm{Acc}(\mathcal{E}) = \log_2\left(\frac{3}{2}\right) = .58496... \tag{5.155}$$

The Holevo bound is not saturated.

Now suppose that Alice has enough cash so that she can afford to send two qubits to Bob, where again each qubit is drawn from the ensemble $\mathcal{E}$. The obvious thing for Alice to do is prepare one of the *nine* states

$$|\varphi_a\rangle|\varphi_b\rangle, \quad a,b = 1,2,3, \tag{5.156}$$

each with $p_{ab} = 1/9$. Then Bob's best strategy is to perform the POVM eq. (5.152) on each of the two qubits, achieving a mutual information of .58496 bits per qubit, as before.

But Alice and Bob are determined to do better. After discussing the problem with A. Peres and W. Wootters, they decide on a different strategy. Alice will prepare one of *three* two-qubit states

$$|\Phi_a\rangle = |\varphi_a\rangle|\varphi_a\rangle, \quad a = 1,2,3, \tag{5.157}$$

each occurring with *a priori* probability $p_a = 1/2$. Considered one-qubit at a time, Alice's choice is governed by the ensemble $\mathcal{E}$, but now her two qubits have (classical) correlations – both are prepared the same way.

The three $|\Phi_a\rangle$'s are linearly independent, and so span a three-dimensional subspace of the four-dimensional two-qubit Hilbert space. In a homework exercise, you will show that the density matrix

$$\boldsymbol{\rho} = \frac{1}{3}\left(\sum_{a=1}^{3}|\Phi_a\rangle\langle\Phi_a|\right), \tag{5.158}$$

has the nonzero eigenvalues $1/2, 1/4, 1/4$, so that

$$S(\boldsymbol{\rho}) = -\frac{1}{2}\log\frac{1}{2} - 2\left(\frac{1}{4}\log\frac{1}{4}\right) = \frac{3}{2}. \tag{5.159}$$

The Holevo bound requires that the accessible information *per qubit* is less than 3/4 bit. This would at least be consistent with the possibility that we can exceed the .58496 bits per qubit attained by the nine-state method.

Naively, it may seem that Alice won't be able to convey as much classical information to Bob, if she chooses to send one of only three possible states instead of nine. But on further reflection, this conclusion is not obvious. True, Alice has fewer signals to choose from, but the signals are *more distinguishable;* we have

$$\langle\Phi_a|\Phi_b\rangle = \frac{1}{4}, \quad a \neq b, \tag{5.160}$$

instead of eq. (5.150). It is up to Bob to exploit this improved distinguishability in his choice of measurement. In particular, Bob will find it advantageous to perform *collective* measurements on the two qubits instead of measuring them one at a time.

It is no longer obvious what Bob's optimal measurement will be. But Bob can invoke a general procedure that, while not guaranteed optimal, is usually at least pretty good. We'll call the POVM constructed by this procedure a "pretty good measurement" (or PGM).

Consider some collection of vectors $|\tilde{\Phi}_a\rangle$ that are not assumed to be orthogonal or normalized. We want to devise a POVM that can distinguish these vectors reasonably well. Let us first construct

$$\boldsymbol{G} = \sum_a |\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|; \tag{5.161}$$

This is a positive operator on the space spanned by the $|\tilde{\Phi}_a\rangle$'s. Therefore, on that subspace, $\boldsymbol{G}$ has an inverse, $\boldsymbol{G}^{-1}$ and that inverse has a positive square root $\boldsymbol{G}^{-1/2}$. Now we define

$$\boldsymbol{F}_a = \boldsymbol{G}^{-1/2}|\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|\boldsymbol{G}^{-1/2}, \tag{5.162}$$

and we see that

$$\begin{aligned}\sum_a \boldsymbol{F}_a &= \boldsymbol{G}^{-1/2}\left(\sum_a |\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|\right)\boldsymbol{G}^{-1/2} \\ &= \boldsymbol{G}^{-1/2}\boldsymbol{G}\boldsymbol{G}^{-1/2} = \boldsymbol{1},\end{aligned} \tag{5.163}$$

on the span of the $|\tilde{\Phi}_a\rangle$'s. If necessary, we can augment these $\boldsymbol{F}_a$'s with one more positive operator, the projection $\boldsymbol{F}_0$ onto the orthogonal complement of the span of the $|\tilde{\Phi}_a\rangle$'s, and so construct a POVM. This POVM is the PGM associated with the vectors $|\tilde{\Phi}_a\rangle$.

In the special case where the $|\tilde{\Phi}_a\rangle$'s are orthogonal,

$$|\tilde{\Phi}_a\rangle = \sqrt{\lambda_a}|\Phi_a\rangle, \tag{5.164}$$

(where the $|\Phi_a\rangle$'s are orthonormal), we have

$$\begin{aligned}\boldsymbol{F}_a &= \sum_{a,b,c}(|\Phi_b\rangle\lambda_b^{-1/2}\langle\Phi_b|)(\lambda_a|\Phi_a\rangle\langle\Phi_a|)(|\Phi_c\rangle\lambda_c^{-1/2}\langle\Phi_c|) \\ &= |\Phi_a\rangle\langle\Phi_a|;\end{aligned} \tag{5.165}$$

this is the orthogonal measurement that perfectly distinguishes the $|\Phi_a\rangle$'s and so clearly is optimal. If the $|\tilde{\Phi}_a\rangle$'s are linearly independent but not orthogonal, then the PGM is again an orthogonal measurement (because n one-dimensional operators in an n-dimensional space can constitute a POVM only if mutually orthogonal), but in that case the measurement may not be optimal.

In the homework, you'll construct the PGM for the vectors $|\Phi_a\rangle$ in eq. (5.157), and you'll show that

$$\begin{aligned}p(a|a) &= \langle\Phi_a|\boldsymbol{F}_a|\Phi_a\rangle = \frac{1}{3}\left(1+\frac{1}{\sqrt{2}}\right)^2 = .971405 \\ p(b|a) &= \langle\Phi_a|\boldsymbol{F}_b|\Phi_a\rangle = \frac{1}{6}\left(1-\frac{1}{\sqrt{2}}\right)^2 = .0142977,\end{aligned} \tag{5.166}$$

(for $b \neq a$). It follows that the conditional entropy of the input is

$$H(X|Y) = .215893, \tag{5.167}$$

and since $H(X) = \log_2 3 = 1.58496$, the information gain is

$$I = H(X) - H(X|Y) = 1.36907, \tag{5.168}$$

a mutual information of .684535 bits per qubit. Thus, the improved distinguishability of Alice's signals has indeed paid off – we have exceeded the

.58496 bits that can be extracted from a single qubit. We still didn't saturate the Holevo bound ($I < 1.5$ in this case), but we came a lot closer than before.

This example, first described by Peres and Wootters, teaches some useful lessons. First, Alice is able to convey more information to Bob by "pruning" her set of codewords. She is better off choosing among fewer signals that are more distinguishable than more signals that are less distinguishable. An alphabet of three letters encodes more than an alphabet of nine letters.

Second, Bob is able to read more of the information if he performs a collective measurement instead of measuring each qubit separately. His optimal orthogonal measurement projects Alice's signal onto a basis of *entangled* states.

The PGM described here is "optimal" in the sense that it gives the best information gain of any *known* measurement. Most likely, this is really the highest I that can be achieved with *any* measurement, but I have not proved it.

5.4.3 Attaining Holevo: pure states

With these lessons in mind, we can proceed to show that, given an ensemble of pure states, we can construct n-letter codewords that asymptotically attain an accessible information of $S(\boldsymbol{\rho})$ per letter.

We must select a code, the ensemble of codewords that Alice can prepare, and a "decoding observable," the POVM that Bob will use to try to distinguish the codewords. Our task is to show that Alice can choose $2^{n(S-\delta)}$ codewords, such that Bob can determine which one was sent, with negligible probability of error as $n \to \infty$. We won't go through all the details of the argument, but will be content to understand why the result is highly plausible.

The main idea, of course, is to invoke random coding. Alice chooses product signal states

$$|\varphi_{x_1}\rangle|\varphi_{x_2}\rangle \ldots |\varphi_{x_n}\rangle, \tag{5.169}$$

by drawing each letter at random from the ensemble $\mathcal{E} = \{|\varphi_x\rangle, p_x\}$. As we have seen, for a typical code each typical codeword has a large overlap with a typical subspace $\Lambda^{(n)}$ that has dimension $\dim \Lambda^{(n)} > 2^{n(S(\boldsymbol{\rho})-\delta)}$. Furthermore, for a typical code, the marginal ensemble governing each letter is close to $\mathcal{E}$.

Because the typical subspace is very large for n large, Alice can choose many codewords, yet be assured that the typical overlap of two typical code-

words is very small. Heuristically, the typical codewords are randomly distributed in the typical subspace, and on average, two random unit vectors in a space of dimension D have overlap $1/D$. Therefore if $|u\rangle$ and $|w\rangle$ are two codewords

$$\langle|\langle u|w\rangle|^2\rangle_\Lambda < 2^{-n(S-\delta)}. \tag{5.170}$$

Here $< \cdot >_\Lambda$ denotes an average over random typical codewords.

You can convince yourself that the typical codewords really are uniformly distributed in the typical subspace as follows: Averaged over the ensemble, the overlap of random codewords $|\varphi_{x_1}\rangle \ldots |\varphi_{x_n}\rangle$ and $|\varphi_{y_1}\rangle \ldots |\varphi_{y_n}\rangle$ is

$$\begin{aligned} &= \sum p_{x_1} \ldots p_{x_n} p_{y_1} \ldots p_{y_n} (|\langle\varphi_{x_1}|\varphi_{y_1}\rangle|^2 \ldots |\langle\varphi_{x_n}|\varphi_{y_n}\rangle|^2) \\ &= \mathrm{tr}(\boldsymbol{\rho} \otimes \ldots \otimes \boldsymbol{\rho})^2. \end{aligned} \tag{5.171}$$

Now suppose we restrict the trace to the typical subspace $\Lambda^{(n)}$; this space has $\dim \Lambda^{(n)} < 2^{n(S+\delta)}$ and the eigenvalues of $\boldsymbol{\rho}^{(n)} = \boldsymbol{\rho} \otimes \ldots \otimes \boldsymbol{\rho}$ restricted to $\Lambda^{(n)}$ satisfy $\lambda < 2^{-n(S-\delta)}$. Therefore

$$\langle|\langle u|w\rangle|^2\rangle_\Lambda = \mathrm{tr}_\Lambda[\boldsymbol{\rho}^{(n)}]^2 < 2^{n(S+\delta)}[2^{-n(S-\delta)}]^2 = 2^{-n(S-3\delta)}, \tag{5.172}$$

where tr_Λ denotes the trace in the typical subspace.

Now suppose that $2^{n(S-\delta)}$ random codewords $\{|u_i\rangle\}$ are selected. Then if $|u_j\rangle$ is any fixed codeword

$$\sum_{i\neq j}\langle|\langle u_i|u_j\rangle|^2\rangle < 2^{n(S-\delta)}2^{-n(S-\delta')} + \varepsilon = 2^{-n(\delta-\delta')} + \varepsilon; \tag{5.173}$$

here the sum is over all codewords, and the average is no longer restricted to the typical codewords – the ε on the right-hand side arises from the atypical case. Now for any fixed δ, we can choose δ' and ε as small as we please for n sufficiently large; we conclude that when we average over both codes and codewords within a code, the codewords become highly distinguishable as $n \to \infty$.

Now we invoke some standard Shannonisms: Since eq. (5.173) holds when we average over codes, it also holds for a particular code. (Furthermore, since nearly all codes have the property that the marginal ensemble for each letter is close to $\mathcal{E}$, there is a code with this property satisfying eq. (5.173).) Now eq. (5.173) holds when we average over the particular codeword $|u_j\rangle$. But by throwing away at most half of the codewords, we can ensure that each and every codeword is highly distinguishable from all the others.

We see that Alice can choose $2^{n(S-\delta)}$ highly distinguishable codewords, which become mutually orthogonal as $n \to \infty$. Bob can perform a PGM at finite n that approaches an optimal orthogonal measurement as $n \to \infty$. Therefore the accessible information per letter

$$\frac{1}{n}\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) = S(\boldsymbol{\rho}) - \delta, \tag{5.174}$$

is attainable, where $\tilde{\mathcal{E}}^{(n)}$ denotes Alice's ensemble of n-letter codewords.

Of course, for any finite n, Bob's POVM will be a complicated collective measurement performed on all n letters. To give an honest proof of attainability, we should analyze the POVM carefully, and bound its probability of error. This has been done by Hausladen, *et al.*[4] The handwaving argument here at least indicates why their conclusion is not surprising.

It also follows from the Holevo bound and the subadditivity of the entropy that the accessible information per letter cannot exceed $S(\boldsymbol{\rho})$ asymptotically. The Holevo bound tells us that

$$\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) \leq S(\tilde{\boldsymbol{\rho}}^{(n)}), \tag{5.175}$$

where $\tilde{\boldsymbol{\rho}}^{(n)}$ denotes the density matrix of the codewords, and subadditivity implies that

$$S(\tilde{\boldsymbol{\rho}}^{(n)}) \leq \sum_{i=1}^{n} S(\tilde{\boldsymbol{\rho}}_i), \tag{5.176}$$

where $\tilde{\boldsymbol{\rho}}_i$ is the reduced density matrix of the ith letter. Since each $\tilde{\boldsymbol{\rho}}_i$ approaches $\boldsymbol{\rho}$ asymptotically, we have

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) \leq \lim_{n\to\infty} \frac{1}{n} S(\tilde{\boldsymbol{\rho}}^{(n)}) \leq S(\boldsymbol{\rho}). \tag{5.177}$$

To derive this bound, we did not assume anything about the code, except that the marginal ensemble for each letter asymptotically approaches $\mathcal{E}$. In

[4] P. Hausladen, R. Jozsa, B. Schumacher, M. Westmoreland, and W. K. Wootters, "Classical information capacity of a quantum channel," *Phys. Rev.* A **54** (1996) 1869-1876.

particular the bound applies even if the codewords are entangled states rather than product states. Therefore we have shown that $S(\boldsymbol{\rho})$ is the optimal accessible information per letter.

We can define a kind of channel capacity associated with a specified alphabet of pure quantum states, the "fixed-alphabet capacity." We suppose that Alice is equipped with a source of quantum states. She can produce any one of the states $|\varphi_x\rangle$, but it is up to her to choose the *a priori* probabilities of these states. The fixed-alphabet capacity C_{f_a} is the maximum accessible information per letter she can achieve with the best possible distribution $\{p_x\}$. We have found that

$$C_{f_a} = \underset{\{p_x\}}{\mathrm{Max}}\, S(\boldsymbol{\rho}). \tag{5.178}$$

C_{f_a} is the optimal number of classical bits we can encode per letter (asymptotically), given the specified quantum-state alphabet of the source.

5.4.4 Attaining Holevo: mixed states

Now we would like to extend the above reasoning to a more general context. We will consider n-letter messages, where the marginal ensemble for each letter is the ensemble of *mixed* quantum states

$$\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}. \tag{5.179}$$

We want to argue that it is possible (asymptotically as $n \to \infty$) to convey $\chi(\mathcal{E})$ bits of classical information per letter. Again, our task is to: (1) specify a code that Alice and Bob can use, where the ensemble of codewords yields the ensemble $\mathcal{E}$ letter by letter (at least asymptotically). (2) Specify Bob's decoding observable, the POVM he will use to attempt to distinguish the codewords. (3) Show that Bob's probability of error approaches zero as $n \to \infty$. As in our discussion of the pure-state case, I will not exhibit the complete proof (see Holevo[5] and Schumacher and Westmoreland[6]). Instead, I'll offer an argument (with even more handwaving than before, if that's possible) indicating that the conclusion is reasonable.

[5] A.S. Holevo, "The Capacity of the Quantum Channel with General Signal States," quant-ph/9611023

[6] B. Schumacher and M.D. Westmoreland, "Sending Classical Information Via Noisy Quantum Channels," *Phys. Rev.* **A 56** (1997) 131-138.

As always, we will demonstrate attainability by a random coding argument. Alice will select mixed-state codewords, with each letter drawn from the ensemble $\mathcal{E}$. That is, the codeword

$$\boldsymbol{\rho}_{x_1} \otimes \boldsymbol{\rho}_{x_2} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}, \tag{5.180}$$

is chosen with probability $p_{x_1} p_{x_2} \ldots p_{x_n}$. The idea is that *each* typical codeword can be regarded as an ensemble of pure states, with nearly all of its support on a certain typical subspace. If the typical subspaces of the various codewords have little overlap, then Bob will be able to perform a POVM that identifies the typical subspace characteristic of Alice's message, with small probability of error.

What is the dimension of the typical subspace of a typical codeword? If we *average* over the codewords, the mean entropy of a codeword is

$$\langle S^{(n)} \rangle = \sum_{x_1 \ldots x_n} p_{x_1} p_{x_2} \ldots p_{x_n} S(\boldsymbol{\rho}_{x_1} \otimes \boldsymbol{\rho}_{x_2} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}). \tag{5.181}$$

Using additivity of the entropy of a product state, and $\Sigma_x \, p_x = 1$, we obtain

$$\langle S^{(n)} \rangle = n \sum_x p_x S(\boldsymbol{\rho}_x) \equiv n \langle S \rangle. \tag{5.182}$$

For n large, the entropy of a codeword is, with high probability, close to this mean, and furthermore, the high probability eigenvalues of $\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_2}$ are close to $2^{-n\langle S \rangle}$. In other words a typical $\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}$ has its support on a typical subspace of dimension $2^{n\langle S \rangle}$.

This statement is closely analogous to the observation (crucial to the proof of Shannon's noisy channel coding theorem) that the number of typical messages received when a typical message is sent through a noisy classical channel is $2^{nH(Y|X)}$.

Now the argument follows a familiar road. For each typical message $x_1 x_2 \ldots x_n$, Bob can construct a "decoding subspace" of dimension $2^{n(\langle S \rangle + \delta)}$, with assurance that Alice's message is highly likely to have nearly all its support on this subspace. His POVM will be designed to determine in which decoding subspace Alice's message lies. Decoding errors will be unlikely if typical decoding subspaces have little overlap.

Although Bob is really interested only in the value of the decoding subspace (and hence $x_1 x_2 \ldots x_n$), let us suppose that he performs the complete PGM determined by all the vectors that span all the typical subspaces of Alice's codewords. (And this PGM will approach an orthogonal measurement for large n, as long as the number of codewords is not too large.) He obtains a particular result which is likely to be in the typical subspace of dimension $2^{nS(\boldsymbol{\rho})}$ determined by the source $\boldsymbol{\rho} \otimes \boldsymbol{\rho} \otimes \ldots \otimes \boldsymbol{\rho}$, and furthermore, is likely to be in the decoding subspace of the message that Alice actually sent. Since Bob's measurement results are uniformly distributed in a space on dimension 2^{nS}, and the pure-state ensemble determined by a particular decoding subspace has dimension $2^{n(\langle S \rangle + \delta)}$, the average overlap of the vector determined by Bob's result with a typical decoding subspace is

$$\frac{2^{n(\langle S \rangle + \delta)}}{2^{nS}} = 2^{-n(S - \langle S \rangle - \delta)} = 2^{-n(\chi - \delta)}. \tag{5.183}$$

If Alice chooses 2^{nR} codewords, the average probability of a decoding error will be

$$2^{nR} 2^{-n(\chi - \delta)} = 2^{-n(\chi - R - \delta)}. \tag{5.184}$$

We can choose any R less than χ, and this error probability will get very small as $n \to \infty$.

This argument shows that the probability of error is small, averaged over both random codes and codewords. As usual, we can choose a particular code, and throw away some codewords to achieve a small probability of error for every codeword. Furthermore, the particular code may be chosen to be typical, so that the marginal ensemble for each codeword approaches $\mathcal{E}$ as $n \to \infty$. We conclude that an accessible information of χ per letter is asymptotically attainable.

The structure of the argument closely follows that for the corresponding classical coding theorem. In particular, the quantity χ arose much as I does in Shannon's theorem. While 2^{-nI} is the probability that a particular typical sequence lies in a specified decoding sphere, $2^{-n\chi}$ is the overlap of a particular typical state with a specified decoding subspace.

5.4.5 Channel capacity

Combining the Holevo bound with the conclusion that χ bits per letter is attainable, we obtain an expression for the *classical* capacity of a quantum channel (But with a caveat: we are assured that this "capacity" cannot be exceeded only if we disallow entangled codewords.)

Alice will prepare n-letter messages and send them through a noisy quantum channel to Bob. The channel is described by a superoperator, and we will assume that the same superoperator \$ acts on each letter independently (*memoryless* quantum channel). Bob performs the POVM that optimizes his information going about what Alice prepared.

It will turn out, in fact, that Alice is best off preparing pure-state messages (this follows from the subadditivity of the entropy). If a particular letter is prepared as the pure state $|\varphi_x\rangle$, Bob will receive

$$|\varphi_x\rangle\langle\varphi_x| \to \$(|\varphi_x\rangle\langle\varphi_x|) \equiv \boldsymbol{\rho}_x. \tag{5.185}$$

And if Alice sends the pure state $|\varphi_{x_1}\rangle \ldots |\varphi_{x_n}\rangle$, Bob receives the mixed state $\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}$. Thus, the ensemble of Alice's codewords determines as ensemble $\tilde{\mathcal{E}}^{(n)}$ of mixed states received by Bob. Hence Bob's optimal information gain is by definition $\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)})$, which satisfies the Holevo bound

$$\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) \leq \chi(\tilde{\mathcal{E}}^{(n)}). \tag{5.186}$$

Now Bob's ensemble is

$$\{\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n},\ p(x_1, x_2, \ldots, x_n)\}, \tag{5.187}$$

where $p(x_1, x_2 \ldots, x_n)$ is a completely arbitrary probability distribution on Alice's codewords. Let us calculate χ for this ensemble. We note that

$$\begin{aligned}
&\sum_{x_1 \ldots x_n} p(x_1, x_2, \ldots, x_n) S(\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}) \\
&= \sum_{x_1 \ldots x_n} p(x_1, x_2, \ldots, x_n) \left[S(\boldsymbol{\rho}_{x_1}) + S(\boldsymbol{\rho}_{x_2}) + \ldots + S(\boldsymbol{\rho}_{x_n})\right] \\
&= \sum_{x_1} p_1(x_1) S(\boldsymbol{\rho}_{x_1}) + \sum_{x_2} p_2(x_2) S(\boldsymbol{\rho}_{x_2}) + \ldots + \sum_{x_n} p_n(x_n) S(\boldsymbol{\rho}_{x_n}),
\end{aligned} \tag{5.188}$$

where, e.g., $p_1(x_1) = \sum_{x_2 \ldots x_n} p(x_1, x_2, \ldots, x_n)$ is the marginal probability distribution for the first letter. Furthermore, from subadditivity we have

$$S(\tilde{\boldsymbol{\rho}}^{(n)}) \leq S(\tilde{\boldsymbol{\rho}}_1) + S(\tilde{\boldsymbol{\rho}}_2) + \ldots + S(\tilde{\boldsymbol{\rho}}_n), \tag{5.189}$$

where $\tilde{\boldsymbol{\rho}}_i$ is the reduced density matrix for the ith letter. Combining eq. (5.188) and eq. (5.189) we find that

$$\chi(\tilde{\mathcal{E}}^{(n)}) \leq \chi(\tilde{\mathcal{E}}_1) + \ldots + \chi(\tilde{\mathcal{E}}_n), \tag{5.190}$$

where $\tilde{\mathcal{E}}_i$ is the marginal ensemble governing the ith letter that Bob receives. Eq. (5.190) applies to any ensemble of product states.

Now, for the channel described by the superoperator \$, we define the product-state *channel capacity*

$$C(\$) = \max_{\mathcal{E}} \chi(\$(\mathcal{E})). \tag{5.191}$$

Therefore, $\chi(\tilde{\mathcal{E}}_i) \leq C$ for each term in eq. (5.190) and we obtain

$$\chi(\tilde{\mathcal{E}}^{(n)}) \leq nC, \tag{5.192}$$

where $\tilde{\mathcal{E}}^{(n)}$ is any ensemble of product states. In particular, we infer from the Holevo bound that Bob's information gain is bounded above by nC. But we have seen that $\chi(\$(\mathcal{E}))$ bits per letter can be attained asymptotically for any $\mathcal{E}$, with the right choice of code and decoding observable. Therefore, C is the optimal number of bits per letter that can be sent through the noisy channel with negligible error probability, *if* the messages that Alice prepares are required to be product states.

We have left open the possibility that the product-state capacity $C(\$)$ might be exceeded if Alice is permitted to prepare *entangled* states of her n letters. It is not known (in January, 1998) whether there are quantum channels for which a higher rate can be attained by using entangled messages. This is one of the many interesting open questions in quantum information theory.

5.5 Entanglement Concentration

Before leaving our survey of quantum information theory, we will visit one more topic where Von Neumann entropy plays a central role: quantifying entanglement.

Consider two bipartite pure states. One is a *maximally* entangled state of two qubits

$$|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle). \tag{5.193}$$

The other is a *partially* entangled state of two *qutrits*

$$|\Psi\rangle = \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{2}|11\rangle + \frac{1}{2}|22\rangle. \tag{5.194}$$

which state is more entangled?

It is not immediately clear that the question has a meaningful answer. Why should it be possible to find an unambiguous way of placing all bipartite states on a continuum, of ordering them according to their degree of entanglement? Can we compare a pair of qutrits with a pair of qubits any more than we can compare an apple and an orange?

A crucial feature of entanglement is that it cannot be created by local operations. In particular, if Alice and Bob share a bipartite pure state, they cannot increase its Schmidt number by any local operations – any unitary transformation or POVM performed by Alice or Bob, even if Alice and Bob exchange classical messages about their actions and measurement outcomes. So a number used to quantify entanglement ought to have the property that local operations do not increase it. An obvious candidate is the Schmidt number, but on reflection it does not seem very satisfactory. Consider

$$|\Psi_\varepsilon\rangle = \sqrt{1-2|\varepsilon|^2}|00\rangle + \varepsilon|11\rangle + \varepsilon|22\rangle, \tag{5.195}$$

which has Schmidt number 3 for any $|\varepsilon| > 0$. Should we really say that $|\Psi_\varepsilon\rangle$ is "more entangled" than $|\phi^+\rangle$? Entanglement, after all, can be regarded as a resource – we might plan to use it for teleportation, for example. It seems clear that $|\Psi_\varepsilon\rangle$ (for $|\varepsilon| \ll 1$) is a less valuable resource than $|\varphi^+\rangle$.

It turns out, though, that there is a natural and sensible way to quantify the entanglement of any bipartite pure state. To compare two states, we perform local operations to change their entanglement to a common currency that can be compared directly. The common currency is a maximally entangled state.

A precise statement about interchangeability (via local operations) of various forms of entanglement will necessarily be an *asymptotic* statement. That is, to precisely quantify the entanglement of a particular bipartite pure state, $|\psi\rangle_{AB}$, let us imagine that we wish to prepare n identical copies of that state. We have available a large supply of maximally entangled *Bell pairs* shared by Alice and Bob. Alice and Bob are to use k of the Bell pairs $\left((|\phi^+\rangle_{AB})^k\right)$, and with local operations and classical communication, to prepare n copies of the desired state $((|\psi\rangle_{AB})^n)$. What is the minimum number $k_{\min}$ of Bell pairs with which they can perform this task?

And now suppose that n copies of $|\psi\rangle_{AB}$ have already been prepared. Alice and Bob are to perform local operations that will transform the entanglement of $(|\psi\rangle_{AB})^n$ back to the standard form; that is, they are to extract k' Bell pairs $\left((|\phi^+\rangle_{AB})^{k'}\right)$. What is the maximum number $k'_{\max}$ of Bell pairs that can be extracted (locally) from $(|\psi\rangle_{AB})^n$?

Since it is an in inviolable principle that local operations cannot create entanglement, it is certain that

$$k'_{\max} \le k_{\min}. \tag{5.196}$$

But we can show that

$$\lim_{n\to\infty} \frac{k_{\min}}{n} = \lim_{n\to\infty} \frac{k'_{\max}}{n} \equiv E(|\psi\rangle_{AB}). \tag{5.197}$$

In this sense, then, locally transforming n copies of the bipartite pure state $|\psi\rangle_{AB}$ into k' maximally entangled pairs is an asymptotically *reversible* process. Since n copies of $|\psi\rangle_{AB}$ can be exchanged for k Bell pairs and vice versa, we see that $\frac{k}{n}$ Bell pairs unambiguously characterizes the amount of entanglement carried by the state $|\psi\rangle_{AB}$. We will call the ratio k/n (in the $n \to \infty$ limit) the *entanglement* E of $|\psi\rangle_{AB}$. The quantity E measures both what we need to pay (in Bell pairs) to create $|\psi\rangle_{AB}$, and the value of $|\psi\rangle_{AB}$ as a resource (e.g., the number of qubits that can be faithfully teleported using $|\psi\rangle_{AB}$).

Now, given a particular pure state $|\psi\rangle_{AB}$, what is the value of E? Can you guess the answer? It is

$$E = S(\boldsymbol{\rho}_A) = S(\boldsymbol{\rho}_B); \tag{5.198}$$

the entanglement is the Von Neumann entropy of Alice's density matrix $\boldsymbol{\rho}_A$ (or Bob's density matrix $\boldsymbol{\rho}_B$). This is clearly the right answer in the case where $|\psi\rangle_{AB}$ is a product of k Bell pairs. In that case $\boldsymbol{\rho}_A$ (or $\boldsymbol{\rho}_B$) is $\frac{1}{2}\mathbf{1}$ for each qubit in Alice's possession

$$\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1} \otimes \frac{1}{2}\mathbf{1} \otimes \ldots \otimes \frac{1}{2}\mathbf{1}, \tag{5.199}$$

and

$$S(\boldsymbol{\rho}_A) = kS\left(\frac{1}{2}\mathbf{1}\right) = k. \tag{5.200}$$

We must now see why $E = S(\boldsymbol{\rho}_A)$ is the right answer for any bipartite pure state.

First we want to show that if Alice and Bob share $k = n(S(\boldsymbol{\rho}_A) + \delta)$ Bell pairs, than they can (by local operations) prepare $(|\psi\rangle_{AB})^n$ with high fidelity. They may perform this task by combining quantum teleportation with Schumacher compression. First, by locally manipulating a bipartite system AC that is under her control, Alice constructs (n copies of) the state $|\psi\rangle_{AC}$. Thus, we may regard the state of system C as a pure state drawn from an ensemble described by $\boldsymbol{\rho}_C$, where $S(\boldsymbol{\rho}_C) = S(\boldsymbol{\rho}_A)$. Next Alice performs Schumacher compression on her n copies of C, retaining good fidelity while squeezing the typical states in $(\mathcal{H}_C)^n$ down to a space $\tilde{\mathcal{H}}_C^{(n)}$ with

$$\dim \tilde{\mathcal{H}}_C^{(n)} = 2^{n(S(\boldsymbol{\rho}_A)+\delta)}. \tag{5.201}$$

Now Alice and Bob can use the $n(S(\boldsymbol{\rho}_A)+\delta)$ Bell pairs they share to teleport the compressed state from Alice's $\tilde{\mathcal{H}}_C^{(n)}$ to Bob's $\tilde{\mathcal{H}}_B^{(n)}$. The teleportation, which in principle has perfect fidelity, requires only local operations and classical communication, if Alice and Bob share the required number of Bell pairs. Finally, Bob Schumacher decompresses the state he receives; then Alice and Bob share $(|\psi\rangle_{AB})^n$ (with arbitrarily good fidelity as $n \to \infty$).

Let us now suppose that Alice and Bob have prepared the state $(|\psi\rangle_{AB})^n$. Since $|\psi\rangle_{AB}$ is, in general, a *partially* entangled state, the entanglement that Alice and Bob share is in a diluted form. They wish to *concentrate* their shared entanglement by squeezing it down to the smallest possible Hilbert space; that is, they want to convert it to maximally-entangled pairs. We will show that Alice and Bob can "distill" at least

$$k' = n(S(\boldsymbol{\rho}_A) - \delta) \tag{5.202}$$

Bell pairs from $(|\psi\rangle_{AB})^n$, with high likelihood of success.

Since we know that Alice and Bob are not able to create entanglement locally, they can't turn k Bell pairs into $k' > k$ pairs through local operations, at least not with high fidelity and success probability. It follows then that $nS(\boldsymbol{\rho}_A)$ is the minimum number of Bell pairs needed to create n copies of $|\psi\rangle_{AB}$, and that $nS(\boldsymbol{\rho}_A)$ is the maximal number of Bell pairs that can be distilled from n copies of $|\psi\rangle_{AB}$. If we could create $|\psi\rangle_{AB}$ from Bell pairs more efficiently, or we could distill Bell pairs from $|\psi\rangle_{AB}$ more efficiently, then we would have a way for Alice and Bob to increase their supply of Bell pairs with local operations, a known impossibility. Therefore, if we can find a way to distill $k' = n(S(\boldsymbol{\rho}_A) - \delta)$ Bell pairs from n copies of $|\psi\rangle_{AB}$, we know that $E = S(\boldsymbol{\rho}_A)$.

To illustrate the concentration of entanglement, imagine that Alice and Bob have n copies of the partially entangled pure state of two qubits

$$|\psi(\theta)\rangle_{AB} = \cos\theta|00\rangle + \sin\theta|11\rangle. \tag{5.203}$$

(Any bipartite pure state of two qubits can be written this way, if we adopt the Schmidt basis and a suitable phase convention.) That is, Alice and Bob share the state

$$(|\psi(\theta)\rangle)^n = (\cos\theta|00\rangle + \sin\theta|11\rangle)^n. \tag{5.204}$$

Now let Alice (or Bob) perform a local measurement on her (his) n qubits. Alice measures the *total* spin of her n qubits along the z-axis

$$\boldsymbol{\sigma}_{3,A}^{(\text{total})} = \sum_{i=1}^{n} \boldsymbol{\sigma}_{3,A}^{(i)}. \tag{5.205}$$

A crucial feature of this measurement is its "*fuzziness.*" The observable $\boldsymbol{\sigma}_{3,A}^{(\text{total})}$ is highly *degenerate*; Alice projects the state of her n spins onto one of the large eigenspaces of this observable. She does not measure the spin of any single qubit; in fact, she is very careful not to acquire any information other than the value of $\boldsymbol{\sigma}_{3,A}^{(\text{total})}$, or equivalently, the number of up spins.

If we expand eq. (5.204), we find altogether 2^n terms. Of these, there are $\binom{n}{m}$ terms in which exactly m of the qubits that Alice holds have the value 1. And each of these terms has a coefficient $(\cos\theta)^{n-m}(\sin\theta)^m$. Thus, the probability that Alice's measurement reveals that m spins are "up" is

$$P(m) = \binom{n}{m}(\cos^2\theta)^{n-m}(\sin^2\theta)^m. \tag{5.206}$$

Furthermore, if she obtains this outcome, then her measurement has prepared an *equally weighted* superposition of all $\binom{n}{m}$ states that have m up spins. (Of course, since Alice's and Bob's spins are perfectly correlated, if Bob were to measure $\sigma_{3,B}^{(\text{total})}$, he would find exactly the same result as Alice. Alternatively, Alice could report her result to Bob in a classical message, and so save Bob the trouble of doing the measurement himself.) No matter what the measurement result, Alice and Bob now share a new state $|\psi'\rangle_{AB}$ such that all the nonzero eigenvalues of $\boldsymbol{\rho}'_A$ (and $\boldsymbol{\rho}'_B$) are equal.

For n large, the probability distribution $P(m)$ in eq. (5.206) peaks sharply – the probability is close to 1 that m/n is close to $\sin^2\theta$ and that

$$\binom{n}{m} \sim \binom{n}{n\sin^2\theta} \sim 2^{nH(\sin^2\theta)}, \tag{5.207}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ is the entropy function. That is, with probability greater than $1-\varepsilon$, the entangled state now shared by Alice and Bob has a Schmidt number $\binom{n}{m}$ with

$$2^{n(H(\sin^2\theta)-\delta)} < \binom{n}{m} < 2^{n(H(\sin^2\theta)+\delta)}. \tag{5.208}$$

Now Alice and Bob want to convert their shared entanglement to standard ($|\phi^+\rangle$) Bell pairs. If the Schmidt number of their shared maximally entangled state happened to be a power of 2, this would be easy. Both Alice and Bob could perform a unitary transformation that would rotate the 2^k-dimensional support of her/his density matrix to the Hilbert space of k-qubits, and then they could discard the rest of their qubits. The k pairs that they retain would then be maximally entangled.

Of course $\binom{n}{m}$ need not be close to a power of 2. But if Alice and Bob share many batches of n copies of the partially entangled state, they can concentrate the entanglement in each batch. After operating on ℓ batches, they will have obtained a maximally entangled state with Schmidt number

$$N_{\text{Schm}} = \binom{n}{m_1}\binom{n}{m_2}\binom{n}{m_3}\cdots\binom{n}{m_\ell}, \tag{5.209}$$

where each m_i is typically close to $n\sin^2\theta$. For any $\varepsilon > 0$, this Schmidt number will eventually, for some ℓ, be close to a power of 2,

$$2^{k_\ell} \le N_{\text{Schm}} < 2^{k_\ell}(1+\varepsilon). \tag{5.210}$$

At that point, either Alice or Bob can perform a measurement that attempts to project the support of dimension $2^{k_\ell}(1+\varepsilon)$ of her/his density matrix to a subspace of dimension 2^{k_ℓ}, succeeding with probability $1-\varepsilon$. *Then* they rotate the support to the Hilbert space of k_ℓ qubits, and discard the rest of their qubits. Typically, k_ℓ is close to $n\ell H(\sin^2\theta)$, so that they distill about $H(\sin^2\theta)$ maximally entangled pairs from each partially entangled state, with a success probability close to 1.

Of course, though the number m of up spins that Alice (or Bob) finds in her (his) measurement is typically close to $n\sin^2\theta$, it can fluctuate about this value. Sometimes Alice and Bob will be lucky, and then will manage to distill more than $H(\sin^2\theta)$ Bell pairs per copy of $|\psi(\theta)\rangle_{AB}$. But the probability of doing substantially better becomes negligible as $n\to\infty$.

These considerations easily generalize to bipartite pure states in larger Hilbert spaces. A bipartite pure state with Schmidt number s can be expressed, in the Schmidt basis, as

$$|\psi(a_1, a_2, \ldots, a_s)\rangle_{AB} = a_1|11\rangle + a_2|22\rangle + \ldots + a_s|ss\rangle. \tag{5.211}$$

Then in the state $(|\psi\rangle_{AB})^n$, Alice (or Bob) can measure the total number of $|1\rangle$'s, the total number of $|2\rangle$'s, etc. in her (his) possession. If she finds $m_1|1\rangle$'s, $m_2|2\rangle$'s, etc., then her measurement prepares a maximally entangled state with Schmidt number

$$N_{\text{Schm}} = \frac{n!}{(m_1)!(m_2)!\cdots(m_s)!}. \tag{5.212}$$

For m large, Alice will typically find

$$m_i \sim |a_i|^2 n, \tag{5.213}$$

and therefore

$$N_{\text{Sch}} \sim 2^{nH}, \tag{5.214}$$

where

$$H = \sum_i -|a_i|^2\log|a_i|^2 = S(\boldsymbol{\rho}_A). \tag{5.215}$$

Thus, asymptotically for $n\to\infty$, close to $nS(\boldsymbol{\rho}_A)$ Bell pairs can be distilled from n copies of $|\psi\rangle_{AB}$.

5.5.1 Mixed-state entanglement

We have found a well-motivated and unambiguous way to quantify the entanglement of a bipartite pure state $|\psi\rangle_{AB}$: $E = S(\boldsymbol{\rho}_A)$, where

$$\boldsymbol{\rho}_A = \text{tr}_B(|\psi\rangle_{AB}\,{}_{AB}\langle\psi|). \tag{5.216}$$

It is also of considerable interest to quantify the entanglement of bipartite mixed states. Unfortunately, mixed-state entanglement is not nearly as well understood as pure-state entanglement, and is the topic of much current research.

Suppose that $\boldsymbol{\rho}_{AB}$ is a mixed state shared by Alice and Bob, and that they have n identical copies of this state. And suppose that, asymptotically as $n \to \infty$, Alice and Bob can prepare $(\boldsymbol{\rho}_{AB})^n$, with good fidelity and high success probability, from k Bell pairs using local operations and classical communication. We define the *entanglement of formation* F of $\boldsymbol{\rho}_{AB}$ as

$$F(\boldsymbol{\rho}_{AB}) = \lim_{n\to\infty} \frac{k_{\text{min}}}{n}. \tag{5.217}$$

Further, suppose that Alice and Bob can use local operations and classical communication to distill k' Bell pairs from n copies of $\boldsymbol{\rho}_{AB}$. We define the *entanglement of distillation* D of $\boldsymbol{\rho}_{AB}$ as

$$D(\boldsymbol{\rho}_{AB}) = \lim_{n\to\infty} \frac{k'_{\text{max}}}{n}. \tag{5.218}$$

For pure states, we found $D = E = F$. But for mixed states, no explicit general formulas for D or F are known. Since entanglement cannot be created locally, we know that $D \le F$, but it is not known (in January, 1998) whether $D = F$. However, one strongly suspects that, for mixed states, $D < F$. To prepare the mixed state $(\boldsymbol{\rho}_{AB})^n$ from the pure state $(|\phi^+\rangle_{AB\ AB}\langle\phi^+|)^k$, we must discard some quantum information. It would be quite surprising if this process turned out to be (asymptotically) reversible.

It is useful to distinguish two different types of entanglement of distillation. D_1 denotes the number of Bell pairs that can be distilled if only one-way classical communication is allowed (e.g., Alice can *send* messages to Bob but she cannot *receive* messages from Bob). $D_2 = D$ denotes the entanglement of distillation if the classical communication is unrestricted. It is known that $D_1 < D_2$, and hence that $D_1 < F$ for some mixed states (while $D_1 = D_2 = F$ for pure states).

One reason for the interest in mixed-state entanglement (and in D_1 in particular) is a connection with the transmission of quantum information through noisy quantum channels. If a quantum channel described by a superoperator \$ is not *too* noisy, then we can construct an n-letter block code such that quantum information can be encoded, sent through the channel $(\$)^n$, decoded, and recovered with arbitrarily good fidelity as $n \to \infty$. The optimal number of encoded qubits per letter that can be transmitted through the channel is called the quantum channel capacity $C(\$)$. It turns out that $C(\$)$ can be related to D_1 of a particular mixed state associated with the channel — but we will postpone further discussion of the quantum channel capacity until later.

5.6 Summary

Shannon entropy and classical data compression. The *Shannon entropy* of an ensemble $X = \{x, p(x)\}$ is $H(x) \equiv \langle -\log p(x)\rangle$; it quantifies the compressibility of classical information. A message n letters long, where each letter is drawn independently from X, can be compressed to $H(x)$ bits per letter (and no further), yet can still be decoded with arbitrarily good accuracy as $n \to \infty$.

Mutual information and classical channel capacity. The *mutual information* $I(X;Y) = H(X) + H(Y) - H(X,Y)$ quantifies how ensembles X and Y are correlated; when we learn the value of y we acquire (on the average) $I(X;Y)$ bits of information about x. The capacity of a memoryless noisy classical communication channel is $C = \max_{\{p(x)\}} I(X;Y)$. This is the highest number of bits per letter that can be transmitted through the channel (using the best possible code) with negligible error probability as $n \to \infty$.

Von Neumann entropy, Holevo information, and quantum data compression. The *Von Neumann entropy* of a density matrix $\boldsymbol{\rho}$ is

$$S(\boldsymbol{\rho}) = -\text{tr}\boldsymbol{\rho}\log\boldsymbol{\rho}, \tag{5.219}$$

and the *Holevo information* of an ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$ of quantum states is

$$\chi(\mathcal{E}) = S(\sum_x p_x \boldsymbol{\rho}_x) - \sum_x p_x S(\boldsymbol{\rho}_x). \tag{5.220}$$

The Von Neumann entropy quantifies the compressibility of an ensemble of pure quantum states. A message n letters long, where each letter is drawn independently from the ensemble $\{|\varphi_x\rangle, p_x\}$, can be compressed to $S(\boldsymbol{\rho})$ qubits per letter (and no further), yet can still be decoded with arbitrarily good fidelity as $n \to \infty$. If the letters are drawn from the ensemble $\mathcal{E}$ of mixed quantum states, then high-fidelity compression to fewer than $\chi(\mathcal{E})$ qubits per letter is not possible.

Accessible information. The *accessible information* of an ensemble $\mathcal{E}$ of quantum states is the maximal number of bits of information that can be acquired about the preparation of the state (on the average) with the best possible measurement. The accessible information cannot exceed the Holevo information of the ensemble. An n-letter code can be constructed such that the marginal ensemble for each letter is close to $\mathcal{E}$, and the accessible information per letter is close to $\chi(\mathcal{E})$. The product-state capacity of a quantum channel \$ is

$$C(\$) = \max_{\mathcal{E}} \chi(\$(\mathcal{E})). \tag{5.221}$$

This is the highest number of classical bits per letter than can be transmitted through the quantum channel, with negligible error probability as $n \to \infty$, assuming that each codeword is a tensor product of letter states.

Entanglement concentration. The *entanglement* E of a bipartite pure state $|\psi\rangle_{AB}$ is $E = S(\boldsymbol{\rho}_A)$ where $\boldsymbol{\rho}_A = \mathrm{tr}_B(|\psi\rangle_{AB\ AB}\langle\psi|)$. With local operations and classical communication, we can prepare n copies of $|\psi\rangle_{AB}$ from nE Bell pairs (but not from fewer), and we can distill nE Bells pairs (but not more) from n copies of $|\psi\rangle_{AB}$ (asymptotically as $n \to \infty$).

5.7 Exercises

5.1 Distinguishing nonorthogonal states.

Alice has prepared a single qubit in one of the two (nonorthogonal) states

$$|u\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ |v\rangle = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix}, \tag{5.222}$$

where $0 < \theta < \pi$. Bob knows the value of θ, but he has no idea whether Alice prepared $|u\rangle$ or $|v\rangle$, and he is to perform a measurement to learn what he can about Alice's preparation.

Bob considers three possible measurements:

a) An orthogonal measurement with

$$\boldsymbol{E}_1 = |u\rangle\langle u|, \quad \boldsymbol{E}_2 = \mathbf{1} - |u\rangle\langle u|. \tag{5.223}$$

(In this case, if Bob obtains outcome 2, he knows that Alice must have prepared $|v\rangle$.)

b) A three-outcome POVM with

$$\boldsymbol{F}_1 = A(\mathbf{1} - |u\rangle\langle u|), \quad \boldsymbol{F}_2 = A(\mathbf{1} - |v\rangle\langle v|)$$

$$\boldsymbol{F}_3 = (1 - 2A)\mathbf{1} + A(|u\rangle\langle u| + |v\rangle\langle v|), \tag{5.224}$$

where A has the largest value consistent with positivity of $\boldsymbol{F}_3$. (In this case, Bob determines the preparation unambiguously if he obtains outcomes 1 or 2, but learns nothing from outcome 3.)

c) An orthogonal measurement with

$$\boldsymbol{E}_1 = |w\rangle\langle w|, \quad \boldsymbol{E}_2 = \mathbf{1} - |w\rangle\langle w|, \tag{5.225}$$

where

$$|w\rangle = \begin{pmatrix} \cos\left[\frac{1}{2}\left(\frac{\theta}{2} + \frac{\pi}{2}\right)\right] \\ \sin\left[\frac{1}{2}\left(\frac{\theta}{2} + \frac{\pi}{2}\right)\right] \end{pmatrix}. \tag{5.226}$$

(In this case $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are projections onto the spin states that are oriented in the $x-z$ plane normal to the axis that bisects the orientations of $|u\rangle$ and $|v\rangle$.)

Find Bob's average information gain $I(\theta)$ (the mutual information of the preparation and the measurement outcome) in all three cases, and plot all three as a function of θ. Which measurement should Bob choose?

5.2 Relative entropy.

The *relative entropy* $S(\boldsymbol{\rho}|\boldsymbol{\sigma})$ of two density matrices $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ is defined by

$$S(\boldsymbol{\rho}|\boldsymbol{\sigma}) = \mathrm{tr}\boldsymbol{\rho}(\log\boldsymbol{\rho} - \log\boldsymbol{\sigma}). \tag{5.227}$$

You will show that $S(\boldsymbol{\rho}|\boldsymbol{\sigma})$ is nonnegative, and derive some consequences of this property.

a) A differentiable real-valued function of a real variable is *concave* if

$$f(y) - f(x) \le (y - x)f'(x), \tag{5.228}$$

for all x and y. Show that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are observables, and f is concave, then

$$\mathrm{tr}(f(\boldsymbol{b}) - f(\boldsymbol{a})) \le \mathrm{tr}[(\boldsymbol{b} - \boldsymbol{a})f'(\boldsymbol{a})]. \tag{5.229}$$

b) Show that $f(x) = -x \log x$ is concave for $x > 0$.

c) Use (a) and (b) to show $S(\boldsymbol{\rho}|\boldsymbol{\sigma}) \geq 0$ for any two density matrices $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$.

d) Use nonnegativity of $S(\boldsymbol{\rho}|\boldsymbol{\sigma})$ to show that if $\boldsymbol{\rho}$ has its support on a space of dimension D, then

$$S(\boldsymbol{\rho}) \leq \log D. \tag{5.230}$$

e) Use nonnegativity of relative entropy to prove the *subadditivity* of entropy

$$S(\boldsymbol{\rho}_{AB}) \leq S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_B). \tag{5.231}$$

[Hint: Consider the relative entropy of $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$ and $\boldsymbol{\rho}_{AB}$.]

f) Use subadditivity to prove the *concavity* of the entropy:

$$S(\sum_i \lambda_i \boldsymbol{\rho}_i) \geq \sum_i \lambda_i S(\boldsymbol{\rho}_i), \tag{5.232}$$

where the λ_i's are positive real numbers summing to one. [Hint: Apply subadditivity to

$$\boldsymbol{\rho}_{AB} = \sum_i \lambda_i \left(\boldsymbol{\rho}_i\right)_A \otimes \left(|e_i\rangle\langle e_i|\right)_B . \,] \tag{5.233}$$

g) Use subadditivity to prove the *triangle inequality* (also called the Araki-Lieb inequality):

$$S(\boldsymbol{\rho}_{AB}) \geq |S(\boldsymbol{\rho}_A) - S(\boldsymbol{\rho}_B)|. \tag{5.234}$$

[Hint: Consider a purification of $\boldsymbol{\rho}_{AB}$; that is, construct a pure state $|\psi\rangle$ such that $\boldsymbol{\rho}_{AB} = \mathrm{tr}_C |\psi\rangle\langle\psi|$. Then apply subadditivity to $\boldsymbol{\rho}_{BC}$.]

5.3 Lindblad–Uhlmann monotonicity.

According to a theorem proved by Lindblad and by Uhlmann, relative entropy on $\mathcal{H}_A \otimes \mathcal{H}_B$ has a property called *monotonicity*:

$$S(\boldsymbol{\rho}_A|\boldsymbol{\sigma}_A) \leq S(\boldsymbol{\rho}_{AB}|\boldsymbol{\sigma}_{AB}); \tag{5.235}$$

The relative entropy of two density matrices on a system AB cannot be less than the induced relative entropy on the subsystem A.

a) Use Lindblad-Uhlmann monotonicity to prove the strong subadditivity property of the Von Neumann entropy. [Hint: On a tripartite system ABC, consider the relative entropy of $\boldsymbol{\rho}_{ABC}$ and $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_{BC}$.]

b) Use Lindblad–Uhlmann monotonicity to show that the action of a superoperator cannot increase relative entropy, that is,

$$S(\$\boldsymbol{\rho}|\$\boldsymbol{\sigma}) \leq S(\boldsymbol{\rho}|\boldsymbol{\sigma}), \tag{5.236}$$

Where \$ is any superoperator (completely positive map). [Hint: Recall that any superoperator has a unitary representation.]

c) Show that it follows from (b) that a superoperator cannot increase the Holevo information of an ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$ of mixed states:

$$\chi(\$(\mathcal{E})) \leq \chi(\mathcal{E}), \tag{5.237}$$

where

$$\chi(\mathcal{E}) = S\left(\sum_x p_x \boldsymbol{\rho}_x\right) - \sum_x p_x S(\boldsymbol{\rho}_x). \tag{5.238}$$

5.4 The Peres–Wootters POVM.

Consider the Peres–Wootters information source described in §5.4.2 of the lecture notes. It prepares one of the three states

$$|\Phi_a\rangle = |\varphi_a\rangle|\varphi_a\rangle, \quad a = 1, 2, 3, \tag{5.239}$$

each occurring with *a priori* probability $\frac{1}{3}$, where the $|\varphi_a\rangle$'s are defined in eq. (5.149).

a) Express the density matrix

$$\boldsymbol{\rho} = \frac{1}{3}\left(\sum_a |\Phi_a\rangle\langle\Phi_a|\right), \tag{5.240}$$

in terms of the Bell basis of maximally entangled states $\{|\phi^\pm\rangle, |\psi^\pm\rangle\}$, and compute $S(\boldsymbol{\rho})$.

b) For the three vectors $|\Phi_a\rangle, a = 1, 2, 3$, construct the "pretty good measurement" defined in eq. (5.162). (Again, expand the $|\Phi_a\rangle$'s in the Bell basis.) In this case, the PGM is an orthogonal measurement. Express the elements of the PGM basis in terms of the Bell basis.

c) Compute the mutual information of the PGM outcome and the preparation.

5.5 Teleportation with mixed states.

An operational way to define entanglement is that an entangled state can be used to teleport an unknown quantum state with better fidelity than could be achieved with local operations and classical communication only. In this exercise, you will show that there are mixed states that are entangled in this sense, yet do not violate any Bell inequality. Hence, for mixed states (in contrast to pure states) "entangled" and "Bell-inequality-violating" are not equivalent.

Consider a "noisy" entangled pair with density matrix.

$$\boldsymbol{\rho}(\lambda) = (1-\lambda)|\psi^-\rangle\langle\psi^-| + \lambda\frac{1}{4}\mathbf{1}. \tag{5.241}$$

a) Find the fidelity F that can be attained if the state $\boldsymbol{\rho}(\lambda)$ is used to teleport a qubit from Alice to Bob. [Hint: Recall that you showed in an earlier exercise that a "random guess" has fidelity $F = \frac{1}{2}$.]

b) For what values of λ is the fidelity found in (a) better than what can be achieved if Alice measures her qubit and sends a classical message to Bob? [Hint: Earlier, you showed that $F = 2/3$ can be achieved if Alice measures her qubit. In fact this is the best possible F attainable with classical communication.]

c) Compute

$$\text{Prob}(\uparrow_{\hat{n}}\uparrow_{\hat{m}}) \equiv \text{tr}\left(\boldsymbol{E}_A(\hat{n})\boldsymbol{E}_B(\hat{m})\boldsymbol{\rho}(\lambda)\right), \tag{5.242}$$

where $\boldsymbol{E}_A(\hat{n})$ is the projection of Alice's qubit onto $|\uparrow_{\hat{n}}\rangle$ and $\boldsymbol{E}_B(\hat{m})$ is the projection of Bob's qubit onto $|\uparrow_{\hat{m}}\rangle$.

d) Consider the case $\lambda = 1/2$. Show that in this case the state $\boldsymbol{\rho}(\lambda)$ violates no Bell inequalities. Hint: It suffices to construct a local hidden variable model that correctly reproduces the spin correlations found in (c), for $\lambda = 1/2$. Suppose that the hidden variable $\hat{\alpha}$ is uniformly distributed on the unit sphere, and that there are functions f_A and f_B such that

$$\text{Prob}_A(\uparrow_{\hat{n}}) = f_A(\hat{\alpha}\cdot\hat{n}), \quad \text{Prob}_B(\uparrow_{\hat{m}}) = f_B(\hat{\alpha}\cdot\hat{m}). \tag{5.243}$$

The problem is to find f_A and f_B (where $0 \le f_{A,B} \le 1$) with the properties

$$\int_{\hat{\alpha}} f_A(\hat{\alpha}\cdot\hat{n}) = 1/2, \quad \int_{\hat{\alpha}} f_B(\hat{\alpha}\cdot\hat{m}) = 1/2,$$
$$\int_{\hat{\alpha}} f_A(\hat{\alpha}\cdot\hat{n}) f_B(\hat{\alpha}\cdot\hat{m}) = \text{Prob}(\uparrow_{\hat{n}}\uparrow_{\hat{m}}). \tag{5.244}$$

Chapter 6

Quantum Computation

6.1 Classical Circuits

The concept of a quantum computer was introduced in Chapter 1. Here we will specify our model of quantum computation more precisely, and we will point out some basic properties of the model. But before we explain what a quantum computer does, perhaps we should say what a classical computer does.

6.1.1 Universal gates

A classical (deterministic) computer evaluates a function: given n-bits of input it produces m-bits of output that are uniquely determined by the input; that is, it finds the value of

$$f : \{0,1\}^n \to \{0,1\}^m, \tag{6.1}$$

for a particular specified n-bit argument. A function with an m-bit value is equivalent to m functions, each with a one-bit value, so we may just as well say that the basic task performed by a computer is the evaluation of

$$f : \{0,1\}^n \to \{0,1\}. \tag{6.2}$$

We can easily count the number of such functions. There are 2^n possible inputs, and for each input there are two possible outputs. So there are altogether 2^{2^n} functions taking n bits to one bit.

The evaluation of any such function can be reduced to a sequence of elementary logical operations. Let us divide the possible values of the input

$$x = x_1 x_2 x_3 \ldots x_n, \tag{6.3}$$

into one set of values for which $f(x) = 1$, and a complementary set for which $f(x) = 0$. For each $x^{(a)}$ such that $f(x^{(a)}) = 1$, consider the function $f^{(a)}$ such that

$$f^{(a)}(x) = \begin{cases} 1 & x = x^{(a)} \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

Then

$$f(x) = f^{(1)}(x) \vee f^{(2)}(x) \vee f^{(3)}(x) \vee \ldots. \tag{6.5}$$

f is the logical OR ($\vee$) of all the $f^{(a)}$'s. In binary arithmetic the $\vee$ operation of two bits may be represented

$$x \vee y = x + y - x \cdot y; \tag{6.6}$$

it has the value 0 if x and y are both zero, and the value 1 otherwise.

Now consider the evaluation of $f^{(a)}$. In the case where $x^{(a)} = 111\ldots 1$, we may write

$$f^{(a)}(x) = x_1 \wedge x_2 \wedge x_3 \ldots \wedge x_n; \tag{6.7}$$

it is the logical AND ($\wedge$) of all n bits. In binary arithmetic, the AND is the product

$$x \wedge y = x \cdot y. \tag{6.8}$$

For any other $x^{(a)}$, $f^{(a)}$ is again obtained as the AND of n bits, but where the NOT ($\neg$) operation is first applied to each x_i such that $x_i^{(a)} = 0$; for example

$$f^{(a)}(x) = (\neg x_1) \wedge x_2 \wedge x_3 \wedge (\neg x_4) \wedge \ldots \tag{6.9}$$

if

$$x^{(a)} = 0110\ldots. \tag{6.10}$$

The NOT operation is represented in binary arithmetic as

$$\neg x = 1 - x. \tag{6.11}$$

We have now constructed the function $f(x)$ from three elementary logical connectives: NOT, AND, OR. The expression we obtained is called the "disjunctive normal form" of $f(x)$. We have also implicitly used another operation, COPY, that takes one bit to two bits:

$$\text{COPY} : x \to xx. \tag{6.12}$$

We need the COPY operation because each $f^{(a)}$ in the disjunctive normal form expansion of f requires its own copy of x to act on.

In fact, we can pare our set of elementary logical connectives to a smaller set. Let us define a NAND ("NOT AND") operation by

$$x \uparrow y = \neg(x \wedge y) = (\neg x) \vee (\neg y). \tag{6.13}$$

In binary arithmetic, the NAND operation is

$$x \uparrow y = 1 - xy. \tag{6.14}$$

If we can COPY, we can use NAND to perform NOT:

$$x \uparrow x = 1 - x^2 = 1 - x = \neg x. \tag{6.15}$$

(Alternatively, if we can prepare the constant $y = 1$, then $x \uparrow 1 = 1-x = \neg x$.) Also,

$$(x \uparrow y) \uparrow (x \uparrow y) = \neg(x \uparrow y) = 1 - (1 - xy) = xy = x \wedge y, \tag{6.16}$$

and

$$\begin{aligned}(x \uparrow x) \uparrow (y \uparrow y) &= (\neg x) \uparrow (\neg y) = 1 - (1 - x)(1 - y) \\ &= x + y - xy = x \vee y.\end{aligned} \tag{6.17}$$

So if we can COPY, NAND performs AND and OR as well. We conclude that the single logical connective NAND, together with COPY, suffices to evaluate any function f. (You can check that an alternative possible choice of the universal connective is NOR:

$$x \downarrow y = \neg(x \vee y) = (\neg x) \wedge (\neg y).) \tag{6.18}$$

If we are able to prepare a constant bit ($x = 0$ or $x = 1$), we can reduce the number of elementary operations from two to one. The NAND/NOT gate

$$(x, y) \to (1 - x, 1 - xy), \tag{6.19}$$

computes NAND (if we ignore the first output bit) and performs copy (if we set the second input bit to $y = 1$, and we subsequently apply NOT to both output bits). We say, therefore, that NAND/NOT is a universal gate. If we have a supply of constant bits, and we can apply the NAND/NOT gates to any chosen pair of input bits, then we can perform a sequence of NAND/NOT gates to evaluate any function $f : \{0,1\}^n \to \{0,1\}$ for any value of the input $x = x_1 x_2 \ldots x_n$.

These considerations motivate the circuit model of computation. A computer has a few basic components that can perform elementary operations on bits or pairs of bits, such as COPY, NOT, AND, OR. It can also prepare a constant bit or input a variable bit. A computation is a finite sequence of such operations, a *circuit*, applied to a specified string of input bits.[1] The result of the computation is the final value of all remaining bits, after all the elementary operations have been executed.

It is a fundamental result in the theory of computation that just a few elementary gates suffice to evaluate any function of a finite input. This result means that with very simple hardware components, we can build up arbitrarily complex computations.

So far, we have only considered a computation that acts on a particular fixed input, but we may also consider *families* of circuits that act on inputs of variable size. Circuit families provide a useful scheme for analyzing and classifying the *complexity* of computations, a scheme that will have a natural generalization when we turn to quantum computation.

6.1.2 Circuit complexity

In the study of complexity, we will often be interested in functions with a one-bit output

$$f : \{0,1\}^n \to \{0,1\}. \tag{6.20}$$

[1]The circuit is required to be *acyclic*, meaning that no *directed* closed loops are permitted.

Such a function f may be said to encode a solution to a "decision problem" — the function examines the input and issues a YES or NO answer. Often, a question that would not be stated colloquially as a question with a YES/NO answer can be "repackaged" as a decision problem. For example, the function that defines the FACTORING problem is:

$$f(x,y) = \begin{cases} 1 & \text{if integer } x \text{ has a divisor less than } y, \\ 0 & \text{otherwise;} \end{cases} \tag{6.21}$$

knowing $f(x,y)$ for all $y < x$ is equivalent to knowing the *least* nontrivial factor of y. Another important example of a decision problem is the HAMILTONIAN path problem: let the input be an ℓ-vertex graph, represented by an $\ell \times \ell$ adjacency matrix (a 1 in the ij entry means there is an edge linking vertices i and j); the function is

$$f(x) = \begin{cases} 1 & \text{if graph } x \text{ has a Hamiltonian path,} \\ 0 & \text{otherwise.} \end{cases} \tag{6.22}$$

(A path is Hamiltonian if it visits each vertex exactly once.)

We wish to gauge how hard a problem is by quantifying the resources needed to solve the problem. For a decision problem, a reasonable measure of hardness is the *size* of the smallest circuit that computes the corresponding function $f : \{0,1\}^n \to \{0,1\}$. By size we mean the number of elementary gates or components that we must wire together to evaluate f. We may also be interested in how much *time* it takes to do the computation if many gates are permitted to execute in parallel. The *depth* of a circuit is the number of time steps required, assuming that gates acting on distinct bits can operate simultaneously (that is, the depth is the maximum length of a directed path from the input to the output of the circuit). The *width* of a circuit is the maximum number of gates that act in any one time step.

We would like to divide the decision problems into two classes: easy and hard. But where should we draw the line? For this purpose, we consider infinite families of decision problems with variable input size; that is, where the number of bits of input can be any integer n. Then we can examine how the size of the circuit that solves the problem scales with n.

If we use the scaling behavior of a circuit family to characterize the difficulty of a problem, there is a subtlety. It would be cheating to hide the difficulty of the problem in the *design* of the circuit. Therefore, we should

restrict attention to circuit families that have acceptable "uniformity" properties — it must be "easy" to build the circuit with $n+1$ bits of input once we have constructed the circuit with an n-bit input.

Associated with a family of functions $\{f_n\}$ (where f_n has n-bit input) are circuits $\{C_n\}$ that compute the functions. We say that a circuit family $\{C_n\}$ is "polynomial size" if the size of C_n grows with n no faster than a power of n,

$$\text{size } (C_n) \le \text{poly } (n), \tag{6.23}$$

where poly denotes a polynomial. Then we define:

$$P = \{\text{decision problem solved by polynomial-size circuit families}\}$$

(P for "polynomial time"). Decision problems in P are "easy." The rest are "hard." Notice that C_n computes $f_n(x)$ for every possible n-bit input, and therefore, if a decision problem is in P we can find the answer even for the "worst-case" input using a circuit of size no greater than poly(n). (As noted above, we implicitly assume that the circuit family is "uniform" so that the *design* of the circuit can itself be solved by a polynomial-time algorithm. Under this assumption, solvability in polynomial time by a circuit family is equivalent to solvability in polynomial time by a universal Turing machine.)

Of course, to determine the size of a circuit that computes f_n, we must know what the elementary components of the circuit are. Fortunately, though, whether a problem lies in P does not depend on what gate set we choose, as long as the gates are universal, the gate set is finite, and each gate acts on a set of bits of bounded size. One universal gate set can *simulate* another.

The vast majority of function families $f : \{0,1\}^n \to \{0,1\}$ are not in P. For most functions, the output is essentially random, and there is no better way to "compute" $f(x)$ than to consult a look-up table of its values. Since there are 2^n n-bit inputs, the look-up table has *exponential* size, and a circuit that encodes the table must also have exponential size. The problems in P belong to a very special class — they have enough structure so that the function f can be computed efficiently.

Of particular interest are decision problems that can be answered by exhibiting an example that is easy to verify. For example, given x and $y < x$, it is hard (in the worst case) to determine if x has a factor less than y. But if someone kindly provides a $z < y$ that divides x, it is easy for us to check that z is indeed a factor of x. Similarly, it is hard to determine if a graph has a Hamiltonian path, but if someone kindly provides a path, it is easy to verify that the path really is Hamiltonian.

This concept that a problem may be hard to solve, but that a solution can be easily verified once found, can be formalized by the notion of a "non-deterministic" circuit. A nondeterministic circuit $\tilde{C}_{n,m}(x^{(n)}, y^{(m)})$ associated with the circuit $C_n(x^{(n)})$ has the property:

$$C_n(x^{(n)}) = 1 \text{ iff } \tilde{C}_{n,m}(x^{(n)}, y^{(m)}) = 1 \quad \text{for } some\ y^{(m)}. \tag{6.24}$$

(where $x^{(n)}$ is n bits and $y^{(m)}$ is m bits.) Thus for a particular $x^{(n)}$ we can use $\tilde{C}_{n,m}$ to *verify* that $C_n(x^{(n)} = 1$, *if* we are fortunate enough to have the right $y^{(m)}$ in hand. We define:

> NP: {decision problems that admit a polynomial-size *nondeterministic* circuit family}

(NP for "nondeterministic polynomial time"). If a problem is in NP, there is no guarantee that the problem is easy, only that a solution is easy to check once we have the right information. Evidently $P \subseteq NP$. Like P, the NP problems are a small subclass of all decision problems.

Much of complexity theory is built on a fundamental conjecture:

$$\text{Conjecture}: P \neq NP; \tag{6.25}$$

there exist hard decision problems whose solutions are easily verified. Unfortunately, this important conjecture still awaits proof. But after 30 years of trying to show otherwise, most complexity experts are firmly confident of its validity.

An important example of a problem in NP is CIRCUIT-SAT. In this case the input is a circuit C with n gates, m input bits, and one output bit. The problem is to find if there is *any* m-bit input for which the output is 1. The function to be evaluated is

$$f(C) = \begin{cases} 1 & \text{if there exists } x^{(m)} \text{ with } C(x^{(m)}) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.26}$$

This problem is in NP because, *given* a circuit, it is easy to *simulate* the circuit and evaluate its output for any particular input.

I'm going to state some important results in complexity theory that will be relevant for us. There won't be time for proofs. You can find out more

by consulting one of the many textbooks on the subject; one good one is *Computers and Intractability: A Guide to the Theory of NP-Completeness*, by M. R. Garey and D. S. Johnson.

Many of the insights engendered by complexity theory flow from Cook's Theorem (1971). The theorem states that *every* problem in NP is *polynomially reducible* to CIRCUIT-SAT. This means that for any PROBLEM $\in NP$, there is a polynomial-size circuit family that maps an "instance" $x^{(n)}$ of PROBLEM to an "instance" $y^{(m)}$ of CIRCUIT-SAT; that is

$$\text{CIRCUIT} - \text{SAT } (y^{(m)}) = 1 \text{ iff PROBLEM } (x^{(n)}) = 1. \tag{6.27}$$

It follows that if we had a magical device that could efficiently solve CIRCUIT-SAT (a CIRCUIT-SAT "oracle"), we could couple that device with the polynomial reduction to efficiently solve PROBLEM. Cook's theorem tells us that if it turns out that CIRCUIT-SAT $\in P$, then $P = NP$.

A problem that, like CIRCUIT-SAT, has the property that every problem in NP is polynomially reducible to it, is called *NP-complete (NPC)*. Since Cook, many other examples have been found. To show that a PROBLEM $\in NP$ is NP-complete, it suffices to find a polynomial reduction to PROBLEM of another problem that is already known to be NP-complete. For example, one can exhibit a polynomial reduction of CIRCUIT-SAT to HAMILTONIAN. It follows from Cook's theorem that HAMILTONIAN is also NP-complete.

If we assume that $P \neq NP$, it follows that there exist problems in NP of intermediate difficulty (the class NPI). These are neither P nor NPC.

Another important complexity class is called co-NP. Heuristically, NP decision problems are ones we can answer by exhibiting an *example* if the answer is YES, while co-NP problems can be answered with a *counter-example* if the answer is NO. More formally:

$$\{C\} \in NP : C(x) = 1 \text{ iff } C(x, y) = 1 \text{ for some } y \tag{6.28}$$

$$\{C\} \in \text{co}-NP : C(x) = 1 \text{ iff } C(x, y) = 1 \text{ for all } y. \tag{6.29}$$

Clearly, there is a symmetry relating the classes NP and co-NP — whether we consider a problem to be in NP or co-NP depends on how we choose to frame the question. ("Is there a Hamiltonian circuit?" is in NP. "Is there no Hamiltonian circuit?" is in co-NP). But the interesting question is: is a problem in *both* NP and co-NP? If so, then we can easily verify the answer (once a suitable example is in hand) regardless of whether the answer is YES or NO. It is believed (though not proved) that $NP \neq$ co$-NP$. (For example, we can show that a graph has a Hamiltonian path by exhibiting an example, but we don't know how to show that it has *no* Hamiltonian path that way!) *Assuming* that $NP \neq$ co$-NP$, there is a theorem that says that no co-NP problems are contained in NPC. Therefore, problems in the intersection of NP and co-NP, if not in P, are good candidates for inclusion in NPI.

In fact, a problem in $NP \cap$ co$-NP$ that is believed not in P is the FACTORING problem. As already noted, FACTORING is in NP because, if we are offered a factor of x, we can easily check its validity. But it is also in co-NP, because it is known that if we are given a prime number then (at least in principle), we can efficiently verify its primality. Thus, if someone tells us the prime factors of x, we can efficiently check that the prime factorization is right, and can *exclude* that any integer less than y is a divisor of x. Therefore, it seems likely that FACTORING is in NPI.

We are led to a crude (conjectured) picture of the structure of $NP \cup$ co$-NP$. NP and co-NP do not coincide, but they have a nontrivial intersection. P lies in $NP \cap$ co$-NP$ (because $P =$ co$-P$), but the intersection also contains problems not in P (like FACTORING). Neither NPC nor co-NPC intersects with $NP \cap$ co$-NP$.

There is much more to say about complexity theory, but we will be content to mention one more element that relates to the discussion of quantum complexity. It is sometimes useful to consider *probabilistic* circuits that have access to a random number generator. For example, a gate in a probabilistic circuit might act in either one of two ways, and flip a fair coin to decide which action to execute. Such a circuit, for a single fixed input, can sample many possible computational paths. An algorithm performed by a probabilistic circuit is said to be "randomized."

If we attack a decision problem using a probabilistic computer, we attain a probability distribution of outputs. Thus, we won't necessarily always get the right answer. But if the probability of getting the right answer is larger than $\frac{1}{2} + \delta$ for every possible input ($\delta > 0$), then the machine is useful. In fact, we can run the computation many times and use majority voting to achieve an error probability less than ε. Furthermore, the number of times we need to repeat the computation is only polylogarithmic in ε^{-1}.

If a problem admits a probabilistic circuit family of polynomial size that always gives the right answer with probability larger than $\frac{1}{2}+\delta$ (for any input, and for fixed $\delta > 0$), we say the problem is in the class BPP ("bounded-error probabilistic polynomial time"). It is evident that

$$P \subseteq BPP, \tag{6.30}$$

but the relation of NP to BPP is not known. In particular, it has not been proved that BPP is contained in NP.

6.1.3 Reversible computation

In devising a model of a quantum computer, we will generalize the circuit model of classical computation. But our quantum logic gates will be unitary transformations, and hence will be invertible, while classical logic gates like the NAND gate are not invertible. Before we discuss quantum circuits, it is useful to consider some features of reversible classical computation.

Aside from the connection with quantum computation, another incentive for studying reversible classical computation arose in Chapter 1. As Landauer observed, because irreversible logic elements erase information, they are necessarily dissipative, and therefore, require an irreducible expenditure of power. But if a computer operates reversibly, then in principle there need be no dissipation and no power requirement. We can compute for free!

A reversible computer evaluates an invertible function taking n bits to n bits

$$f : \{0,1\}^n \to \{0,1\}^n, \tag{6.31}$$

the function must be invertible so that there is a unique input for each output; then we are able in principle to run the computation backwards and recover the input from the output. Since it is a 1-1 function, we can regard it as a permutation of the 2^n strings of n bits — there are $(2^n)!$ such functions.

Of course, any irreversible computation can be "packaged" as an evaluation of an invertible function. For example, for any $f : \{0,1\}^n \to \{0,1\}^m$, we can construct $\tilde{f} : \{0,1\}^{n+m} \to \{0,1\}^{n+m}$ such that

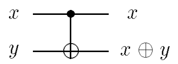
$$\tilde{f}(x; 0^{(m)}) = (x; f(x)), \tag{6.32}$$

(where $0^{(m)}$ denotes m-bits initially set to zero). Since $\tilde{f}$ takes each $(x; 0^{(m)})$ to a distinct output, it can be extended to an invertible function of $n+m$ bits. So for any f taking n bits to m, there is an invertible $\tilde{f}$ taking $n+m$ to $n+m$, which evaluates $f(x)$ acting on $(x, 0^{(m)})$

Now, how do we build up a complicated reversible computation from elementary components — that is, what constitutes a universal gate set? We will see that one-bit and two-bit reversible gates do not suffice; we will need three-bit gates for universal reversible computation.

Of the four 1-bit $\to$ 1-bit gates, two are reversible; the trivial gate and the NOT gate. Of the $(2^4)^2 = 256$ possible 2-bit $\to$ 2-bit gates, $4! = 24$ are reversible. One of special interest is the controlled-NOT or reversible XOR gate that we already encountered in Chapter 4:

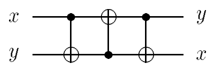
$$\text{XOR}: \ (x,y) \mapsto (x, x \oplus y), \tag{6.33}$$



This gate flips the second bit if the first is 1, and does nothing if the first bit is 0 (hence the name controlled-NOT). Its square is trivial, that is, it inverts itself. Of course, this gate performs a NOT on the second bit if the first bit is set to 1, and it performs the copy operation if y is initially set to zero:

$$\text{XOR}: \ (x,0) \mapsto (x,x). \tag{6.34}$$

With the circuit



constructed from three X0R's, we can swap two bits:

$$(x,y) \to (x, x \oplus y) \to (y, x \oplus y) \to (y,x). \tag{6.35}$$

With these swaps we can shuffle bits around in a circuit, bringing them together if we want to act on them with a particular component in a fixed location.

To see that the one-bit and two-bit gates are nonuniversal, we observe that all these gates are *linear*. Each reversible two-bit gate has an action of the form

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x' \\ y' \end{pmatrix} = \mathcal{M}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}, \tag{6.36}$$

where the constant $\binom{a}{b}$ takes one of four possible values, and the matrix $\mathcal{M}$ is one of the six invertible matrices

$$\mathcal{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}. \tag{6.37}$$

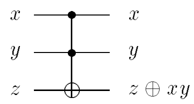
(All addition is performed modulo 2.) Combining the six choices for $\mathcal{M}$ with the four possible constants, we obtain 24 distinct gates, which exhausts all the reversible $2 \to 2$ gates.

Since the linear transformations are closed under composition, any circuit composed from reversible $2 \to 2$ (and $1 \to 1$) gates will compute a linear function

$$x \to \mathcal{M}x + a. \tag{6.38}$$

But for $n \geq 3$, there are invertible functions on n-bits that are nonlinear. An important example is the 3-bit *Toffoli gate* (or controlled-controlled-NOT) $\theta^{(3)}$

$$\theta^{(3)}: (x,y,z) \to (x,y,z \oplus xy); \tag{6.39}$$



it flips the third bit if the first two are 1 and does nothing otherwise. Like the XOR gate, it is its own inverse.

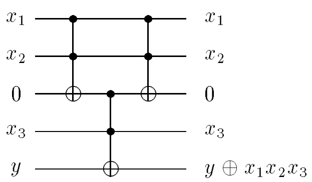
Unlike the reversible 2-bit gates, the Toffoli gate serves as a universal gate for Boolean logic, if we can provide fixed input bits and ignore output bits. If z is initially 1, then $x \uparrow y = 1 - xy$ appears in the third output — we can perform NAND. If we fix $x = 1$, the Toffoli gate functions like an XOR gate, and we can use it to copy.

The Toffoli gate $\theta^{(3)}$ is universal in the sense that we can build a circuit to compute any reversible function using Toffoli gates alone (if we can fix input bits and ignore output bits). It will be instructive to show this directly, without relying on our earlier argument that NAND/NOT is universal for Boolean functions. In fact, we can show the following: From the NOT gate and the Toffoli gate $\theta^{(3)}$, we can construct any invertible function on n bits, provided we have one extra bit of scratchpad space available.

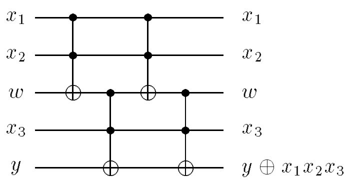
The first step is to show that from the three-bit Toffoli-gate $\theta^{(3)}$ we can construct an n-bit Toffoli gate $\theta^{(n)}$ that acts as

$$(x_1, x_2, \ldots x_{n-1}, y) \to (x_1, x_2, \ldots, x_{n-1} y \oplus x_1 x_2 \ldots x_{n-1}). \tag{6.40}$$

The construction requires one extra bit of scratch space. For example, we construct $\theta^{(4)}$ from $\theta^{(3)}$'s with the circuit



The purpose of the last $\theta^{(3)}$ gate is to reset the scratch bit back to its original value zero. Actually, with one more gate we can obtain an implementation of $\theta^{(4)}$ that works irrespective of the initial value of the scratch bit:

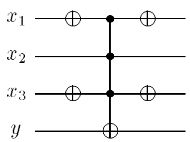


Again, we can eliminate the last gate if we don't mind flipping the value of the scratch bit.

We can see that the scratch bit really is necessary, because $\theta^{(4)}$ is an odd permutation (in fact a transposition) of the 24 4-bit strings — it transposes 1111 and 1110. But $\theta^{(3)}$ acting on any three of the four bits is an even permutation; *e.g.*, acting on the last three bits it transposes 0111 with 0110,

and 1111 with 1110. Since a product of even permutations is also even, we cannot obtain $\theta^{(4)}$ as a product of $\theta^{(3)}$'s that act on four bits only.

The construction of $\theta^{(4)}$ from four $\theta^{(3)}$'s generalizes immediately to the construction of $\theta^{(n)}$ from two $\theta^{(n-1)}$'s and two $\theta^{(3)}$'s (just expand x_1 to several control bits in the above diagram). Iterating the construction, we obtain $\theta^{(n)}$ from a circuit with $2^{n-2}+2^{n-3}-2$ $\theta^{(3)}$'s. Furthermore, just one bit of scratch space is sufficient.[2]) (When we need to construct $\theta^{(k)}$, any available extra bit will do, since the circuit returns the scratch bit to its original value. The next step is to note that, by conjugating $\theta^{(n)}$ with NOT gates, we can in effect modify the value of the control string that "triggers" the gate. For example, the circuit



flips the value of y if $x_1 x_2 x_3 = 010$, and it acts trivially otherwise. Thus this circuit transposes the two strings 0100 and 0101. In like fashion, with $\theta^{(n)}$ and NOT gates, we can devise a circuit that transposes any two n-bit strings that differ in only one bit. (The location of the bit where they differ is chosen to be the *target* of the $\theta^{(n)}$ gate.)

But in fact a transposition that exchanges any two n-bit strings can be expressed as a product of transpositions that interchange strings that differ in only one bit. If a_0 and a_s are two strings that are Hamming distance s apart (differ in s places), then there is a chain

$$a_0, a_1, a_2, a_3, \ldots, a_s, \tag{6.41}$$

such that each string in the chain is Hamming distance one from its neighbors. Therefore, each of the transpositions

$$(a_0 a_1), (a_1 a_2), (a_2 a_3), \ldots (a_{s-1} a_s), \tag{6.42}$$

[2]With more scratch space, we can build $\theta^{(n)}$ from $\theta^{(3)}$'s much more efficiently — see the exercises.

can be implemented as a $\theta^{(n)}$ gate conjugated by NOT gates. By composing transpositions we find

$$(a_0 a_s) = (a_{s-1}a_s)(a_{s-2}a_{s-1})\ldots(a_2a_3)(a_1a_2)(a_0a_1)(a_1a_2)(a_2a_3)$$
$$\ldots(a_{s-2}a_{s-1})(a_{s-1}a_s); \qquad (6.43)$$

we can construct the Hamming-distance-s transposition from $2s-1$ Hamming-distance-one transpositions. It follows that we can construct (a_0a_s) from $\theta^{(n)}$'s and NOT gates.

Finally, since every permutation is a product of transpositions, we have shown that every invertible function on n-bits (every permutation on n-bit strings) is a product of $\theta^{(3)}$'s and NOT's, using just one bit of scratch space.

Of course, a NOT can be performed with a $\theta^{(3)}$ gate if we fix two input bits at 1. Thus the Toffoli gate $\theta^{(3)}$ is universal for reversible computation, if we can fix input bits and discard output bits.

6.1.4 Billiard ball computer

Two-bit gates suffice for universal irreversible computation, but three-bit gates are needed for universal reversible computation. One is tempted to remark that "three-body interactions" are needed, so that building reversible hardware is more challenging than building irreversible hardware. However, this statement may be somewhat misleading.

Fredkin described how to devise a universal reversible computer in which the fundamental interaction is an elastic collision between two billiard balls. Balls of radius $\frac{1}{\sqrt{2}}$ move on a square lattice with unit lattice spacing. At each integer valued time, the center of each ball lies at a lattice site; the presence or absence of a ball at a particular site (at integer time) encodes a bit of information. In each unit of time, each ball moves unit distance along one of the lattice directions. Occasionally, at integer-valued times, 90^o elastic collisions occur between two balls that occupy sites that are distance $\sqrt{2}$ apart (joined by a lattice diagonal).

The device is programmed by nailing down balls at certain sites, so that those balls act as perfect reflectors. The program is executed by fixing initial positions and directions for the moving balls, and evolving the system according to Newtonian mechanics for a finite time. We read the output by observing the final positions of all the moving balls. The collisions are nondissipative, so that we can run the computation backward by reversing the velocities of all the balls.

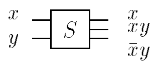
To show that this machine is a universal reversible computer, we must explain how to operate a universal gate. It is convenient to consider the three-bit *Fredkin gate*

$$(x, y, z) \rightarrow (x, xz + \bar{x}y, xy + \bar{x}z), \qquad (6.44)$$

which swaps y and z if $x = 0$ (we have introduced the notation $\bar{x} = \neg x$). You can check that the Fredkin gate can simulate a NAND/NOT gate if we fix inputs and ignore outputs.

We can build the Fredkin gate from a more primitive object, the *switch gate*. A switch gate taking two bits to three acts as

$$(x, y) \rightarrow (x, xy, \bar{x}y). \qquad (6.45)$$



The gate is "reversible" in that we can run it backwards acting on a constrained 3-bit input taking one of the four values

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad (6.46)$$

Furthermore, the switch gate is itself universal; fixing inputs and ignoring outputs, it can do NOT ($y = 1$, third output) AND (second output), and COPY ($y = 1$, first and second output). It is not surprising, then, that we can compose switch gates to construct a universal reversible $3 \rightarrow 3$ gate. Indeed, the circuit

builds the Fredkin gate from four switch gates (two running forward and two running backward). Time delays needed to maintain synchronization are not explicitly shown.

In the billiard ball computer, the switch gate is constructed with two reflectors, such that (in the case $x = y = 1$) two moving balls collide twice. The trajectories of the balls in this case are:

A ball labeled x emerges from the gate along the same trajectory (and at the same time) regardless of whether the other ball is present. But for $x = 1$, the position of the other ball (if present) is shifted down compared to its final position for $x = 0$ — this is a switch gate. Since we can perform a switch gate, we can construct a Fredkin gate, and implement universal reversible logic with a billiard ball computer.

An evident weakness of the billiard-ball scheme is that initial errors in the positions and velocities of the ball will accumulate rapidly, and the computer will eventually fail. As we noted in Chapter 1 (and Landauer has insistently pointed out) a similar problem will afflict any proposed scheme for dissipationless computation. To control errors we must be able to compress the phase space of the device, which will necessarily be a dissipative process.

6.1.5 Saving space

But even aside from the issue of error control there is another key question about reversible computation. How do we manage the scratchpad space needed to compute reversibly?

In our discussion of the universality of the Toffoli gate, we saw that in principle we can do any reversible computation with very little scratch space. But in practice it may be impossibly difficult to figure out how to do a particular computation with minimal space, and in any case economizing on space may be costly in terms of the run time.

There is a general strategy for simulating an irreversible computation on a reversible computer. Each irreversible NAND or COPY gate can be simulated by a Toffoli gate by fixing inputs and ignoring outputs. We accumulate and save all "garbage" output bits that are needed to reverse the steps of the computation. The computation proceeds to completion, and then a copy of the output is generated. (This COPY operation is logically reversible.) Then the computation runs in reverse, cleaning up all garbage bits, and returning all registers to their original configurations. With this procedure the reversible circuit runs only about twice as long as the irreversible circuit that it simulates, and all garbage generated in the simulation is disposed of without any dissipation and hence no power requirement.

This procedure works, but demands a huge amount of scratch space – the space needed scales linearly with the length T of the irreversible computation being simulated. In fact, it is possible to use space far more efficiently (with only a minor slowdown), so that the space required scales like $\log T$ instead

of T. (That is, there is a general-purpose scheme that requires space $\propto \log T$; of course, we might do even better in the simulation of a particular computation.)

To use space more effectively, we will divide the computation into smaller steps of roughly equal size, and we will run these steps backward when possible during the course of the computation. However, just as we are unable to perform step k of the computation unless step $k-1$ has already been completed, we are unable to run step k *in reverse* if step $k-1$ has previously been executed in reverse.[3] The amount of space we require (to store our garbage) will scale like the maximum value of the number of forward steps minus the number of backward steps that have been executed.

The challenge we face can be likened to a game — the *reversible pebble game*.[4] The steps to be executed form a one-dimension directed graph with sites labeled $1, 2, 3 \ldots T$. Execution of step k is modeled by placing a pebble on the kth site of the graph, and executing step k in reverse is modeled as removal of a pebble from site k. At the beginning of the game, no sites are covered by pebbles, and in each turn we add or remove a pebble. But we cannot place a pebble at site k (except for $k = 1$) unless site $k-1$ is already covered by a pebble, and we cannot remove a pebble from site k (except for $k = 1$) unless site $k-1$ is covered. The object is to cover site T (complete the computation) without using more pebbles than necessary (generating a minimal amount of garbage).

In fact, with n pebbles we can reach site $T = 2^n - 1$, but we can go no further.

We can construct a recursive procedure that enables us to reach site $T = 2^{n-1}$ with n pebbles, leaving only one pebble in play. Let $F_1(k)$ denote placing a pebble at site k, and $F_1(k)^{-1}$ denote removing a pebble from site k. Then

$$F_2(1, 2) = F_1(1)F_1(2)F_1(1)^{-1}, \tag{6.47}$$

leaves a pebble at site $k = 2$, using a maximum of two pebbles at intermediate

[3]We make the conservative assumption that we are not clever enough to know ahead of time what portion of the output from step $k-1$ might be needed later on. So we store a complete record of the configuration of the machine after step $k-1$, which is not to be erased until an updated record has been stored after the completion of a subsequent step.

[4]as pointed out by Bennett. For a recent discussion, see M. Li and P. Vitanyi, `quant-ph/9703022`.

stages. Similarly

$$F_3(1,4) = F_2(1,2)F_2(3,4)F_2(1,2)^{-1}, \tag{6.48}$$

reaches site $k = 4$ using a maximum of three pebbles, and

$$F_4(1,8) = F_3(1,4)F_3(5,8)F_3(1,4)^{-1}, \tag{6.49}$$

reaches $k = 8$ using four pebbles. Evidently we can construct $F_n(1, 2^{n-1})$ which uses a maximum of n pebbles and leaves a single pebble in play. (The routine

$$F_n(1, 2^{n-1})F_{n-1}(2^{n-1}+1, 2^{n-1}+2^{n-2}) \ldots F_1(2^n - 1), \tag{6.50}$$

leaves all n pebbles in play, with the maximal pebble at site $k = 2^n - 1$.)

Interpreted as a routine for executing $T = 2^{n-1}$ steps of a computation, this strategy for playing the pebble game represents a simulation requiring space scaling like $n \sim \log T$. How long does the simulation take? At each level of the recursive procedure described above, two steps forward are replaced by two steps forward and one step back. Therefore, an irreversible computation with $T_{\rm irr} = 2^n$ steps is simulated in $T_{\rm rev} = 3^n$ steps, or

$$T_{\rm rev} = (T_{\rm irr})^{\log 3/\log 2}, = (T_{\rm irr})^{1.58}, \tag{6.51}$$

a modest power law slowdown.

In fact, we can improve the slowdown to

$$T_{\rm rev} \sim (T_{\rm irr})^{1+\varepsilon}, \tag{6.52}$$

for any $\varepsilon > 0$. Instead of replacing two steps forward with two forward and one back, we replace ℓ forward with ℓ forward and $\ell - 1$ back. A recursive procedure with n levels reaches site ℓ^n using a maximum of $n(\ell - 1) + 1$ pebbles. Now we have $T_{\rm irr} = \ell^n$ and $T_{\rm rev} = (2\ell - 1)^n$, so that

$$T_{\rm rev} = (T_{\rm irr})^{\log(2\ell-1)/\log \ell}; \tag{6.53}$$

the power characterizing the slowdown is

$$\frac{\log(2\ell-1)}{\log \ell} = \frac{\log 2\ell + \log\left(1 - \frac{1}{2\ell}\right)}{\log \ell} \simeq 1 + \frac{\log 2}{\log \ell}, \tag{6.54}$$

and the space requirement scales as

$$S \simeq n\ell \simeq \ell \frac{\log T}{\log \ell}. \tag{6.55}$$

Thus, for any fixed $\varepsilon > 0$, we can attain S scaling like $\log T$, and a slowdown no worse than $(T_{\rm irr})^{1+\varepsilon}$. (This is not the optimal way to play the Pebble game if our objective is to get as far as we can with as few pebbles as possible. We use more pebbles to get to step T, but we get there faster.)

We have now seen that a reversible circuit can simulate a circuit composed of irreversible gates efficiently — without requiring unreasonable memory resources or causing an unreasonable slowdown. Why is this important? You might worry that, because reversible computation is "harder" than irreversible computation, the classification of complexity depends on whether we compute reversibly or irreversibly. But this is not the case, because a reversible computer can simulate an irreversible computer pretty easily.

6.2 Quantum Circuits

Now we are ready to formulate a mathematical model of a quantum computer. We will generalize the circuit model of classical computation to the quantum circuit model of quantum computation.

A classical computer processes bits. It is equipped with a finite set of gates that can be applied to sets of bits. A quantum computer processes qubits. We will assume that it too is equipped with a discrete set of fundamental components, called *quantum gates*. Each quantum gate is a unitary transformation that acts on a fixed number of qubits. In a quantum computation, a finite number n of qubits are initially set to the value $|00\ldots0\rangle$. A circuit is executed that is constructed from a finite number of quantum gates acting on these qubits. Finally, a Von Neumann measurement of all the qubits (or a subset of the qubits) is performed, projecting each onto the basis $\{|0\rangle, |1\rangle\}$. The outcome of this measurement is the result of the computation.

Several features of this model require comment:

(1) It is implicit but important that the Hilbert space of the device has a preferred decomposition into a tensor product of low-dimensional spaces, in this case the two-dimensional spaces of the qubits. Of course, we could have considered a tensor product of, say, qutrits instead. But

anyway we assume there is a natural decomposition into subsystems that is respected by the quantum gates — which act on only a few subsystems at a time. Mathematically, this feature of the gates is crucial for establishing a clearly defined notion of quantum complexity. Physically, the fundamental reason for a natural decomposition into subsystems is *locality*; feasible quantum gates must act in a bounded spatial region, so the computer decomposes into subsystems that interact only with their neighbors.

(2) Since unitary transformations form a continuum, it may seem unnecessarily restrictive to postulate that the machine can execute only those quantum gates chosen from a discrete set. We nevertheless accept such a restriction, because we do not want to invent a new physical implementation each time we are faced with a new computation to perform.

(3) We might have allowed our quantum gates to be superoperators, and our final measurement to be a POVM. But since we can easily simulate a superoperator by performing a unitary transformation on an extended system, or a POVM by performing a Von Neumann measurement on an extended system, the model as formulated is of sufficient generality.

(4) We might allow the final measurement to be a collective measurement, or a projection into a different basis. But any such measurement can be implemented by performing a suitable unitary transformation followed by a projection onto the standard basis $\{|0\rangle, |1\rangle\}^n$. Of course, complicated collective measurements can be transformed into measurements in the standard basis only with some difficulty, but it is appropriate to take into account this difficulty when characterizing the complexity of an algorithm.

(5) We might have allowed measurements at intermediate stages of the computation, with the subsequent choice of quantum gates conditioned on the outcome of those measurements. But in fact the same result can always be achieved by a quantum circuit with all measurements postponed until the end. (While we can postpone the measurements in principle, it might be very useful in practice to perform measurements at intermediate stages of a quantum algorithm.)

A quantum gate, being a unitary transformation, is reversible. In fact, a classical reversible computer is a special case of a quantum computer. A

classical reversible gate

$$x^{(n)} \to y^{(n)} = f(x^{(n)}), \tag{6.56}$$

implementing a permutation of n-bit strings, can be regarded as a unitary transformation that acts on the "computational basis $\{|x_i\rangle\}$ according to

$$U : |x_i\rangle \to |y_i\rangle. \tag{6.57}$$

This action is unitary because the 2^n strings $|y_i\rangle$ are all mutually orthogonal. A quantum computation constructed from such classical gates takes $|0\dots0\rangle$ to one of the computational basis states, so that the final measurement is deterministic.

There are three main issues concerning our model that we would like to address. The first issue is *universality*. The most general unitary transformation that can be performed on n qubits is an element of $U(2^n)$. Our model would seem incomplete if there were transformations in $U(2^n)$ that we were unable to reach. In fact, we will see that there are many ways to choose a discrete set of *universal quantum gates*. Using a universal gate set we can construct circuits that compute a unitary transformation that comes as close as we please to any element in $U(2^n)$.

Thanks to universality, there is also a machine independent notion of *quantum complexity*. We may define a new complexity class BQP — the class of decision problems that can be solved, with high probability, by polynomial-size quantum circuits. Since one universal quantum computer can simulate another efficiently, the class does not depend on the details of our hardware (on the universal gate set that we have chosen).

Notice that a quantum computer can easily simulate a probabilistic classical computer: it can prepare $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and then project to $\{|0\rangle, |1\rangle\}$, generating a random bit. Therefore BQP certainly contains the class BPP. But as we discussed in Chapter 1, it seems to be quite reasonable to expect that BQP is actually larger than BPP, because a probabilistic classical computer cannot easily simulate a quantum computer. The fundamental difficulty is that the Hilbert space of n qubits is huge, of dimension 2^n, and hence the mathematical description of a typical vector in the space is exceedingly complex. Our second issue is to better characterize the resources needed to simulate a quantum computer on a classical computer. We will see that, despite the vastness of Hilbert space, a classical computer can simulate an n-qubit quantum computer even if limited to an amount of memory space that is polynomial in n. This means the BQP is contained in the complexity class $PSPACE$, the decision problems that can be solved with polynomial space, but may require exponential time. (We know that NP is also contained in $PSPACE$, since checking if $C(x^{(n)}, y^{(m)}) = 1$ for each $y^{(m)}$ can be accomplished with polynomial space.[5]

The third important issue we should address is *accuracy*. The class BQP is defined formally under the idealized assumption that quantum gates can be executed with perfect precision. Clearly, it is crucial to relax this assumption in any realistic implementation of quantum computation. A polynomial size quantum circuit family that solves a hard problem would not be of much interest if the quantum gates in the circuit were required to have exponential accuracy. In fact, we will show that this is not the case. An idealized T-gate quantum circuit can be simulated with acceptable accuracy by noisy gates, provided that the error probability per gate scales like $1/T$.

We see that quantum computers pose a serious challenge to the strong Church–Turing thesis, which contends that any physically reasonable model of computation can be simulated by probabilistic classical circuits with at worst a polynomial slowdown. But so far there is no firm proof that

$$BPP \neq BQP. \tag{6.58}$$

Nor is such a proof necessarily soon to be expected.[6] Indeed, a corollary would be

$$BPP \neq PSPACE, \tag{6.59}$$

which would settle one of the long-standing and pivotal open questions in complexity theory.

It might be less unrealistic to hope for a proof that $BPP \neq BQP$ follows from another standard conjecture of complexity theory such as $P \neq NP$. So far no such proof has been found. But while we are not yet able to prove that quantum computers have capabilities far beyond those of conventional computers, we nevertheless might uncover evidence suggesting that $BPP \neq BQP$. We will see that there are problems that seem to be hard (in classical computation) yet can be efficiently solved by quantum circuits.

[5] Actually there is another rung of the complexity hierarchy that may separate BQP and $PSPACE$; we can show that $BQP \subseteq P^{\#P} \subseteq PSPACE$, but we won't consider $P^{\#P}$ any further here.

[6] That is, we ought not to expect a "nonrelativized proof." A separation between BPP and BQP "relative to an oracle" will be established later in the chapter.

Thus it seems likely that the classification of complexity will be different depending on whether we use a classical computer or a quantum computer to solve a problem. If such a separation really holds, it is the quantum classification that should be regarded as the more fundamental, for it is better founded on the physical laws that govern the universe.

6.2.1 Accuracy

Let's discuss the issue of accuracy. We imagine that we wish to implement a computation in which the quantum gates $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_T$ are applied sequentially to the initial state $|\varphi_0\rangle$. The state prepared by our ideal quantum circuit is

$$|\varphi_T\rangle = \boldsymbol{U}_T \boldsymbol{U}_{T-1} \ldots \boldsymbol{U}_2 \boldsymbol{U}_1 |\varphi_0\rangle. \tag{6.60}$$

But in fact our gates do not have perfect accuracy. When we attempt to apply the unitary transformation $\boldsymbol{U}_t$, we instead apply some "nearby" unitary transformation $\tilde{\boldsymbol{U}}_t$. (Of course, this is not the most general type of error that we might contemplate – the unitary $\boldsymbol{U}_t$ might be replaced by a *superoperator*. Considerations similar to those below would apply in that case, but for now we confine our attention to "unitary errors.")

The errors cause the actual state of the computer to wander away from the ideal state. How far does it wander? Let $|\varphi_t\rangle$ denote the ideal state after t quantum gates are applied, so that

$$|\varphi_t\rangle = \boldsymbol{U}_t |\varphi_{t-1}\rangle. \tag{6.61}$$

But if we apply the actual transformation $\tilde{\boldsymbol{U}}_t$, then

$$\tilde{\boldsymbol{U}}_t |\varphi_{t-1}\rangle = |\varphi_t\rangle + |E_t\rangle, \tag{6.62}$$

where

$$|E_t\rangle = (\tilde{\boldsymbol{U}}_t - \boldsymbol{U}_t)|\varphi_{t-1}\rangle, \tag{6.63}$$

is an unnormalized vector. If $|\tilde{\varphi}_t\rangle$ denotes the actual state after t steps, then we have

$$\begin{aligned} |\tilde{\varphi}_1\rangle &= |\varphi_1\rangle + |E_1\rangle, \\ |\tilde{\varphi}_2\rangle &= \tilde{\boldsymbol{U}}_2 |\tilde{\varphi}_1\rangle = |\varphi_2\rangle + |E_2\rangle + \tilde{\boldsymbol{U}}_2 |E_1\rangle, \end{aligned} \tag{6.64}$$

and so forth; we ultimately obtain

$$|\tilde{\varphi}_T\rangle = |\varphi_T\rangle + |E_T\rangle + \tilde{\boldsymbol{U}}_T|E_{T-1}\rangle + \tilde{\boldsymbol{U}}_T\tilde{\boldsymbol{U}}_{T-1}|E_{T-2}\rangle \\ + \ldots + \tilde{\boldsymbol{U}}_T\tilde{\boldsymbol{U}}_{T-1}\ldots\tilde{\boldsymbol{U}}_2|E_1\rangle. \quad (6.65)$$

Thus we have expressed the difference between $|\tilde{\varphi}_T\rangle$ and $|\varphi_T\rangle$ as a sum of T remainder terms. The worst case yielding the largest deviation of $|\tilde{\varphi}_T\rangle$ from $|\varphi_T\rangle$ occurs if all remainder terms line up in the same direction, so that the errors interfere constructively. Therefore, we conclude that

$$\| \, |\tilde{\varphi}_T\rangle - |\varphi_T\rangle \, \| \;\leq\; \| \, |E_T\rangle \, \| + \| \, |E_{T-1}\rangle \, \| \\ + \ldots + \| \, |E_2\rangle \, \| + \| \, |E_1\rangle \, \|, \quad (6.66)$$

where we have used the property $\| \boldsymbol{U}|E_i\rangle \| = \| \, |E_i\rangle \, \|$ for any unitary $\boldsymbol{U}$.

Let $\| \boldsymbol{A} \|_{\text{sup}}$ denote the sup norm of the operator $\boldsymbol{A}$ — that is, the maximum modulus of an eigenvalue of $\boldsymbol{A}$. We then have

$$\| \, |E_t\rangle \, \| = \| \left(\tilde{\boldsymbol{U}}_t - \boldsymbol{U}_t\right) |\varphi_{t-1}\rangle \, \| \leq \| \tilde{\boldsymbol{U}}_t - \boldsymbol{U}_t \|_{\text{sup}} \quad (6.67)$$

(since $|\varphi_{t-1}\rangle$ is normalized). Now suppose that, for each value of t, the error in our quantum gate is bounded by

$$\| \tilde{\boldsymbol{U}}_t - \boldsymbol{U}_t \|_{\text{sup}} < \varepsilon. \quad (6.68)$$

Then after T quantum gates are applied, we have

$$\| \, |\tilde{\varphi}_T\rangle - |\varphi_T\rangle \, \| < T\varepsilon; \quad (6.69)$$

in this sense, the accumulated error in the state grows linearly with the length of the computation.

The distance bounded in eq. (6.68) can equivalently be expressed as $\| \boldsymbol{W}_t - \boldsymbol{1} \|_{\text{sup}}$, where $\boldsymbol{W}_t = \tilde{\boldsymbol{U}}_t\boldsymbol{U}_t^\dagger$. Since $\boldsymbol{W}_t$ is unitary, each of its eigenvalues is a phase $e^{i\theta}$, and the corresponding eigenvalue of $\boldsymbol{W}_t - \boldsymbol{1}$ has modulus

$$|e^{i\theta} - 1| = (2 - 2\cos\theta)^{1/2}, \quad (6.70)$$

so that eq. (6.68) is the requirement that each eigenvalue satisfies

$$\cos\theta > 1 - \varepsilon^2/2, \quad (6.71)$$

(or $|\theta| \lesssim \varepsilon$, for ε small). The origin of eq. (6.69) is clear. In each time step, $|\tilde{\varphi}\rangle$ rotates relative to $|\varphi\rangle$ by (at worst) an angle of order ε, and the distance between the vectors increases by at most of order ε.

How much accuracy is good enough? In the final step of our computation, we perform an orthogonal measurement, and the probability of outcome a, in the ideal case, is

$$P(a) = |\langle a|\varphi_T\rangle|^2. \quad (6.72)$$

Because of the errors, the actual probability is

$$\tilde{P}(a) = |\langle a|\tilde{\varphi}_T\rangle|^2. \quad (6.73)$$

If the actual vector is close to the ideal vector, then the probability distributions are close, too. If we sum over an orthonormal basis $\{|a\rangle\}$, we have

$$\sum_a |\tilde{P}(a) - P(a)| \leq 2 \, \| \, |\tilde{\varphi}_T\rangle - |\varphi_T\rangle \, \|, \quad (6.74)$$

as you will show in a homework exercise. Therefore, if we keep $T\varepsilon$ fixed (and small) as T gets large, the error in the probability distribution also remains fixed. In particular, if we have designed a quantum algorithm that solves a decision problem correctly with probability greater $\frac{1}{2} + \delta$ (in the ideal case), then we can achieve success probability greater than $\frac{1}{2}$ with our noisy gates, if we can perform the gates with an accuracy $T\varepsilon < O(\delta)$. A quantum circuit family in the BQP class can really solve hard problems, as long as we can improve the accuracy of the gates linearly with the computation size T.

6.2.2 BQP $\subseteq$ PSPACE

Of course a classical computer can simulate any quantum circuit. But how much memory does the classical computer require? Naively, since the simulation of an n-qubit circuit involves manipulating matrices of size 2^n, it may seem that an amount of memory space exponential in n is needed. But we will now show that the simulation can be done to acceptable accuracy (albeit very slowly!) in polynomial space. This means that the quantum complexity class BQP is contained in the class PSPACE of problems that can be solved with polynomial space.

The object of the classical simulation is to compute the probability for each possible outcome a of the final measurement

$$\text{Prob}(a) = |\langle a|\boldsymbol{U}_T|0\rangle|^2, \quad (6.75)$$

where

$$\boldsymbol{U}_T = \boldsymbol{U}_T \boldsymbol{U}_{T-1} \ldots \boldsymbol{U}_2 \boldsymbol{U}_1, \tag{6.76}$$

is a product of T quantum gates. Each $\boldsymbol{U}_t$, acting on the n qubits, can be represented by a $2^n \times 2^n$ unitary matrix, characterized by the complex matrix elements

$$\langle y|\boldsymbol{U}_t|x\rangle, \tag{6.77}$$

where $x, y \in \{0, 1 \ldots, 2^n - 1\}$. Writing out the matrix multiplication explicitly, we have

$$\begin{aligned}\langle a|\boldsymbol{U}_T|0\rangle = \sum_{\{x_t\}} \langle a|\boldsymbol{U}_T|x_{T-1}\rangle\langle x_{T-1}|\boldsymbol{U}_{T-1}|x_{T-2}\rangle \ldots \\ \ldots \langle x_2|\boldsymbol{U}_2|x_1\rangle\langle x_1|\boldsymbol{U}_1|0\rangle.\end{aligned} \tag{6.78}$$

Eq. (6.78) is a sort of "path integral" representation of the quantum computation – the probability amplitude for the final outcome a is expressed as a coherent sum of amplitudes for each of a vast number $(2^{n(T-1)})$ of possible computational paths that begin at 0 and terminate at a after T steps.

Our classical simulator is to add up the $2^{n(T-1)}$ complex numbers in eq. (6.78) to compute $\langle a|\boldsymbol{U}_T|0\rangle$. The first problem we face is that finite size classical circuits do integer arithmetic, while the matrix elements $\langle y|\boldsymbol{U}_t|x\rangle$ need not be rational numbers. The classical simulator must therefore settle for an approximate calculation to reasonable accuracy. Each term in the sum is a product of T complex factors, and there are $2^{n(T-1)}$ terms in the sum. The accumulated errors are sure to be small if we express the matrix elements to m bits of accuracy, with m large compared to $n(T-1)$. Therefore, we can replace each complex matrix element by pairs of signed integers, taking values in $\{0, 1, 2, \ldots, 2^{m-1}\}$. These integers give the binary expansion of the real and imaginary part of the matrix element, expressed to precision 2^{-m}.

Our simulator will need to compute each term in the sum eq. (6.78) and accumulate a total of all the terms. But each addition requires only a modest amount of scratch space, and furthermore, since only the accumulated subtotal need be stored for the next addition, not much space is needed to sum all the terms, even though there are exponentially many.

So it only remains to consider the evaluation of a typical term in the sum, a product of T matrix elements. We will require a classical circuit that

evaluates

$$\langle y|\boldsymbol{U}_t|x\rangle; \tag{6.79}$$

this circuit accepts the $2n$ bit input (x, y), and outputs the $2m$-bit value of the (complex) matrix element. Given a circuit that performs this function, it will be easy to build a circuit that multiplies the complex numbers together without using much space.

Finally, at this point, we appeal to the properties we have demanded of our quantum gate set — the gates from a discrete set, and each gate acts on a bounded number of qubits. Because there are a fixed (and finite) number of gates, there are only a fixed number of gate subroutines that our simulator needs to be able to call. And because the gate acts on only a few qubits, nearly all of its matrix elements vanish (when n is large), and the value $\langle y|\boldsymbol{U}|x\rangle$ can be determined (to the required accuracy) by a simple circuit requiring little memory.

For example, in the case of a single-qubit gate acting on the first qubit, we have

$$\langle y_1 y_2 \ldots y_n|\boldsymbol{U}|x_1 x_2 \ldots x_n\rangle = 0 \text{ if } x_2 x_3 \ldots x_n \neq y_2 y_3 \ldots y_n. \tag{6.80}$$

A simple circuit can compare x_2 with y_2, x_3 with y_3, *etc.*, and output zero if the equality is not satisfied. In the event of equality, the circuit outputs one of the four complex numbers

$$\langle y_1|\boldsymbol{U}|x_1\rangle, \tag{6.81}$$

to m bits of precision. A simple circuit can encode the $8m$ bits of this 2×2 complex-valued matrix. Similarly, a simple circuit, requiring only space polynomial in n and m, can evaluate the matrix elements of any gate of fixed size.

We conclude that a classical computer with space bounded above by poly(n) can simulate an n-qubit universal quantum computer, and therefore that BQP $\subseteq$ PSPACE. Of course, it is also evident that the simulation we have described requires exponential time, because we need to evaluate the sum of $2^{n(T-1)}$ complex numbers. (Actually, most of the terms vanish, but there are still an exponentially large number of nonvanishing terms.)

6.2.3 Universal quantum gates

We must address one more fundamental question about quantum computation; how do we construct an adequate set of quantum gates? In other words, what constitutes a universal quantum computer?

We will find a pleasing answer. Any generic two-qubit gate suffices for universal quantum computation. That is, for all but a set of measure zero of 4×4 unitary matrices, if we can apply that matrix to any pair of qubits, then we can construct an n-qubit circuit that computes a transformation that comes as close as we please to any element of $U(2^n)$.

Mathematically, this is not a particularly deep result, but physically it is very interesting. It means that, in the quantum world, as long as we can devise a generic interaction between two qubits, and we can implement that interaction accurately between any two qubits, we can compute anything, no matter how complex. Nontrivial computation is ubiquitous in quantum theory.

Aside from this general result, it is also of some interest to exhibit particular universal gate sets that might be particularly easy to implement physically. We will discuss a few examples.

There are a few basic elements that enter the analysis of any universal quantum gate set.

(1) Powers of a generic gate

Consider a "generic" k-qubit gate. This is a $2^k \times 2^k$ unitary matrix $\boldsymbol{U}$ with eigenvalues $e^{i\theta_1}, e^{i\theta_2}, \ldots e^{i\theta_{2^k}}$. For all but a set of measure zero of such matrices, each θ_i is an irrational multiple of π, and all the θ_i's are incommensurate (each θ_i/θ_j is also irrational). The positive integer power $\boldsymbol{U}^n$ of $\boldsymbol{U}$ has eigenvalues

$$e^{in\theta_1}, e^{in\theta_2}, \ldots, e^{in\theta_{2^k}}. \tag{6.82}$$

Each such list of eigenvalues defines a point in a 2^k-dimensional torus (the product of 2^k circles). As n ranges over positive integer values, these points densely fill the whole torus, if $\boldsymbol{U}$ is generic. If $\boldsymbol{U} = e^{iA}$, positive integer powers of $\boldsymbol{U}$ come as close as we please to $\boldsymbol{U}(\lambda) = e^{i\lambda A}$, for any real λ. We say that any $\boldsymbol{U}(\lambda)$ is *reachable* by positive integer powers of $\boldsymbol{U}$.

(2) Switching the leads

There are a few (classical) transformations that we can implement just by switching the labels on k qubits, or in other words, by applying the gate $\boldsymbol{U}$ to the qubits in a different order. Of the $(2^k)!$ permutations of the length-k strings, $k!$ can be realized by swapping qubits. If a gate applied to k qubits with a standard ordering is $\boldsymbol{U}$, and P is a permutation implemented by swapping qubits, then we can construct the gate

$$\boldsymbol{U}' = P\boldsymbol{U}P^{-1}, \tag{6.83}$$

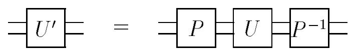
just by switching the leads on the gate. For example, swapping two qubits implements the transposition

$$P : |01\rangle \leftrightarrow |10\rangle, \tag{6.84}$$

or

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{6.85}$$

acting on basis $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$. By switching leads, we obtain a gate



We can also construct any positive integer power of $\boldsymbol{U}'$, $(P\boldsymbol{U}P^{-1})^n = P\boldsymbol{U}^nP^{-1}$.

(3) Completing the Lie algebra

We already remarked that if $\boldsymbol{U} = e^{iA}$ is generic, then powers of $\boldsymbol{U}$ are dense in the torus $\{e^{i\lambda A}\}$. We can further argue that if $\boldsymbol{U} = e^{iA}$ and $\boldsymbol{U}' = e^{iB}$ are generic gates, we can compose them to come arbitrarily close to

$$e^{i(\alpha A + \beta B)} \text{ or } e^{-\gamma[A,B]}, \tag{6.86}$$

for any real α, β, γ. Thus, the "reachable" transformations have a closed *Lie algebra*. We say that $\boldsymbol{U} = e^{iA}$ is generated by A; then if A and B are both generic generators of reachable transformations, so are real linear combinations of A and B, and (i times) the commutator of A and B.

We first note that

$$\lim_{n\to\infty}(e^{i\alpha A/n}e^{i\beta B/n})^n = \lim_{n\to\infty}\left(1+\frac{i}{n}(\alpha A+\beta B)\right)^n = e^{i(\alpha A+\beta B)}. \tag{6.87}$$

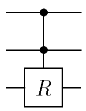
Therefore, any $e^{i(\alpha A+\beta B)}$ is reachable if each $e^{i\alpha A/n}$ and $e^{i\beta B/n}$ is. Furthermore

$$\lim_{n\to\infty}\left(e^{iA/\sqrt{n}}e^{iB/\sqrt{n}}e^{-iA/\sqrt{n}}e^{-iB/\sqrt{n}}\right)^n = \lim_{n\to\infty}\left[1-\frac{1}{n}(AB-BA)\right]^n = e^{-[A,B]}, \tag{6.88}$$

so $e^{-[A,B]}$ is also reachable.

By invoking the observations (1), (2), and (3) above, we will be able to show that a generic two-qubit gate is universal.

Deutsch gate. It was David Deutsch (1989) who first pointed out the existence of a universal quantum gate. Deutsch's three-bit universal gate is a quantum cousin of the Toffoli gate. It is the controlled-controlled-$\boldsymbol{R}$ transformation



that applies $\boldsymbol{R}$ to the third qubit if the first two qubits have the value 1; otherwise it acts trivially. Here

$$\boldsymbol{R} = -i\boldsymbol{R}_x(\theta) = (-i)\exp\left(i\frac{\theta}{2}\boldsymbol{\sigma}_x\right) = (-i)\left(\cos\frac{\theta}{2}+i\boldsymbol{\sigma}_x\sin\frac{\theta}{2}\right) \tag{6.89}$$

is, up to a phase, a rotation by θ about the x-axis, where θ is a particular angle incommensurate with π.

The nth power of the Deutsch gate is the controlled-controlled-$\boldsymbol{R}^n$. In particular, $\boldsymbol{R}^4 = \boldsymbol{R}_x(4\theta)$, so that all one-qubit transformations generated by $\boldsymbol{\sigma}_x$ are reachable by integer powers of $\boldsymbol{R}$. Furthermore the $(4n+1)$st power is

$$(-i)\left[\cos\frac{(4n+1)\theta}{2}+i\boldsymbol{\sigma}_x\sin\frac{(4n+1)\theta}{2}\right], \tag{6.90}$$

which comes as close as we please to $\boldsymbol{\sigma}_x$. Therefore, the Toffoli gate is reachable with integer powers of the Deutsch gate, and the Deutsch gate is universal for classical computation.

Acting on the three-qubit computational basis

$$\{|000\rangle, |001\rangle, |010\rangle, |011\rangle, |100\rangle, |101\rangle, |110\rangle, |111\rangle\}, \tag{6.91}$$

the generator of the Deutsch gate transposes the last two elements

$$|110\rangle \leftrightarrow |111\rangle. \tag{6.92}$$

We denote this 8×8 matrix as

$$(\boldsymbol{\sigma}_x)_{67} = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \boldsymbol{\sigma}_x \end{array}\right). \tag{6.93}$$

With Toffoli gates, we can perform any permutation of these eight elements, in particular

$$P = (6m)(7n), \tag{6.94}$$

for any m and n. So we can also reach any transformation generated by

$$P(\boldsymbol{\sigma}_x)_{67}P = (\boldsymbol{\sigma}_x)_{mn}. \tag{6.95}$$

Furthermore,

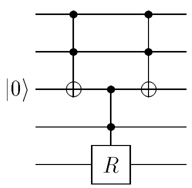
$$[(\boldsymbol{\sigma}_x)_{56}, (\boldsymbol{\sigma}_x)_{67}] = \left[\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\right] = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} = i(\boldsymbol{\sigma}_y)_{57}, \tag{6.96}$$

and similarly, we can reach any unitary generated by $(\boldsymbol{\sigma}_y)_{mn}$. Finally

$$[(\boldsymbol{\sigma}_x)_{mn}, (\boldsymbol{\sigma}_y)_{mn}] = i(\boldsymbol{\sigma}_z)_{mn}, \tag{6.97}$$

So we can reach any transformation generated by a linear combination of the $(\boldsymbol{\sigma}_{x,y,z})_{mn}$'s. These span the whole $SU(8)$ Lie Algebra, so we can generate any three-qubit unitary (aside from an irrelevant overall phase).

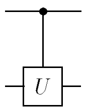
Now recall that we have already found that we can construct the n-bit Toffoli gate by composing three-bit Toffoli gates. The circuit



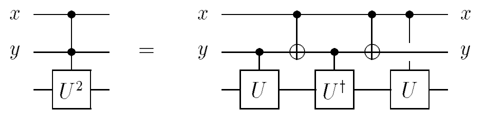
uses one scratch bit to construct a four-bit Deutsch gate ((controlled)^3-$\boldsymbol{R}$) from the three-bit Deutsch gate and two three-bit Toffoli gates, and a similar circuit constructs the n-bit Deutsch gate from a three-bit Deutsch gate and two $(n-1)$-bit Toffoli gates. Once we have an n-bit Deutsch gate, and universal classical computation, exactly the same argument as above shows that we can reach any transformation in $SU(2^n)$.

Universal two-qubit gates. For reversible classical computation, we saw that three-bit gates are needed for universality. But in quantum computation, two-bit gates turn out to be adequate. Since we already know that the Deutsch gate is universal, we can establish this by showing that the Deutsch gate can be constructed by composing two-qubit gates.

In fact, if



denotes the controlled-$\boldsymbol{U}$ gate (the 2×2 unitary $\boldsymbol{U}$ is applied to the second qubit if the first qubit is 1; otherwise the gate acts trivially) then a controlled-controlled-$\boldsymbol{U}^2$ gate is obtained from the circuit



the power of $\boldsymbol{U}$ applied to the third qubit is

$$y - (x \oplus y) + x = x + y - (x + y - 2xy) = 2xy. \tag{6.98}$$

Therefore, we can construct Deutsch's gate from the controlled-$\boldsymbol{U}$, controlled $\boldsymbol{U}^{-1}$ and controlled-NOT gates, where

$$\boldsymbol{U}^2 = -i\boldsymbol{R}_x(\theta); \tag{6.99}$$

we may choose

$$\boldsymbol{U} = e^{-i\frac{\pi}{4}}\boldsymbol{R}_x\left(\frac{\theta}{2}\right). \tag{6.100}$$

Positive powers of $\boldsymbol{U}$ came as close as we please to $\boldsymbol{\sigma}_x$ and $\boldsymbol{U}^{-1}$, so from the controlled-$\boldsymbol{U}$ alone we can construct the Deutsch gate. Therefore, the controlled-$\left(e^{-i\frac{\pi}{4}}\boldsymbol{R}_x\left(\frac{\theta}{2}\right)\right)$ is itself a universal gate, for θ/π irrational.

(Note that the above construction shows that, while we cannot construct the Toffoli gate from two-bit reversible classical gates, we *can* construct it from a controlled "square root of NOT" — a controlled-$\boldsymbol{U}$ with $\boldsymbol{U}^2 = \boldsymbol{\sigma}_x$.)

Generic two-bit gates. Now we have found particular two-bit gates (controlled rotations) that are universal gates. Therefore, for universality, it is surely sufficient if we can construct transformations that are dense in the $U(4)$ acting on a pair of qubits.

In fact, though, any generic two-qubit gate is sufficient to generate all of $U(4)$. As we have seen, if $e^{i\boldsymbol{A}}$ is a generic element of $U(4)$, we can reach any transformation generated by $\boldsymbol{A}$. Furthermore, we can reach any transformations generated by an element of the minimal Lie algebra containing $\boldsymbol{A}$ and

$$\boldsymbol{B} = P\boldsymbol{A}P^{-1} \tag{6.101}$$

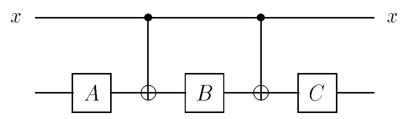
where P is the permutation ($|01\rangle \leftrightarrow |10\rangle$) obtained by switching the leads.

Now consider a general $\boldsymbol{A}$, (expanded in terms of a basis for the Lie algebra of $U(4)$), and consider a particular scheme for constructing 16 elements of the algebra by successive commutations, starting from $\boldsymbol{A}$ and $\boldsymbol{B}$. The elements so constructed are linearly independent (and it follows that any transformation in $U(4)$ is reachable) if the determinant of a particular 16×16 matrix vanishes. Unless this vanishes identically, its zeros occur only on a submanifold of vanishing measure. But in fact, we can choose, say

$$\boldsymbol{A} = (\alpha I + \beta\sigma_x + \gamma\sigma_y)_{23}, \tag{6.102}$$

(for incommensurate α, β, γ), and show by explicit computation that the entire 16-dimension Lie Algebra is actually generated by successive commutations, starting with $\boldsymbol{A}$ and $\boldsymbol{B}$. Hence we conclude that failure to generate the entire $U(4)$ algebra is nongeneric, and find that almost all two-qubit gates are universal.

Other adequate sets of gates. One can also see that universal quantum computation can be realized with a gate set consisting of *classical* multi-qubit gates and quantum single-qubit gates. For example, we can see that the XOR gate, combined with one-qubit gates, form a universal set. Consider the circuit



which applies $\boldsymbol{ABC}$ to the second qubit if $x = 0$, and $\boldsymbol{A\sigma_x B\sigma_x C}$ to the second qubit if $x = 1$. If we can find $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ such that

$$\begin{aligned} \boldsymbol{ABC} &= \mathbf{1} \\ \boldsymbol{A\sigma_x B\sigma_x C} &= \boldsymbol{U}, \end{aligned} \tag{6.103}$$

then this circuit functions as a controlled-$\boldsymbol{U}$ gate. In fact unitary 2×2 $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ with this property exist for any unitary $\boldsymbol{U}$ with determinant one (as you'll show in an exercise). Therefore, the XOR plus arbitrary one-qubit transformations form a universal set. Of course, two generic (noncommuting) one-qubit transformations are sufficient to reach all. In fact, with an XOR

and a *single* generic one-qubit rotation, we can construct a second one-qubit rotation that does not commute with the first. Hence, an XOR together with just one generic single-qubit gate constitutes a universal gate set.

If we are able to perform a Toffoli gate, then even certain nongeneric one-qubit transformations suffice for universal computation. For example (another exercise) the Toffoli gate, together with $\pi/2$ rotations about the x and z axes, are a universal set.

Precision. Our discussion of universality has focused on *reachability* without any regard for *complexity*. We have only established that we can construct a quantum circuit that comes as close as we please to a desired element of $U(2^n)$, and we have not considered the size of the circuit that we need. But from the perspective of quantum complexity theory, universality is quite significant because it implies that one quantum computer can simulate another to reasonable accuracy without an unreasonable slowdown.

Actually, we have not been very precise up until now about what it means for one unitary transformation to be "close" to another; we should define a topology. One possibility is to use the sup norm as in our previous discussion of accuracy — the distance between matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ is then $\| \boldsymbol{U} - \boldsymbol{W} \|_{\text{sup}}$. Another natural topology is associated with the inner product

$$\langle \boldsymbol{W} | \boldsymbol{U} \rangle \equiv \text{tr } \boldsymbol{W}^\dagger \boldsymbol{U} \tag{6.104}$$

(if $\boldsymbol{U}$ and $\boldsymbol{W}$ are $N \times N$ matrices, this is just the usual inner product on C^{N^2}, where we regard $\boldsymbol{U}_{ij}$ as a vector with N^2 components). Then we may define the distance squared between matrices as

$$\| \boldsymbol{U} - \boldsymbol{W} \|^2 \equiv \langle \boldsymbol{U} - \boldsymbol{W} | \boldsymbol{U} - \boldsymbol{W} \rangle. \tag{6.105}$$

For the purpose of analyzing complexity, just about any reasonable topology will do.

The crucial point is that given any universal gate set, we can reach within distance ε of any desired unitary transformation that acts on a fixed number of qubits, using a quantum circuit whose size is bounded above by a polynomial in ε^{-1}. Therefore, one universal quantum computer can simulate another, to accuracy ε, with a slowdown no worse than a factor that is polynomial in ε^{-1}. Now we have already seen that to have a high probability of getting the right answer when we perform a quantum circuit of size T, we should implement each quantum gate to an accuracy that scales like T^{-1}. Therefore, if you have a quantum circuit family of polynomial size that runs on your quantum computer, I can devise a polynomial size circuit family that runs on my machine, and that emulates your machine to acceptable accuracy.

Why can a poly(ε^{-1})-size circuit reach a given k-qubit $\boldsymbol{U}$ to within distance ε? We know for example that the positive integer powers of a generic k-qubit $e^{i\boldsymbol{A}}$ are dense in the 2^k-torus $\{e^{i\lambda\boldsymbol{A}}\}$. The region of the torus within distance ε of any given point has volume of order ε^{2^k}, so (asymptotically for ε sufficiently small) we can reach any $\{e^{i\lambda\boldsymbol{A}}\}$ to within distance ε with $\left(e^{i\lambda\boldsymbol{A}}\right)^n$, for some integer n of order ε^{-2^k}. We also know that we can obtain transformations $\{e^{i\boldsymbol{A}_a}\}$ where the $\boldsymbol{A}_a$'s span the full $U(2^k)$ Lie algebra, using circuits of fixed size (independent of ε). We may then approach any $\exp\left(i\sum_a \alpha_a \boldsymbol{A}_a\right)$ as in eq. (6.87), also with polynomial convergence.

In principle, we should be able to do much better, reaching a desired k-qubit unitary within distance ε using just poly($\log(\varepsilon^{-1})$) quantum gates. Since the number of size-T circuits that we can construct acting on k qubits is exponential in T, and the circuits fill $U(2^k)$ roughly uniformly, there should be a size-T circuit reaching within a distance of order e^{-T} of any point in $U(2^k)$. However, it might be a computationally hard problem *classically* to work out the circuit that comes exponentially close to the unitary we are trying to reach. Therefore, it would be dishonest to rely on this more efficient construction in an asymptotic analysis of quantum complexity.

6.3 Some Quantum Algorithms

While we are not yet able to show that $BPP \neq BQP$, there are three approaches that we can pursue to study the differences between the capabilities of classical and quantum computers:

(1) Nonexponential speedup. We can find quantum algorithms that are demonstrably faster than the best classical algorithm, but not *exponentially* faster. These algorithms shed no light on the conventional classification of complexity. But they do demonstrate a type of separation between tasks that classical and quantum computers can perform. Example: Grover's quantum speedup of the search of an unsorted data base.

(2) "Relativized" exponential speedup. We can consider the problem of analyzing the contents of a "quantum black box." The box performs an

a priori unknown) unitary transformation. We can prepare an input for the box, and we can measure its output; our task is to find out what the box does. It is possible to prove that quantum black boxes (computer scientists call them oracles[7]) exist with this property: By feeding quantum superpositions to the box, we can learn what is inside with an *exponential* speedup, compared to how long it would take if we were only allowed classical inputs. A computer scientist would say that $BPP \neq BQP$ "relative to the oracle." Example: Simon's exponential quantum speedup for finding the period of a 2 to 1 function.

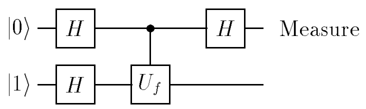
(3) Exponential speedup for "apparently" hard problems. We can exhibit a quantum algorithm that solves a problem in polynomial time, where the problem appears to be hard classically, so that it is strongly suspected (though not proved) that the problem is not in BPP. Example: Shor's factoring algorithm.

Deutsch's problem. We will discuss examples from all three approaches. But first, we'll warm up by recalling an example of a simple quantum algorithm that was previously discussed in §1.5: Deutsch's algorithm for distinguishing between constant and balanced functions $f : \{0,1\} \to \{0,1\}$. We are presented with a quantum black box that computes $f(x)$; that is, it enacts the two-qubit unitary transformation

$$U_f : |x\rangle|y\rangle \to |x\rangle|y \oplus f(x)\rangle, \tag{6.106}$$

which flips the second qubit iff f(first qubit) $= 1$. Our assignment is to determine whether $f(0) = f(1)$. If we are restricted to the "classical" inputs $|0\rangle$ and $|1\rangle$, we need to access the box twice ($x = 0$ and $x = 1$) to get the answer. But if we are allowed to input a coherent superposition of these "classical" states, then once is enough.

The quantum circuit that solves the problem (discussed in §1.5) is:



[7]The term "oracle" signifies that the box responds to a query *immediately*; that is, the time it takes the box to operate is not included in the complexity analysis.

Here H denotes the Hadamard transform

$$\boldsymbol{H}: |x\rangle \rightarrow \frac{1}{\sqrt{2}}\sum_y (-1)^{xy}|y\rangle, \tag{6.107}$$

or

$$\begin{aligned}\boldsymbol{H}: |0\rangle &\rightarrow \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\\ |1\rangle &\rightarrow \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle);\end{aligned} \tag{6.108}$$

that is, $\boldsymbol{H}$ is the 2×2 matrix

$$\boldsymbol{H}: \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}. \tag{6.109}$$

The circuit takes the input $|0\rangle|1\rangle$ to

$$\begin{aligned}|0\rangle|1\rangle \rightarrow& \frac{1}{2}(|0\rangle + |1\rangle)(|0\rangle - |1\rangle)\\ \rightarrow& \frac{1}{2}\left((-1)^{f(0)}|0\rangle + (-1)^{f(1)}|1\rangle\right)(|0\rangle - |1\rangle)\\ \rightarrow& \frac{1}{2}\left[\left((-1)^{f(0)} + (-1)^{f(1)}\right)|0\rangle\right.\\ &\left. + \left((-1)^{f(0)} - (-1)^{f(1)}\right)|1\rangle\right]\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle).\end{aligned} \tag{6.110}$$

Then when we measure the first qubit, we find the outcome $|0\rangle$ with probability one if $f(0) = f(1)$ (constant function) and the outcome $|1\rangle$ with probability one if $f(0) \neq f(1)$ (balanced function).

A quantum computer enjoys an advantage over a classical computer because it can invoke *quantum parallelism*. Because we input a superposition of $|0\rangle$ and $|1\rangle$, the output is sensitive to both the values of $f(0)$ and $f(1)$, even though we ran the box just once.

Deutsch–Jozsa problem. Now we'll consider some generalizations of Deutsch's problem. We will continue to assume that we are to analyze a quantum black box ("quantum oracle"). But in the hope of learning something about complexity, we will imagine that we have a family of black boxes, with variable input size. We are interested in how the time needed to find out what is inside the box scales with the size of the input (where "time" is measured by how many times we query the box).

In the *Deutsch–Jozsa problem*, we are presented with a quantum black box that computes a function taking n bits to 1,

$$f: \{0,1\}^n \rightarrow \{0,1\}, \tag{6.111}$$

and we have it on good authority that f is either constant ($f(x) = c$ for all x) or balanced ($f(x) = 0$ for exactly $\frac{1}{2}$ of the possible input values). We are to solve the decision problem: Is f constant or balanced?

In fact, we can solve this problem, too, accessing the box only once, using the same circuit as for Deutsch's problem (but with x expanded from one bit to n bits). We note that if we apply n Hadamard gates in parallel to n-qubits.

$$\boldsymbol{H}^{(n)} = \boldsymbol{H}\otimes\boldsymbol{H}\otimes\ldots\otimes\boldsymbol{H}, \tag{6.112}$$

then the n-qubit state transforms as

$$\boldsymbol{H}^{(n)}: |x\rangle \rightarrow \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2}}\sum_{y_i=\{0,1\}}(-1)^{x_i y_i}|y_i\rangle\right) \equiv \frac{1}{2^{n/2}}\sum_{y=0}^{2^n-1}(-1)^{x\cdot y}|y\rangle, \tag{6.113}$$

where x, y represent n-bit strings, and $x\cdot y$ denotes the *bitwise* AND (or mod 2 scalar product)

$$x\cdot y = (x_1\wedge y_1)\oplus(x_2\wedge y_2)\oplus\ldots\oplus(x_n\wedge y_n). \tag{6.114}$$

Acting on the input $(|0\rangle)^n|1\rangle$, the action of the circuit is

$$\begin{aligned}(|0\rangle)^n|1\rangle &\rightarrow \left(\frac{1}{2^{n/2}}\sum_{x=0}^{2^n-1}|x\rangle\right)\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\\ &\rightarrow \left(\frac{1}{2^{n/2}}\sum_{x=0}^{2^n-1}(-1)^{f(x)}|x\rangle\right)\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\\ &\rightarrow \left(\frac{1}{2^n}\sum_{x=0}^{2^n-1}\sum_{y=0}^{2^n-1}(-1)^{f(x)}(-1)^{x\cdot y}|y\rangle\right)\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\end{aligned} \tag{6.115}$$

Now let us evaluate the sum

$$\frac{1}{2^n}\sum_{x=0}^{2^n-1}(-1)^{f(x)}(-1)^{x\cdot y}. \tag{6.116}$$

If f is a constant function, the sum is

$$(-1)^{f(x)}\left(\frac{1}{2^n}\sum_{x=0}^{2^n-1}(-1)^{x\cdot y}\right) = (-1)^{f(x)}\delta_{y,0}; \qquad (6.117)$$

it vanishes unless $y = 0$. Hence, when we measure the n-bit register, we obtain the result $|y = 0\rangle \equiv (|0\rangle)^n$ with probability one. But if the function is balanced, then for $y = 0$, the sum becomes

$$\frac{1}{2^n}\sum_{x=0}^{2^n-1}(-1)^{f(x)} = 0, \qquad (6.118)$$

(because half of the terms are $(+1)$ and half are (-1)). Therefore, the probability of obtaining the measurement outcome $|y = 0\rangle$ is zero.

We conclude that one query of the quantum oracle suffices to distinguish constant and balanced function with 100% confidence. The measurement result $y = 0$ means constant, any other result means balanced.

So quantum computation solves this problem neatly, but is the problem really hard classically? If we are restricted to classical input states $|x\rangle$, we can query the oracle repeatedly, choosing the input x at random (without replacement) each time. Once we obtain distinct outputs for two different queries, we have determined that the function is balanced (not constant). But if the function is in fact constant, we will not be *certain* it is constant until we have submitted $2^{n-1}+1$ queries and have obtained the same response every time. In contrast, the quantum computation gives a definite response in only one go. So in this sense (if we demand absolute certainty) the classical calculation requires a number of queries exponential in n, while the quantum computation does not, and we might therefore claim an exponential quantum speedup.

But perhaps it is not reasonable to demand absolute certainty of the classical computation (particularly since any real quantum computer will be susceptible to errors, so that the quantum computer will also be unable to attain absolute certainty.) Suppose we are satisfied to guess balanced or constant, with a probability of success

$$P(\text{success}) > 1 - \varepsilon. \qquad (6.119)$$

If the function is actually balanced, then if we make k queries, the probability of getting the same response every time is $p = 2^{-(k-1)}$. If after receiving the same response k consecutive times we guess that the function is balanced, then a quick Bayesian analysis shows that the probability that our guess is wrong is $\frac{1}{2^{k-1}+1}$ (assuming that balanced and constant are a priori equally probable). So if we guess after k queries, the probability of a wrong guess is

$$1 - P(\text{success}) = \frac{1}{2^{k-1}(2^{k-1}+1)}. \qquad (6.120)$$

Therefore, we can achieve success probability $1-\varepsilon$ for $\varepsilon^{-1} = 2^{k-1}(2^{k-1}+1)$ or $k \sim \frac{1}{2}\log\left(\frac{1}{\varepsilon}\right)$. Since we can reach an exponentially good success probability with a polynomial number of trials, it is not really fair to say that the problem is hard.

Bernstein–Vazirani problem. Exactly the same circuit can be used to solve another variation on the Deutsch–Jozsa problem. Let's suppose that our quantum black box computes one of the functions f_a, where

$$f_a(x) = a \cdot x, \qquad (6.121)$$

and a is an n-bit string. Our job is to determine a.

The quantum algorithm can solve this problem with certainty, given just one (n-qubit) quantum query. For this particular function, the quantum state in eq. (6.115) becomes

$$\frac{1}{2^n}\sum_{x=0}^{2^n-1}\sum_{y=0}^{2^n-1}(-1)^{a\cdot x}(-1)^{x\cdot y}|y\rangle. \qquad (6.122)$$

But in fact

$$\frac{1}{2^n}\sum_{x=0}^{2^n-1}(-1)^{a\cdot x}(-1)^{x\cdot y} = \delta_{a,y}, \qquad (6.123)$$

so this state is $|a\rangle$. We can execute the circuit once and measure the n-qubit register, finding the n-bit string a with probability one.

If only classical queries are allowed, we acquire only one bit of information from each query, and it takes n queries to determine the value of a. Therefore, we have a clear separation between the quantum and classical difficulty of the problem. Even so, this example does not probe the relation of BPP to BQP, because the classical problem is not hard. The number of queries required classically is only linear in the input size, not exponential.

Simon's problem. Bernstein and Vazirani managed to formulate a variation on the above problem that *is* hard classically, and so establish for the first time a "relativized" separation between quantum and classical complexity. We will find it more instructive to consider a simpler example proposed somewhat later by Daniel Simon.

Once again we are presented with a quantum black box, and this time we are assured that the box computes a function

$$f : \{0,1\}^n \to \{0,1\}^n, \tag{6.124}$$

that is 2-to-1. Furthermore, the function has a "period" given by the n-bit string a; that is

$$f(x) = f(y) \qquad \text{iff} \qquad y = x \oplus a, \tag{6.125}$$

where here $\oplus$ denotes the bitwise XOR operation. (So a is the period if we regard x as taking values in $(Z_2)^n$ rather than Z_{2^n}.) This is all we know about f. Our job is to determine the value of a.

Classically this problem is *hard*. We need to query the oracle an exponentially large number of times to have any reasonable probability of finding a. We don't learn anything until we are fortunate enough to choose two queries x and y that happen to satisfy $x \oplus y = a$. Suppose, for example, that we choose $2^{n/4}$ queries. The number of pairs of queries is less than $(2^{n/4})^2$, and for each pair $\{x, y\}$, the probability that $x \oplus y = a$ is 2^{-n}. Therefore, the probability of successfully finding a is less than

$$2^{-n}(2^{n/4})^2 = 2^{-n/2}; \tag{6.126}$$

even with exponentially many queries, the success probability is exponentially small.

If we wish, we can frame the question as a decision problem: Either f is a 1-1 function, or it is 2-to-1 with some randomly chosen period a, each occurring with an a priori probability $\frac{1}{2}$. We are to determine whether the function is 1-to-1 or 2-to-1. Then, after $2^{n/4}$ classical queries, our probability of making a correct guess is

$$P(\text{success}) < \frac{1}{2} + \frac{1}{2^{n/2}}, \tag{6.127}$$

which does not remain bounded away from $\frac{1}{2}$ as n gets large.

But with quantum queries the problem is easy! The circuit we use is essentially the same as above, but now *both* registers are expanded to n qubits. We prepare the equally weighted superposition of all n-bit strings (by acting on $|0\rangle$ with $\boldsymbol{H}^{(n)}$), and then we query the oracle:

$$U_f : \left(\sum_{x=0}^{2^n-1} |x\rangle\right)|0\rangle \to \sum_{x=0}^{2^n-1} |x\rangle|f(x)\rangle. \tag{6.128}$$

Now we measure the second register. (This step is not actually necessary, but I include it here for the sake of pedagogical clarity.) The measurement outcome is selected at random from the 2^{n-1} possible values of $f(x)$, each occurring equiprobably. Suppose the outcome is $f(x_0)$. Then because both x_0 and $x_0 \oplus a$, and only these values, are mapped by f to $f(x_0)$, we have prepared the state

$$\frac{1}{\sqrt{2}}(|x_0\rangle + |x_0 \oplus a\rangle) \tag{6.129}$$

in the first register.

Now we want to extract some information about a. Clearly it would do us no good to measure the register (in the computational basis) at this point. We would obtain either the outcome x_0 or $x_0 \oplus a$, each occurring with probability $\frac{1}{2}$, but neither outcome would reveal anything about the value of a.

But suppose we apply the Hadamard transform $\boldsymbol{H}^{(n)}$ to the register before we measure:

$$\begin{aligned}\boldsymbol{H}^{(n)} :& \frac{1}{\sqrt{2}}(|x_0\rangle + |x_0 + a\rangle) \\ &\to \frac{1}{2^{(n+1)/2}} \sum_{y=0}^{2^n-1} \left[(-1)^{x_0\cdot y} + (-1)^{(x_0 \oplus a)\cdot y}\right] |y\rangle \\ &= \frac{1}{2^{(n-1)/2}} \sum_{a\cdot y=0} (-1)^{x_0\cdot y}|y\rangle. \end{aligned} \tag{6.130}$$

If $a \cdot y = 1$, then the terms in the coefficient of $|y\rangle$ interfere destructively. Hence only states $|y\rangle$ with $a \cdot y = 0$ survive in the sum over y. The measurement outcome, then, is selected at random from all possible values of y such that $a \cdot y = 0$, each occurring with probability $2^{-(n-1)}$.

We run this algorithm repeatedly, each time obtaining another value of y satisfying $y \cdot a = 0$. Once we have found n such linearly independent values $\{y_1, y_2, y_3 \ldots y_n\}$ (that is, linearly independent over $(Z_2)^n$), we can solve the equations

$$\begin{aligned} y_1 \cdot a &= 0 \\ y_2 \cdot a &= 0 \\ &\vdots \\ y_n \cdot a &= 0, \end{aligned} \tag{6.131}$$

to determine a unique value of a, and our problem is solved. It is easy to see that with $O(n)$ repetitions, we can attain a success probability that is exponentially close to 1.

So we finally have found an example where, given a particular type of quantum oracle, we can solve a problem in polynomial time by exploiting quantum superpositions, while exponential time is required if we are limited to classical queries. As a computer scientist might put it:

> There exists an oracle relative to which $BQP \neq BPP$.

Note that whenever we compare classical and quantum complexity relative to an oracle, we are considering a quantum oracle (queries and replies are states in Hilbert space), but with a preferred orthonormal basis. If we submit a classical query (an element of the preferred basis) we always receive a classical response (another basis element). The issue is whether we can achieve a significant speedup by choosing more general quantum queries.

6.4 Quantum Database Search

The next algorithm we will study also exhibits, like Simon's algorithm, a speedup with respect to what can be achieved with a classical algorithm. But in this case the speedup is merely quadratic (the quantum time scales like the square root of the classical time), in contrast to the exponential speedup in the solution to Simon's problem. Nevertheless, the result (discovered by Lov Grover) is extremely interesting, because of the broad utility of the algorithm.

Heuristically, the problem we will address is: we are confronted by a very large unsorted database containing $N \gg 1$ items, and we are to locate one particular item, to find a needle in the haystack. Mathematically, the database is represented by a table, or a function $f(x)$, with $x \in \{0, 1, 2, \ldots, N-1\}$. We have been assured that the entry a occurs in the table exactly once; that is, that $f(x) = a$ for only one value of x. The problem is, given a, to find this value of x.

If the database has been properly *sorted*, searching for x is easy. Perhaps someone has been kind enough to list the values of a in ascending order. Then we can find x by looking up only $\log_2 N$ entries in the table. Let's suppose $N \equiv 2^n$ is a power of 2. First we look up $f(x)$ for $x = 2^{n-1} - 1$, and check if $f(x)$ is greater than a. If so, we next look up f at $x = 2^{n-2} - 1$, *etc.* With each table lookup, we reduce the number of candidate values of x by a factor of 2, so that n lookups suffice to sift through all 2^n sorted items. You can use this algorithm to look up a number in the Los Angeles phone book, because the names are listed in lexicographic order.

But now suppose that you know someone's phone number, and you want to look up her *name*. Unless you are fortunate enough to have access to a reverse directory, this is a tedious procedure. Chances are you will need to check quite a few entries in the phone book before you come across her number.

In fact, if the N numbers are listed in a random order, you will need to look up $\frac{1}{2}N$ numbers before the probability is $P = \frac{1}{2}$ that you have found her number (and hence her name). What Grover discovered is that, if you have a quantum phone book, you can learn her name with high probability by consulting the phone book only about $\sqrt{N}$ times.

This problem, too, can be formulated as an oracle or "black box" problem. In this case, the oracle is the phone book, or lookup table. We can input a name (a value of x) and the oracle outputs either 0, if $f(x) \neq a$, or 1, if $f(x) = a$. Our task is to find, as quickly as possible, the value of x with

$$f(x) = a. \tag{6.132}$$

Why is this problem important? You may have never tried to find in the phone book the name that matches a given number, but if it weren't so hard you might try it more often! More broadly, a rapid method for searching an unsorted database could be invoked to solve any problem in NP. Our oracle could be a subroutine that interrogates every potential "witness" y that could potentially testify to certify a solution to the problem. For example, if we are confronted by a graph and need to know if it admits a Hamiltonian path, we could submit a path to the "oracle," and it could quickly answer whether the path is Hamiltonian or not. If we knew a fast way to query the oracle about all the possible paths, we would be able to find a Hamiltonian path efficiently (if one exists).

6.4.1 The oracle

So "oracle" could be shorthand for a subroutine that quickly evaluates a function to check a proposed solution to a decision problem, but let us continue to regard the oracle abstractly, as a black box. The oracle "knows" that of the 2^n possible strings of length n, one (the "marked" string or "solution" ω) is special. We submit a query x to the oracle, and it tells us whether $x=\omega$ or not. It returns, in other words, the value of a function $f_\omega(x)$, with

$$\begin{aligned} f_\omega(x) &= 0, \qquad x \neq \omega, \\ f_\omega(x) &= 1, \qquad x = \omega. \end{aligned} \tag{6.133}$$

But furthermore, it is a *quantum* oracle, so it can respond to queries that are superpositions of strings. The oracle is a quantum black box that implements the unitary transformation

$$\boldsymbol{U}_{f_\omega} : |x\rangle|y\rangle \to |x\rangle|y \oplus f_\omega(x)\rangle, \tag{6.134}$$

where $|x\rangle$ is an n-qubit state, and $|y\rangle$ is a single-qubit state.

As we have previously seen in other contexts, we may choose the state of the single-qubit register to be $\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$, so that the oracle acts as

$$\begin{aligned} \boldsymbol{U}_{f_\omega} &: |x\rangle\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \\ &\to (-1)^{f_\omega(x)}|x\rangle\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle). \end{aligned} \tag{6.135}$$

We may now ignore the second register, and obtain

$$\boldsymbol{U}_\omega : |x\rangle \to (-1)^{f_\omega(x)}|x\rangle, \tag{6.136}$$

or

$$\boldsymbol{U}_\omega = \boldsymbol{1} - 2|\omega\rangle\langle\omega|. \tag{6.137}$$

The oracle flips the sign of the state $|\omega\rangle$, but acts trivially on any state orthogonal to $|\omega\rangle$. This transformation has a simple geometrical interpretation. Acting on any vector in the 2^n-dimensional Hilbert space, U_ω *reflects* the vector about the hyperplane orthogonal to $|\omega\rangle$ (it preserves the component in the hyperplane, and flips the component along $|\omega\rangle$).

We know that the oracle performs this reflection for some particular computational basis state $|\omega\rangle$, but we know nothing *a priori* about the value of the string ω. Our job is to determine ω, with high probability, consulting the oracle a minimal number of times.

6.4.2 The Grover iteration

As a first step, we prepare the state

$$|s\rangle = \frac{1}{\sqrt{N}}\left(\sum_{x=0}^{N-1}|x\rangle\right), \tag{6.138}$$

The equally weighted superposition of all computational basis states – this can be done easily by applying the Hadamard transformation to each qubit of the initial state $|x=0\rangle$. Although we do not know the value of ω, we *do* know that $|\omega\rangle$ is a computational basis state, so that

$$|\langle\omega|s\rangle| = \frac{1}{\sqrt{N}}, \tag{6.139}$$

irrespective of the value of ω. Were we to measure the state $|s\rangle$ by projecting onto the computational basis, the probability that we would "find" the marked state $|\omega\rangle$ is only $\frac{1}{N}$. But following Grover, we can repeatedly iterate a transformation that enhances the probability amplitude of the unknown state $|\omega\rangle$ that we are seeking, while suppressing the amplitude of all of the undesirable states $|x \neq \omega\rangle$. We construct this Grover iteration by combining the unknown reflection $\boldsymbol{U}_\omega$ performed by the oracle with a known reflection that we can perform ourselves. This known reflection is

$$\boldsymbol{U}_s = 2|s\rangle\langle s| - \boldsymbol{1}, \tag{6.140}$$

which preserves $|s\rangle$, but flips the sign of any vector orthogonal to $|s\rangle$. Geometrically, acting on an arbitrary vector, it preserves the component along $|s\rangle$ and flips the component in the hyperplane orthogonal to $|s\rangle$.

We'll return below to the issue of constructing a quantum circuit that implements $\boldsymbol{U}_s$; for now let's just assume that we can perform $\boldsymbol{U}_s$ efficiently.

One Grover iteration is the unitary transformation

$$\boldsymbol{R}_{\text{grov}} = \boldsymbol{U}_s \boldsymbol{U}_\omega, \tag{6.141}$$

one oracle query followed by our reflection. Let's consider how $\boldsymbol{R}_{\text{grov}}$ acts in the plane spanned by $|\omega\rangle$ and $|s\rangle$. This action is easiest to understand if we visualize it geometrically. Recall that

$$|\langle s|\omega\rangle| = \frac{1}{\sqrt{N}} \equiv \sin\theta, \tag{6.142}$$

so that $|s\rangle$ is rotated by θ from the axis $|\omega^\perp\rangle$ normal to $|\omega\rangle$ in the plane. $\boldsymbol{U}_\omega$ reflects a vector in the plane about the axis $|\omega^\perp\rangle$, and $\boldsymbol{U}_s$ reflects a vector about the axis $|s\rangle$. Together, the two reflections rotate the vector by 2θ:

The Grover iteration, then, is nothing but a rotation by 2θ in the plane determined by $|s\rangle$ and $|\omega\rangle$.

6.4.3 Finding 1 out of 4

Let's suppose, for example, that there are $N = 4$ items in the database, with one marked item. With classical queries, the marked item could be found in the 1st, 2nd, 3rd, or 4th query; on the average $2\frac{1}{2}$ queries will be needed before we are successful and four are needed in the worst case.[8] But since $\sin\theta = \frac{1}{\sqrt{N}} = \frac{1}{2}$, we have $\theta = 30^o$ and $2\theta = 60^o$. After one Grover iteration, then, we rotate $|s\rangle$ to a 90^o angle with $|\omega^\perp\rangle$; that is, it lines up with $|\omega\rangle$. When we measure by projecting onto the computational basis, we obtain the result $|\omega\rangle$ *with certainty*. Just one quantum query suffices to find the marked state, a notable improvement over the classical case.

[8]Of course, if we know there is one marked state, the 4th query is actually superfluous, so it might be more accurate to say that at most three queries are needed, and $2\frac{1}{4}$ queries are required on the average.

There is an alternative way to visualize the Grover iteration that is sometimes useful, as an "inversion about the average." If we expand a state $|\psi\rangle$ in the computational basis

$$|\psi\rangle = \sum_x a_x |x\rangle, \tag{6.143}$$

then its inner product with $|s\rangle = \frac{1}{\sqrt{N}} \sum_x |x\rangle$ is

$$\langle s|\psi\rangle = \frac{1}{\sqrt{N}} \sum_x a_x = \sqrt{N}\langle a\rangle, \tag{6.144}$$

where

$$\langle a\rangle = \frac{1}{N} \sum_x a_x, \tag{6.145}$$

is the mean of the amplitude. Then if we apply $\boldsymbol{U}_s = 2|s\rangle\langle s| - \mathbf{1}$ to $|\psi\rangle$, we obtain

$$\boldsymbol{U}_s|\psi\rangle = \sum_x (2\langle a\rangle - a_x)|x\rangle; \tag{6.146}$$

the amplitudes are transformed as

$$\boldsymbol{U}_s : a_x - \langle a\rangle \to \langle a\rangle - a_x, \tag{6.147}$$

that is the coefficient of $|x\rangle$ is inverted about the mean value of the amplitude.

If we consider again the case $N = 4$, then in the state $|s\rangle$ each amplitude is $\frac{1}{2}$. One query of the oracle flips the sign of the amplitude of marked state, and so reduces the mean amplitude to $\frac{1}{4}$. Inverting about the mean then brings the amplitudes of all unmarked states from $\frac{1}{2}$ to zero, and raises the amplitude of the marked state from $-\frac{1}{2}$ to 1. So we recover our conclusion that one query suffices to find the marked state with certainty.

We can also easily see that one query is sufficient to find a marked state if there are N entries in the database, and exactly $\frac{1}{4}$ of them are marked. Then, as above, one query reduces the mean amplitude from $\frac{1}{\sqrt{N}}$ to $\frac{1}{2\sqrt{N}}$, and inversion about the mean then reduces the amplitude of each unmarked state to zero.

(When we make this comparison between the number of times we need to consult the oracle if the queries can be quantum rather than classical, it may be a bit unfair to say that only one query is needed in the quantum case. If the oracle is running a routine that computes a function, then some scratch space will be filled with garbage during the computation. We will need to erase the garbage by running the computation backwards in order to maintain quantum coherence. If the classical computation is irreversible there is no need to run the oracle backwards. In this sense, one query of the quantum oracle may be roughly equivalent, in terms of complexity, to two queries of a classical oracle.)

6.4.4 Finding 1 out of N

Let's return now to the case in which the database contains N items, and exactly one item is marked. Each Grover iteration rotates the quantum state in the plane determined by $|s\rangle$ and $|\omega\rangle$; after T iterations, the state is rotated by $\theta + 2T\theta$ from the $|\omega^\perp\rangle$ axis. To optimize the probability of finding the marked state when we finally perform the measurement, we will iterate until this angle is close to 90^o, or

$$(2T+1)\theta \simeq \frac{\pi}{2} \Rightarrow 2T+1 \simeq \frac{\pi}{2\theta}, \tag{6.148}$$

we recall that $\sin\theta = \frac{1}{\sqrt{N}}$, or

$$\theta \simeq \frac{1}{\sqrt{N}}, \tag{6.149}$$

for N large; if we choose

$$T = \frac{\pi}{4}\sqrt{N}(1 + O(N^{-1/2})), \tag{6.150}$$

then the probability of obtaining the measurement result $|\omega\rangle$ will be

$$\text{Prob}(\omega) = \sin^2\left((2T+1)\theta\right) = 1 - O\left(\frac{1}{N}\right). \tag{6.151}$$

We conclude that only about $\frac{\pi}{4}\sqrt{N}$ queries are needed to determine ω with high probability, a quadratic speedup relative to the classical result.

6.4.5 Multiple solutions

If there are $r > 1$ marked states, and r is known, we can modify the number of iterations so that the probability of finding one of the marked states is still very close to 1. The analysis is just as above, except that the oracle induces a reflection in the hyperplane orthogonal to the vector

$$|\tilde{\omega}\rangle = \frac{1}{\sqrt{r}}\left(\sum_{i=1}^{r} |\omega_i\rangle\right), \tag{6.152}$$

the equally weighted superposition of the marked computational basis states $|\omega_i\rangle$. Now

$$\langle s|\tilde{\omega}\rangle = \sqrt{\frac{r}{N}} \equiv \sin\theta, \tag{6.153}$$

and a Grover iteration rotates a vector by 2θ in the plane spanned by $|s\rangle$ and $|\tilde{\omega}\rangle$; we again conclude that the state is close to $|\tilde{\omega}\rangle$ after a number of iterations

$$T \simeq \frac{\pi}{4\theta} = \frac{\pi}{4}\sqrt{\frac{N}{r}}. \tag{6.154}$$

If we then measure by projecting onto the computational basis, we will find one of the marked states (each occurring equiprobably) with probability close to one. (As the number of solutions increases, the time needed to find one of them declines like $r^{-1/2}$, as opposed to r^{-1} in the classical case.)

Note that if we continue to perform further Grover iterations, the vector continues to rotate, and so the probability of finding a marked state (when we finally measure) begins to decline. The Grover algorithm is like baking a soufflé – if we leave it in the oven for too long, it starts to fall. Therefore, if we don't know anything about the number of marked states, we might fail to find one of them. For example, $T \sim \frac{\pi}{4}\sqrt{N}$ iterations is optimal for $r = 1$, but for $r = 4$, the probability of finding a marked state after this many iterations is quite close to zero.

But even if we don't know r *a priori*, we can still find a solution with a quadratic speed up over classical algorithms (for $r \ll N$). For example, we might choose the number of iterations to be random in the range 0 to $\frac{\pi}{4}\sqrt{N}$. Then the expected probability of finding a marked state is close to $1/2$ for each r, so we are unlikely to fail to find a marked state after several repetitions. And each time we measure, we can submit the state we find to the oracle as a classical query to confirm whether that state is really marked.

In particular, if we don't find a solution after several attempts, there probably is no solution. Hence with high probability we can correctly answer the yes/no question, "Is there a marked state?" Therefore, we can adopt the Grover algorithm to solve any NP problem, where the oracle checks a proposed solution, with a quadratic speedup over a classical exhaustive search.

6.4.6 Implementing the reflection

To perform a Grover iteration, we need (aside from the oracle query) a unitary transformation

$$\boldsymbol{U}_s = 2|s\rangle\langle s| - \mathbf{1}, \tag{6.155}$$

that reflects a vector about the axis defined by the vector $|s\rangle$. How do we build this transformation efficiently from quantum gates? Since $|s\rangle = \boldsymbol{H}^{(n)}|0\rangle$, where $\boldsymbol{H}^{(n)}$ is the bitwise Hadamard transformation, we may write

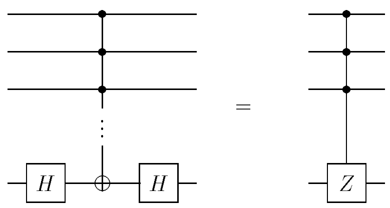
$$\boldsymbol{U}_s = \boldsymbol{H}^{(n)}(2|0\rangle\langle 0| - \mathbf{1})\boldsymbol{H}^{(n)}, \tag{6.156}$$

so it will suffice to construct a reflection about the axis $|0\rangle$. We can easily build this reflection from an n-bit Toffoli gate $\theta^{(n)}$.

Recall that

$$\boldsymbol{H}\boldsymbol{\sigma}_x\boldsymbol{H} = \boldsymbol{\sigma}_z; \tag{6.157}$$

a bit flip in the Hadamard rotated basis is equivalent to a flip of the relative phase of $|0\rangle$ and $|1\rangle$. Therefore:



after conjugating the last bit by $\boldsymbol{H}$, $\theta^{(n)}$ becomes controlled$^{(n-1)}$-$\boldsymbol{\sigma}_z$, which flips the phase of $|11\ldots|1\rangle$ and acts trivially on all other computational basis states. Conjugating by NOT$^{(n)}$, we obtain $\boldsymbol{U}_s$, aside from an irrelevant overall minus sign.

You will show in an exercise that the n-bit Toffoli gate $\theta^{(n)}$ can be constructed from $2n - 5$ 3-bit Toffoli gates $\theta^{(3)}$ (if sufficient scratch space is available). Therefore, the circuit that constructs $\boldsymbol{U}_s$ has a size *linear* in $n = \log N$. Grover's database search (assuming the oracle answers a query instantaneously) takes a time of order $\sqrt{N}\log N$. If we regard the oracle as a subroutine that performs a function evaluation in polylog time, then the search takes time of order $\sqrt{N}\text{poly}(\log N)$.

6.5 The Grover Algorithm Is Optimal

Grover's quadratic quantum speedup of the database search is already interesting and potentially important, but surely with more cleverness we can do better, can't we? No, it turns out that we can't. Grover's algorithm provides the fastest possible quantum search of an unsorted database, if "time" is measured according to the number of queries of the oracle.

Considering the case of a single marked state $|\omega\rangle$, let $\boldsymbol{U}(\omega, T)$ denote a quantum circuit that calls the oracle T times. We place *no* restriction on the circuit aside from specifying the number of queries; in particular, we place no limit on the number of quantum gates. This circuit is applied to an initial state $|\psi(0)\rangle$, producing a final state

$$|\psi_\omega(t)\rangle = \boldsymbol{U}(\omega, T)|\psi(0)\rangle. \tag{6.158}$$

Now we are to perform a measurement designed to distinguish among the N possible values of ω. If we are to be able to perfectly distinguish among the possible values, the states $|\psi_\omega(t)\rangle$ must all be mutually orthogonal, and if we are to distinguish correctly with high probability, they must be nearly orthogonal.

Now, if the states $\{|\psi_\omega\rangle$ are an orthonormal basis, then, for any fixed normalized vector $|\varphi\rangle$,

$$\sum_{\omega=0}^{N-1} \| \, |\psi_\omega\rangle - |\varphi\rangle \, \|^2 \geq 2N - 2\sqrt{N}. \tag{6.159}$$

(The sum is minimized if $|\varphi\rangle$ is the equally weighted superposition of all the basis elements, $|\varphi\rangle = \frac{1}{\sqrt{N}}\sum_\omega |\psi_\omega\rangle$, as you can show by invoking a Lagrange multiplier to perform the constrained extremization.) Our strategy will be to choose the state $|\varphi\rangle$ suitably so that we can use this inequality to learn something about the number T of oracle calls.

Our circuit with T queries builds a unitary transformation

$$\boldsymbol{U}(\omega, T) = \boldsymbol{U}_\omega \boldsymbol{U}_T \boldsymbol{U}_\omega \boldsymbol{U}_{T-1} \ldots \boldsymbol{U}_\omega \boldsymbol{U}_1, \tag{6.160}$$

where $\boldsymbol{U}_\omega$ is the oracle transformation, and the $\boldsymbol{U}_t$'s are arbitrary non-oracle transformations. For our state $|\varphi(T)\rangle$ we will choose the result of applying $\boldsymbol{U}(\omega, T)$ to $|\psi(0)\rangle$, except with each $\boldsymbol{U}_\omega$ replaced by $\mathbf{1}$; that is, the same circuit, but with all queries submitted to the "empty oracle." Hence,

$$|\varphi(T)\rangle = \boldsymbol{U}_T \boldsymbol{U}_{T-1} \ldots \boldsymbol{U}_2 \boldsymbol{U}_1 |\psi(0)\rangle, \tag{6.161}$$

while

$$|\psi_\omega(T)\rangle = \boldsymbol{U}_\omega \boldsymbol{U}_T \boldsymbol{U}_\omega \boldsymbol{U}_{T-1} \ldots \boldsymbol{U}_\omega \boldsymbol{U}_1 |\psi(0)\rangle. \tag{6.162}$$

To compare $|\varphi(T)\rangle$ and $|\psi_\omega(T)\rangle$, we appeal to our previous analysis of the effect of errors on the accuracy of a circuit, regarding the ω oracle as an "erroneous" implementation of the empty oracle. The error vector in the t-th step (cf. eq. (6.63)) is

$$\begin{aligned} \| \, |E(\omega, t)\rangle \, \| &= \| \, (\boldsymbol{U}_\omega - \mathbf{1})|\varphi(t)\rangle \, \| \\ &= 2|\langle\omega|\varphi(t)\rangle|, \end{aligned} \tag{6.163}$$

since $\boldsymbol{U}_\omega = \mathbf{1} - 2|\omega\rangle\langle\omega|$. After T queries we have (cf. eq. (6.66))

$$\| \, |\psi_\omega(T)\rangle - |\varphi(T)\rangle \, \| \leq 2\sum_{t=1}^{T} |\langle\omega|\varphi(t)\rangle|. \tag{6.164}$$

From the identity

$$\begin{aligned} &\left(\sum_{t=1}^{T} c_t\right)^2 + \frac{1}{2}\sum_{s,t=1}^{T}(c_s - c_t)^2 \\ &= \sum_{s,t=1}^{T}\left(c_t c_s + \frac{1}{2}c_s^2 - c_t c_s + \frac{1}{2}c_s^2\right) = T\sum_{t=1}^{T} c_t^2, \end{aligned} \tag{6.165}$$

we obtain the inequality

$$\left(\sum_{t=1}^{T} c_t\right)^2 \leq T\sum_{t=1}^{T} c_t^2, \tag{6.166}$$

which applied to eq. (6.164) yields

$$\| \, |\psi_\omega(T)\rangle - |\varphi(T)\rangle \, \|^2 \leq 4T\left(\sum_{t=1}^{T} |\langle\omega|\varphi(t)\rangle|^2\right). \tag{6.167}$$

Summing over ω we find

$$\sum_\omega \| \, |\psi_\omega(T)\rangle - |\varphi(T)\rangle \, \|^2 \leq 4T\sum_{t=1}^{T}\langle\varphi(t)|\varphi(t)\rangle \ = \ 4T^2. \tag{6.168}$$

Invoking eq. (6.159) we conclude that

$$4T^2 \geq 2N - 2\sqrt{N}, \tag{6.169}$$

if the states $|\psi_\omega(T)\rangle$ are mutually orthogonal. We have, therefore, found that any quantum algorithm that can distinguish all the possible values of the marked state must query the oracle T times where

$$T \geq \sqrt{\frac{N}{2}}, \tag{6.170}$$

(ignoring the small correction as $N \to \infty$). Grover's algorithm finds ω in $\frac{\pi}{4}\sqrt{N}$ queries, which exceeds this bound by only about 11%. In fact, it is possible to refine the argument to improve the bound to $T \geq \frac{\pi}{4}\sqrt{N}(1-\varepsilon)$, which is asymptotically saturated by the Grover algorithm.[9] Furthermore, we can show that Grover's circuit attains the optimal success probability in $T \leq \frac{\pi}{4}\sqrt{N}$ queries.

One feels a twinge of disappointment (as well as a surge of admiration for Grover) at the realization that the database search algorithm cannot be improved. What are the implications for quantum complexity?

For many optimization problems in the NP class, there is no better method known than exhaustive search of all the possible solutions. By exploiting quantum parallelism, we can achieve a quadratic speedup of exhaustive search. Now we have learned that the quadratic speedup is the best possible if we rely on the power of sheer quantum parallelism, if we don't design our quantum algorithm to exploit the specific structure of the problem we wish to solve. Still, we might do better if we are sufficiently clever.

The optimality of the Grover algorithm might be construed as evidence that $BQP \not\supseteq NP$. At least, if it turns out that $NP \subseteq BQP$ and $P \neq NP$, then the NP problems must share a deeply hidden structure (for which there is currently no evidence) that is well-matched to the peculiar capabilities of quantum circuits.

Even the quadratic speedup may prove useful for a variety of NP-complete optimization problems. But a quadratic speedup, unlike an exponential one, does not really move the frontier between solvability and intractability. Quantum computers may someday outperform classical computers in performing exhaustive search, but only if the clock speed of quantum devices does not lag too far behind that of their classical counterparts.

6.6 Generalized Search and Structured Search

In the Grover iteration, we perform the transformation $\boldsymbol{U}_s = 2|s\rangle\langle s| - \boldsymbol{1}$, the reflection in the axis defined by $|s\rangle = \frac{1}{\sqrt{N}}\sum_{x=0}^{N-1}|x\rangle$. Why this axis? The advantage of the state $|s\rangle$ is that it has the same overlap with each and every computational basis state. Therefore, the overlap of any marked state $|\omega\rangle$ with $|s\rangle$ is guaranteed to be $|\langle\omega|s\rangle| = 1/\sqrt{N}$. Hence, if we know the number of marked states, we can determine how many iterations are required to find a marked state with high probability – the number of iterations needed does

[9]C. Zalka, "Grover's Quantum Searching Algorithm is Optimal," quant-ph/9711070.

not depend on which states are marked.

But of course, we could choose to reflect about a different axis. If we can build the unitary $\boldsymbol{U}$ (with reasonable efficiency) then we can construct

$$\boldsymbol{U}(2|0\rangle\langle 0| - \boldsymbol{1})\boldsymbol{U}^\dagger = 2\boldsymbol{U}|0\rangle\langle 0|\boldsymbol{U}^\dagger - \boldsymbol{1}, \tag{6.171}$$

which reflects in the axis $\boldsymbol{U}|0\rangle$.

Suppose that

$$|\langle\omega|\boldsymbol{U}|0\rangle| = \sin\theta, \tag{6.172}$$

where $|\omega\rangle$ is the marked state. Then if we replace $\boldsymbol{U}_s$ in the Grover iteration by the reflection eq. (6.171), one iteration performs a rotation by 2θ in the plane determined by $|\omega\rangle$ and $\boldsymbol{U}|0\rangle$ (by the same argument we used for $\boldsymbol{U}_s$). Thus, after T iterations, with $(2T+I)\theta \cong \pi/2$, a measurement in the computational basis will find $|\omega\rangle$ with high probability. Therefore, we can still search a database if we replace $\boldsymbol{H}^{(n)}$ by $\boldsymbol{U}$ in Grover's quantum circuit, as long as $\boldsymbol{U}|0\rangle$ is not orthogonal to the marked state.[10] But if we have no *a priori* information about which state is marked, then $\boldsymbol{H}^{(n)}$ is the best choice, not only because $|s\rangle$ has a known overlap with each marked state, but also because it has the largest *average* overlap with all the possible marked states.

But sometimes when we are searching a database, we *do* have some information about where to look, and in that case, the generalized search strategy described above may prove useful.[11]

As an example of a problem with some auxiliary structure, suppose that $f(x,y)$ is a one-bit-valued function of the two n-bit strings x and y, and we are to find the unique solution to $f(x,y) = 1$. With Grover's algorithm, we can search through the N^2 possible values ($N = 2^n$) of (x,y) and find the solution (x_0, y_0) with high probability after $\frac{\pi}{4}N$ iterations, a quadratic speedup with respect to classical search.

But further suppose that $g(x)$ is a function of x only, and that it is known that $g(x) = 1$ for exactly M values of x, where $1 \ll M \ll N$. And furthermore, it is known that $g(x_0) = 1$. Therefore, we can use g to help us find the solution (x_0, y_0).

[10]L.K. Grover "Quantum Computers Can Search Rapidly By Using Almost Any Transformation," quant-ph/9712011.

[11]E. Farhi and S. Gutmann, "Quantum-Mechanical Square Root Speedup in a Structured Search Problem," quant-ph/9711035; L.K. Grover, "Quantum Search On Structured Problems," quant-ph/9802035.

Now we have two oracles to consult, one that returns the value of $f(x,y)$, and the other returning the value of $g(x)$. Our task is to find (x_0,y_0) with a minimal number of queries.

Classically, we need of order NM queries to find the solution with reasonable probability. We first evaluate $g(x)$ for each x; then we restrict our search for a solution to $f(x,y)=1$ to only those M values of x such that $g(x)=1$. It is natural to wonder whether there is a way to perform a quantum search in a time of order the square root of the classical time. Exhaustive search that queries only the f oracle requires time $N \gg \sqrt{NM}$, and so does not do the job. We need to revise our method of quantum search to take advantage of the structure provided by g.

A better method is to first apply Grover's algorithm to $g(x)$. In about $\frac{\pi}{4}\sqrt{\frac{N}{M}}$ iterations, we prepare a state that is close to the equally weighted superposition of the M solutions to $g(x)=1$. In particular, the state $|x_0\rangle$ appears with amplitude $\frac{1}{\sqrt{M}}$. Then we apply Grover's algorithm to $f(x,y)$ with x fixed. In about $\frac{\pi}{4}\sqrt{N}$ iterations, the state $|x_0\rangle|s\rangle$ evolves to a state quite close to $|x_0\rangle|y_0\rangle$. Therefore $|x_0,y_0\rangle$ appears with amplitude $\frac{1}{\sqrt{M}}$.

The unitary transformation we have constructed so far, in about $\frac{\pi}{4}\sqrt{N}$ queries, can be regarded as the transformation $\boldsymbol{U}$ that defines a generalized search. Furthermore, we know that

$$\langle x_0,y_0|U|0,0\rangle \cong \frac{1}{\sqrt{M}}. \tag{6.173}$$

Therefore, if we iterate the generalized search about $\frac{\pi}{4}\sqrt{M}$ times, we will have prepared a state that is quite close to $|x_0,y_0\rangle$. With altogether about

$$\left(\frac{\pi}{4}\right)^2\sqrt{NM}, \tag{6.174}$$

queries, then, we can find the solution with high probability. This is indeed a quadratic speedup with respect to the classical search.

6.7 Some Problems Admit No Speedup

The example of structured search illustrates that quadratic quantum speedups over classical algorithms can be attained for a variety of problems, not just for an exhaustive search of a structureless database. One might even dare

to hope that quantum parallelism enables us to significantly speedup any classical algorithm. This hope will now be dashed – for many problems, no quantum speedup is possible.

We continue to consider problems with a quantum black box, an oracle, that computes a function f taking n bits to 1. But we will modify our notation a little. The function f can be represented as a string of $N=2^n$ bits

$$X = X_{N-1}X_{N-2}\ldots X_1X_0, \tag{6.175}$$

where X_i denotes $f(i)$. Our problem is to evaluate some one-bit-valued function of X, that is, to answer a yes/no question about the properties of the oracle. What we will show is that for some functions of X, we can't evaluate the function with low error probability using a quantum algorithm, unless the algorithm queries the oracle as many times (or nearly as many times) as required with a classical algorithm.[12]

The key idea is that any Boolean function of the X_i's can be represented as a polynomial in the X_i's. Furthermore, the probability distribution for a quantum measurement can be expressed as a polynomial in X, where the degree of the polynomial is $2T$, if the measurement follows T queries of the oracle. The issue, then, is whether a polynomial of degree $2T$ can provide a reasonable approximation to the Boolean function of interest.

The action of the oracle can be represented as

$$\boldsymbol{U}_O : |i,y;z\rangle \to |i,y\oplus X_i;z\rangle, \tag{6.176}$$

where i takes values in $\{0,1,\ldots,N-1\}$, $y\in\{0,1\}$, and z denotes the state of auxiliary qubits not acted upon by the oracle. Therefore, in each 2×2 block spanned by $|i,0,z\rangle$ and $|i,1,z\rangle$, $\boldsymbol{U}_O$ is the 2×2 matrix

$$\begin{pmatrix} 1-X_i & X_i \\ X_i & 1-X_i \end{pmatrix}. \tag{6.177}$$

Quantum gates other than oracle queries have no dependence on X. Therefore after a circuit with T queries acts on any initial state, the resulting state $|\psi\rangle$ has amplitudes that are (at most) Tth-degree polynomials in X. If we perform a POVM on $|\psi\rangle$, then the probability $\langle\psi|\boldsymbol{F}|\psi\rangle$ of the outcome associated with the positive operator $\boldsymbol{F}$ can be expressed as a polynomial in X of degree at most $2T$.

[12] E. Farhi, *et al.*, quant-ph/9802045; R. Beals, *et al.*, quant-ph/9802049.

Now any Boolean function of the X_i's can be expressed (uniquely) as a polynomial of degree $\leq N$ in the X_i's. For example, consider the OR function of the N X_i's; it is

$$\mathrm{OR}(X) = 1 - (1 - X_0)(1 - X_1)\cdots(1 - X_{N-1}), \tag{6.178}$$

a polynomial of degree N.

Suppose that we would like our quantum circuit to evaluate the OR function *with certainty*. Then we must be able to perform a measurement with two outcomes, 0 and 1, where

$$\begin{aligned}\mathrm{Prob}(0) &= 1 - \mathrm{OR}(X),\\ \mathrm{Prob}(1) &= \mathrm{OR}(X).\end{aligned} \tag{6.179}$$

But these expressions are polynomials of degree N, which can arise only if the circuit queries the oracle at least T times, where

$$T \geq \frac{N}{2}. \tag{6.180}$$

We conclude that no quantum circuit with fewer than $N/2$ oracle calls can compute OR exactly. In fact, for this function (or any function that takes the value 0 for just one of its N possible arguments), there is a stronger conclusion (exercise): we require $T \geq N$ to evaluate OR with certainty.

On the other hand, evaluating the OR function (answering the yes/no question, "Is there a marked state?") is just what the Grover algorithm can achieve in a number of queries of order $\sqrt{N}$. Thus, while the conclusion is correct that N queries are needed to evaluate OR *with certainty*, this result is a bit misleading. We can evaluate OR *probabilistically* with far fewer queries. Apparently, the Grover algorithm can construct a polynomial in X that, though only of degree $O(\sqrt{N})$, provides a very adequate approximation to the N-th degree polynomial $\mathrm{OR}(X)$.

But OR, which takes the value 1 for every value of X except $X = \vec{0}$, is a very simple Boolean function. We should consider other functions that might pose a more serious challenge for the quantum computer.

One that comes to mind is the PARITY function: $\mathrm{PARITY}(X)$ takes the value 0 if the string X contains an even number of 1's, and the value 1 if the string contains an odd number of 1's. Obviously, a classical algorithm must query the oracle N times to determine the parity. How much better can we do by submitting quantum queries? In fact, we can't do much better at all – at least $N/2$ quantum queries are needed to find the correct value of $\mathrm{PARITY}(X)$, with probability of success greater than $\frac{1}{2} + \delta$.

In discussing PARITY it is convenient to use new variables

$$\tilde{X}_i = 1 - 2X_i, \tag{6.181}$$

that take values ± 1, so that

$$\mathrm{PARITY}(\tilde{X}) = \prod_{i=0}^{N-1} \tilde{X}_i, \tag{6.182}$$

also takes values ± 1. Now, after we execute a quantum circuit with altogether T queries of the oracle, we are to perform a POVM with two possible outcomes $\boldsymbol{F}_{\text{even}}$ and $\boldsymbol{F}_{\text{odd}}$; the outcome will be our estimate of $\mathrm{PARITY}(\tilde{X})$. As we have already noted, the probability of obtaining the outcome even (say) can be expressed as a polynomial $P^{(2T)}_{\text{even}}$ of degree (at most) $2T$ in $\tilde{X}$,

$$\langle \boldsymbol{F}_{\text{even}} \rangle = P^{(2T)}_{\text{even}}(\tilde{X}). \tag{6.183}$$

How often is our guess correct? Consider the sum

$$\begin{aligned}&\sum_{\{\tilde{X}\}} P^{(2T)}_{\text{even}}(\tilde{X}) \cdot \mathrm{PARITY}(\tilde{X})\\ &= \sum_{\{\tilde{X}\}} P^{(2T)}_{\text{even}}(\tilde{X}) \prod_{i=0}^{N-1} \tilde{X}_i.\end{aligned} \tag{6.184}$$

Since each term in the polynomial $P^{(2T)}_{\text{even}}(\tilde{X})$ contains at most $2T$ of the $\tilde{X}_i$'s, we can invoke the identity

$$\sum_{\tilde{X}_i \in \{0,1\}} \tilde{X}_i = 0, \tag{6.185}$$

to see that the sum in eq. (6.184) must vanish if $N > 2T$. We conclude that

$$\sum_{\mathrm{par}(\tilde{X})=1} P^{(2T)}_{\text{even}}(\tilde{X}) = \sum_{\mathrm{par}(\tilde{X})=-1} P^{(2T)}_{\text{even}}(\tilde{X}); \tag{6.186}$$

hence, for $T < N/2$, we are just as likely to guess "even" when the actual $\mathrm{PARITY}(\tilde{X})$ is odd as when it is even (on average). Our quantum algorithm fails to tell us anything about the value of PARITY($\tilde{X}$); that is, averaged over the (a priori equally likely) possible values of X_i, we are just as likely to be right as wrong.

We can also show, by exhibiting an explicit algorithm (exercise), that $N/2$ queries (assuming N even) are *sufficient* to determine PARITY (either probabilistically or deterministically.) In a sense, then, we can achieve a factor of 2 speedup compared to classical queries. But that is the best we can do.

6.8 Distributed database search

We will find it instructive to view the quantum database search algorithm from a fresh perspective. We imagine that two parties, Alice and Bob, need to arrange to meet on a mutually agreeable day. Alice has a calendar that lists $N = 2^n$ days, with each day marked by either a 0, if she is unavailable that day, or a 1, if she is available. Bob has a similar calendar. Their task is to find a day when they will both be available.

Alice and Bob both have quantum computers, but they are very far apart from one another. (Alice is on earth, and Bob has traveled to the Andromeda galaxy). Therefore, it is very expensive for them to communicate. They urgently need to arrange their date, but they must economize on the amount of information that they send back and forth.

Even if there exists a day when both are available, it might not be easy to find it. If Alice and Bob communicate by sending classical bits back and forth, then in the worst case they will need to exchange of order $N = 2^n$ calendar entries to have a reasonable chance of successfully arranging their date.. We will ask: can they do better by exchanging qubits instead?[13] (The quantum information highway from earth to Andromeda was carefully designed and constructed, so it does not cost much more to send qubits instead of bits.)

To someone familiar with the basics of quantum information theory, this sounds like a foolish question. Holevo's theorem told us once and for all that a single qubit can convey no more than one bit of classical information. On further reflection, though, we see that Holevo's theorem does not really settle the issue. While it bounds the mutual information of a state preparation with a measurement outcome, it does not assure us (at least not directly) that Alice and Bob need to exchange as many qubits as bits to compare their calendars. Even so, it comes as a refreshing surprise[14] to learn that Alice and Bob can do the job by exchanging $O(\sqrt{N}\log N)$ qubits, as compared to $O(N)$ classical bits.

To achieve this Alice and Bob must work in concert, implementing a distributed version of the database search. Alice has access to an oracle (her calendar) that computes a function $f_A(x)$, and Bob has an oracle (his calendar) that computes $f_B(x)$. Together, they can query the oracle

$$f_{AB}(x) = f_A(x) \wedge f_B(x) \ . \tag{6.187}$$

Either one of them can implement the reflection $\boldsymbol{U}_s$, so they can perform a complete Grover iteration, and can carry out exhaustive search for a suitable day x such that $f_{AB}(x) = 1$ (when Alice and Bob are both available). If a mutually agreeable day really exists, they will succeed in finding it after of order $\sqrt{N}$ queries.

How do Alice and Bob query f_{AB}? We'll describe how they do it acting on any one of the computational basis states $|x\rangle$. First Alice performs

$$|x\rangle|0\rangle \to |x\rangle|f_A(x)\rangle, \tag{6.188}$$

and then she sends the $n+1$ qubits to Bob. Bob performs

$$|x\rangle|f_A(x)\rangle \to (-1)^{f_A(x)\wedge f_B(x)}|x\rangle|f_A(x)\rangle. \tag{6.189}$$

This transformation is evidently unitary, and you can easily verify that Bob can implement it by querying his oracle. Now the phase multiplying $|x\rangle$ is $(-1)^{f_{AB}(x)}$ as desired, but $|f_A(x)\rangle$ remains stored in the other register, which

[13] In an earlier version of these notes, I proposed a different scenario, in which Alice and Bob had nearly identical tables, but with a single mismatched entry; their task was to find the location of the mismatched bit. However, that example was poorly chosen, because the task can be accomplished with only $\log N$ bits of classical communication. (Thanks to Richard Cleve for pointing out this blunder.) We want Alice to learn the address (a binary string of length n) of the one entry where her table differs from Bob's. So Bob computes the parity of the $N/2$ entries in his table with a label that takes the value 0 in its least significant bit, and he sends that one parity bit to Alice. Alice compares to the parity of the same entries in her table, and she infers one bit (the least significant bit) of the address of the mismatched entry. Then they do the same for each of the remaining $n-1$ bits, until Alice knows the complete address of the "error". Altogether just n bits are sent (and all from Bob to Alice).

[14] H. Burhman, *et al.*, "Quantum vs. Classical Communication and Computation," quant-ph/9802040.

would spoil the coherence of a superposition of x values. Bob cannot erase that register, but Alice can. So Bob sends the $n+1$ qubits back to Alice, and she consults her oracle once more to perform

$$(-1)^{f_A(x)\wedge f_B(x)}|x\rangle|f_A(x)\rangle \rightarrow (-1)^{f_A(x)\wedge f_B(x)}|x\rangle|0\rangle. \tag{6.190}$$

By exchanging $2(n+1)$ qubits, the have accomplished one query of the f_{AB} oracle, and so can execute one Grover iteration.

Suppose, for example, that Alice and Bob know that there is only one mutually agreeable date, but they have no *a priori* information about which date it is. After about $\frac{\pi}{4}\sqrt{N}$ iterations, requiring altogether

$$Q \cong \frac{\pi}{4}\sqrt{N}\cdot 2(\log N+1), \tag{6.191}$$

qubit exchanges, Alice measures, obtaining the good date with probability quite close to 1.

Thus, at least in this special context, exchanging $O(\sqrt{N}\log N)$ qubits is as good as exchanging $O(N)$ classical bits. Apparently, we have to be cautious in interpreting the Holevo bound, which ostensibly tells us that a qubit has no more information-carrying capacity than a bit!

If Alice and Bob don't know in advance how many good dates there are, they can still perform the Grover search (as we noted in §6.4.5), and will find a solution with reasonable probability. With $2\cdot\log N$ bits of classical communication, they can verify whether the date that they found is really mutually agreeable.

6.8.1 Quantum communication complexity

More generally, we may imagine that several parties each possess an n-bit input, and they are to evaluate a function of all the inputs, with one party eventually learning the value of the function. What is the minimum amount of communication needed to compute the function (either deterministically or probabilistically)? The well-studied branch of classical complexity theory that addresses this question is called *communication complexity.* What we established above is a quadratic separation between quantum and classical communication complexity, for a particular class of two-party functions.

Aside from replacing the exchange of classical bits by the exchange of qubits, there are other interesting ways to generalize classical communication complexity. One is to suppose that the parties share some preexisting entangled state (either Bell pairs or multipartite entanglement), and that they may exploit that entanglement along with classical communication to perform the function evaluation. Again, it is not immediately clear that the shared entanglement will make things any easier, since entanglement alone doesn't permit the parties to exchange classical messages. But it turns out that the entanglement *does* help, at least a little bit.[15]

The analysis of communication complexity is a popular past time among complexity theorists, but this discipline does not yet seem to have assumed a prominent position in practical communications engineering. Perhaps this is surprising, considering the importance of efficiently distributing the computational load in parallelized computing, which has become commonplace. Furthermore, it seems that nearly all communication in real life can be regarded as a form of remote computation. I don't really need to receive all the bits that reach me over the telephone line, especially since I will probably retain only a few bits of information pertaining to the call tomorrow (the movie we decided to go to). As a less prosaic example, we on earth may need to communicate with a robot in deep space, to instruct it whether to enter and orbit around a distant star system. Since bandwidth is extremely limited, we would like it to compute the correct answer to the Yes/No question "Enter orbit?" with minimal exchange of information between earth and robot.

Perhaps a future civilization will exploit the known quadratic separation between classical and quantum communication complexity, by exchanging qubits rather than bits with its flotilla of spacecraft. And perhaps an exponential separation will be found, at least in certain contexts, which would significantly boost the incentive to develop the required quantum communications technology.

6.9 Periodicity

So far, the one case for which we have found an exponential separation between the speed of a quantum algorithm and the speed of the corresponding

[15]R. Cleve, et al., "Quantum Entanglement and the Communication Complexity of the Inner Product Function," quant-ph/9708019; W. van Dam, et al., "Multiparty Quantum Communication Complexity," quant-ph/9710054.

classical algorithm is the case of Simon's problem. Simon's algorithm exploits quantum parallelism to speed up the search for the period of a function. Its success encourages us to seek other quantum algorithms designed for other kinds of period finding.

Simon studied periodic functions taking values in $(Z_2)^n$. For that purpose the n-bit Hadamard transform $\boldsymbol{H}^{(n)}$ was a powerful tool. If we wish instead to study periodic functions taking values in Z_{2^n}, the (discrete) Fourier transform will be a tool of comparable power.

The moral of Simon's problem is that, while finding needles in a haystack may be difficult, finding *periodically* spaced needles in a haystack can be far easier. For example, if we scatter a photon off of a periodic array of needles, the photon is likely to be scattered in one of a set of preferred directions, where the Bragg scattering condition is satisfied. These preferred directions depend on the spacing between the needles, so by scattering just one photon, we can already collect some useful information about the spacing. We should further explore the implications of this metaphor for the construction of efficient quantum algorithms.

So imagine a quantum oracle that computes a function

$$f : \{0,1\}^n \to \{0,1\}^m, \tag{6.192}$$

that has an unknown period r, where r is a positive integer satisfying

$$1 \ll r \ll 2^n. \tag{6.193}$$

That is,

$$f(x) = f(x + mr), \tag{6.194}$$

where m is any integer such that x and $x + mr$ lie in $\{0, 1, 2, \ldots, 2^n - 1\}$. We are to find the period r. Classically, this problem is *hard*. If r is, say, of order $2^{n/2}$, we will need to query the oracle of order $2^{n/4}$ times before we are likely to find two values of x that are mapped to the same value of $f(x)$, and hence learn something about r. But we will see that there is a quantum algorithm that finds r in time poly (n).

Even if we know how to compute efficiently the function $f(x)$, it may be a hard problem to determine its period. Our quantum algorithm can be applied to finding, in poly(n) time, the period of any function that we can compute in poly(n) time. Efficient period finding allows us to efficiently solve a variety of (apparently) hard problems, such as factoring an integer, or evaluating a discrete logarithm.

The key idea underlying quantum period finding is that the Fourier transform can be evaluated by an efficient quantum circuit (as discovered by Peter Shor). The quantum Fourier transform (QFT) exploits the power of quantum parallelism to achieve an exponential speedup of the well-known (classical) fast Fourier transform (FFT). Since the FFT has such a wide variety of applications, perhaps the QFT will also come into widespread use someday.

6.9.1 Finding the period

The QFT is the unitary transformation that acts on the computational basis according to

$$QFT : |x\rangle \to \frac{1}{\sqrt{N}} \sum_{y=0}^{N-1} e^{2\pi i x y/N} |y\rangle, \tag{6.195}$$

where $N = 2^n$. For now let's suppose that we can perform the QFT efficiently, and see how it enables us to extract the period of $f(x)$.

Emulating Simon's algorithm, we first query the oracle with the input $\frac{1}{\sqrt{N}} \sum_x |x\rangle$ (easily prepared by applying $\boldsymbol{H}^{(n)}$ to $|0\rangle$), and so prepare the state

$$\frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} |x\rangle |f(x)\rangle. \tag{6.196}$$

Then we measure the output register, obtaining the result $|f(x_0)\rangle$ for some $0 \le x_0 < r$. This measurement prepares in the input register the coherent superposition of the A values of x that are mapped to $f(x_0)$:

$$\frac{1}{\sqrt{A}} \sum_{j=0}^{A-1} |x_0 + jr\rangle, \tag{6.197}$$

where

$$N - r \le x_0 + (A-1)r < N, \tag{6.198}$$

or

$$A - 1 < \frac{N}{r} < A + 1. \tag{6.199}$$

Actually, the measurement of the output register is unnecessary. If it is omitted, the state of the input register is an incoherent superposition (summed over $x_0 \in \{0, 1, \ldots r-1\}$) of states of the form eq. (6.197). The rest of the algorithm works just as well acting on this initial state.

Now our task is to extract the value of r from the state eq. (6.197) that we have prepared. Were we to measure the input register by projecting onto the computational basis at this point, we would learn nothing about r. Instead (cf. Simon's algorithm), we should Fourier transform first and *then* measure.

By applying the QFT to the state eq. (6.197) we obtain

$$\frac{1}{\sqrt{NA}} \sum_{y=0}^{N-1} e^{2\pi i x_0 y} \sum_{j=0}^{A-1} e^{2\pi i j r y/N} |y\rangle. \tag{6.200}$$

If we now measure in the computational basis, the probability of obtaining the outcome y is

$$\text{Prob}(y) = \frac{A}{N} \left| \frac{1}{A} \sum_{j=0}^{A-1} e^{2\pi i j r y/N} \right|^2. \tag{6.201}$$

This distribution strongly favors values of y such that yr/N is close to an integer. For example, if N/r happened to be an integer (and therefore equal to A), we would have

$$\text{Prob}(y) = \frac{1}{r} \left| \frac{1}{A} \sum_{j=0}^{A-1} e^{2\pi i j y/A} \right| = \begin{cases} \frac{1}{r} & y = A \cdot \text{(integer)} \\ 0 & \text{otherwise.} \end{cases} \tag{6.202}$$

More generally, we may sum the geometric series

$$\sum_{j=0}^{A-1} e^{i\theta j} = \frac{e^{iA\theta} - 1}{e^{i\theta} - 1}, \tag{6.203}$$

where

$$\theta_y = \frac{2\pi y r (\text{mod } N)}{N}. \tag{6.204}$$

There are precisely r values of y in $\{0, 1, \ldots, N-1\}$ that satisfy

$$-\frac{r}{2} \leq yr(\text{mod } N) \leq \frac{r}{2}. \tag{6.205}$$

(To see this, imagine marking the multiples of r and N on a number line ranging from 0 to $rN - 1$. For each multiple of N, there is a multiple of r no more than distance $r/2$ away.) For each of these values, the corresponding θ_y satisfies.

$$-\pi \frac{r}{N} \leq \theta_y \leq \pi \frac{r}{N}. \tag{6.206}$$

Now, since $A - 1 < \frac{N}{r}$, for these values of θ_y all of the terms in the sum over j in eq. (6.203) lie in the same half-plane, so that the terms interfere constructively and the sum is substantial.

We know that

$$|1 - e^{i\theta}| \leq |\theta|, \tag{6.207}$$

because the straight-line distance from the origin is less than the arc length along the circle, and for $A|\theta| \leq \pi$, we know that

$$|1 - e^{iA\theta}| \geq \frac{2A|\theta|}{\pi}, \tag{6.208}$$

because we can see (either graphically or by evaluating its derivative) that this distance is a convex function. We actually have $A < \frac{N}{r} + 1$, and hence $A\theta_y < \pi \left(1 + \frac{r}{N}\right)$, but by applying the above bound to

$$\left| \frac{e^{i(A-1)\theta} - 1}{e^{i\theta} - 1} + e^{i(A-1)\theta} \right| \geq \left| \frac{e^{i(A-1)\theta} - 1}{e^{i\theta} - 1} \right| - 1, \tag{6.209}$$

we can still conclude that

$$\left| \frac{e^{iA\theta} - 1}{e^{i\theta} - 1} \right| \geq \frac{2(A-1)|\theta|}{\pi|\theta|} - 1 = \frac{2}{\pi}A - \left(1 + \frac{2}{\pi}\right). \tag{6.210}$$

Ignoring a possible correction of order $2/A$, then, we find

$$\text{Prob}(y) \geq \left(\frac{4}{\pi^2}\right) \frac{1}{r}, \tag{6.211}$$

for each of the r values of y that satisfy eq. (6.205). Therefore, with a probability of at least $4/\pi^2$, the measured value of y will satisfy

$$k\frac{N}{r} - \frac{1}{2} \leq y \leq k\frac{N}{r} + \frac{1}{2}, \tag{6.212}$$

or

$$\frac{k}{r} - \frac{1}{2N} \leq \frac{y}{N} \leq \frac{k}{r} + \frac{1}{2N}, \tag{6.213}$$

where k is an integer chosen from $\{0, 1, \ldots, r-1\}$. The output of the computation is reasonable likely to be within distance $1/2$ of an integer multiple of N/r.

Suppose that we know that $r < M \ll N$. Thus N/r is a rational number with a denominator less than M. Two distinct rational numbers, each with denominator less than M, can be no closer together than $1/M^2$, since $\frac{a}{b} - \frac{c}{d} = \frac{ad-bc}{bd}$. If the measurement outcome y satisfies eq. (6.212), then there is a *unique* value of k/r (with $r < M$) determined by y/N, provided that $N \geq M^2$. This value of k/r can be efficiently extracted from the measured y/N, by the continued fraction method.

Now, with probability exceeding $4/\pi^2$, we have found a value of k/r where k is selected (roughly equiprobably) from $\{0, 1, 2, \ldots, r-1\}$. It is reasonably likely that k and r are relatively prime (have no common factor), so that we have succeeded in finding r. With a query of the oracle, we may check whether $f(x) = f(x+r)$. But if $\mathrm{GCD}(k, r) \neq 1$, we have found only a factor (r_1) of r.

If we did not succeed, we could test some nearby values of y (the measured value might have been close to the range $-r/2 \leq yr (\mathrm{mod}\ N) \leq r/2$ without actually lying inside), or we could try a few multiples of r (the value of $\mathrm{GCD}(k, r)$, if not 1, is probably not large). That failing, we resort to a repetition of the quantum circuit, this time (with probability at least $4/\pi^2$) obtaining a value k'/r. Now k', too, may have a common factor with r, in which case our procedure again determines a factor (r_2) of r. But it is reasonably likely that $\mathrm{GCD}(k, k') = 1$, in which case $r = \mathrm{LCM}, (r_1, r_2)$. Indeed, we can estimate the probability that randomly selected k and k' are relatively prime as follows: Since a prime number p divides a fraction $1/p$ of all numbers, the probability that p divides both k and k' is $1/p^2$. And k and k' are coprime if and only if there is no prime p that divides both. Therefore,

$$\mathrm{Prob}(k, k' \text{ coprime}) = \prod_{\text{prime } p} \left(1 - \frac{1}{p^2}\right) = \frac{1}{\zeta(2)} = \frac{6}{\pi^2} \simeq .607 \tag{6.214}$$

(where $\zeta(z)$ denotes the Riemann zeta function). Therefore, we are likely to succeed in finding the period r after some constant number (independent of N) of repetitions of the algorithm.

6.9.2 From FFT to QFT

Now let's consider the implementation of the quantum Fourier transform. The Fourier transform

$$\sum_x f(x)|x\rangle \to \sum_y \left(\frac{1}{\sqrt{N}} \sum_x e^{2\pi i x y/N} f(x)\right) |y\rangle, \tag{6.215}$$

is multiplication by an $N \times N$ unitary matrix, where the (x, y) matrix element is $(e^{2\pi i/N})^{xy}$. Naively, this transform requires $O(N^2)$ elementary operations. But there is a well-known and very useful (classical) procedure that reduces the number of operations to $O(N \log N)$. Assuming $N = 2^n$, we express x and y as binary expansions

$$\begin{aligned} x &= x_{n-1} \cdot 2^{n-1} + x_{n-2} \cdot 2^{n-2} + \ldots + x_1 \cdot 2 + x_0 \\ y &= y_{n-1} \cdot 2^{n-1} + y_{n-2} \cdot 2^{n-2} + \ldots + y_1 \cdot 2 + y_0. \end{aligned} \tag{6.216}$$

In the product of x and y, we may discard any terms containing n or more powers of 2, as these make no contribution to $e^{2\pi i xy}/2^n$. Hence

$$\begin{aligned} \frac{xy}{2^n} \equiv\ & y_{n-1}(.x_0) + y_{n-2}(.x_1 x_0) + y_{n-3}(.x_2 x_1 x_0) + \ldots \\ & + y_1(.x_{n-2} x_{n-3} \ldots x_0) + y_0(.x_{n-1} x_{n-2} \ldots x_0), \end{aligned} \tag{6.217}$$

where the factors in parentheses are binary expansions; *e.g.*,

$$.x_2 x_1 x_0 = \frac{x_2}{2} + \frac{x_1}{2^2} + \frac{x_0}{2^3}. \tag{6.218}$$

We can now evaluate

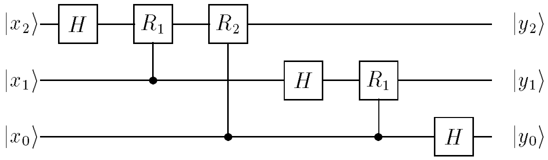
$$\tilde{f}(x) = \frac{1}{\sqrt{N}} \sum_y e^{2\pi i x y/N} f(y), \tag{6.219}$$

for each of the N values of x. But the sum over y factors into n sums over $y_k = 0, 1$, which can be done sequentially in a time of order n.

With quantum parallelism, we can do far better. From eq. (6.217) we obtain

$$\begin{aligned} QFT : |x\rangle \to\ & \frac{1}{\sqrt{N}} \sum_y e^{2\pi i x y/N} |y\rangle \\ =\ & \frac{1}{\sqrt{2^n}} \left(|0\rangle + e^{2\pi i(.x_0)}|1\rangle\right) \left(|0\rangle + e^{2\pi i(.x_1 x_0)}|1\rangle\right) \\ & \ldots \left(|0\rangle + e^{2\pi i(.x_{n-1} x_{n-2} \ldots x_0)}|1\rangle\right). \end{aligned} \tag{6.220}$$

The QFT takes each computational basis state to an *unentangled* state of n qubits; thus we anticipate that it can be efficiently implemented. Indeed, let's consider the case $n = 3$. We can readily see that the circuit



does the job (but note that the order of the bits has been reversed in the output). Each Hadamard gate acts as

$$\boldsymbol{H} : |x_k\rangle \to \frac{1}{\sqrt{2}}\left(|0\rangle + e^{2\pi i(.x_k)}|1\rangle\right). \tag{6.221}$$

The other contributions to the relative phase of $|0\rangle$ and $|1\rangle$ in the kth qubit are provided by the two-qubit conditional rotations, where

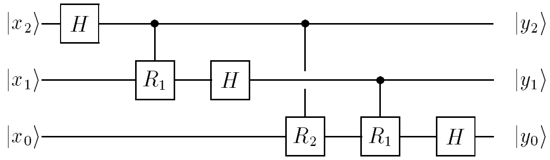
$$\boldsymbol{R}_d = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\pi/2^d} \end{pmatrix}, \tag{6.222}$$

and $d = (k - j)$ is the "distance" between the qubits.

In the case $n = 3$, the QFT is constructed from three $\boldsymbol{H}$ gates and three controlled-$\boldsymbol{R}$ gates. For general n, the obvious generalization of this circuit requires n $\boldsymbol{H}$ gates and $\binom{n}{2} = \frac{1}{2}n(n-1)$ controlled R's. A two qubit gate is applied to each pair of qubits, again with controlled relative phase $\pi/2^d$, where d is the "distance" between the qubits. Thus the circuit family that implements QFT has a size of order $(\log N)^2$.

We can reduce the circuit complexity to linear in $\log N$ if we are willing to settle for an implementation of fixed accuracy, because the two-qubit gates acting on distantly separated qubits contribute only exponentially small phases. If we drop the gates acting on pairs with distance greater than m, than each term in eq. (6.217) is replaced by an approximation to m bits of accuracy; the total error in $xy/2^n$ is certainly no worse than $n2^{-m}$, so we can achieve accuracy ε in $xy/2^n$ with $m \geq \log n/\varepsilon$. If we retain only the gates acting on qubit pairs with distance m or less, then the circuit size is $mn \sim n \log n/\varepsilon$.

In fact, if we are going to measure in the computational basis immediately after implementing the QFT (or its inverse), a further simplification is possible – no two-qubit gates are needed at all! We first remark that the controlled – $\boldsymbol{R}_d$ gate acts symmetrically on the two qubits – it acts trivially on $|00\rangle, |01\rangle$, and $|10\rangle$, and modifies the phase of $|11\rangle$ by $e^{i\theta_d}$. Thus, we can interchange the "control" and "target" bits without modifying the gate. With this change, our circuit for the 3-qubit QFT can be redrawn as:



Once we have measured $|y_0\rangle$, we *know* the value of the control bit in the controlled-$\boldsymbol{R}_1$ gate that acted on the first two qubits. Therefore, we will obtain the same probability distribution of measurement outcomes if, instead of applying controlled-$\boldsymbol{R}_1$ and then measuring, we instead measure y_0 first, and then apply $(\boldsymbol{R}_1)^{y_0}$ to the next qubit, conditioned on the outcome of the measurement of the first qubit. Similarly, we can replace the controlled-$\boldsymbol{R}_1$ and controlled-$\boldsymbol{R}_2$ gates acting on the third qubit by the single qubit rotation

$$(\boldsymbol{R}_2)^{y_0}(\boldsymbol{R}_1)^{y_1}, \tag{6.223}$$

(that is, a rotation with relative phase $\pi(.y_1y_0)$) *after* the values of y_1 and y_0 have been measured.

Altogether then, if we are going to measure after performing the QFT, only n Hadamard gates and $n-1$ single-qubit rotations are needed to implement it. The QFT is remarkably simple!

6.10 Factoring

6.10.1 Factoring as period finding

What does the factoring problem (finding the prime factors of a large composite positive integer) have to do with periodicity? There is a well-known (randomized) reduction of factoring to determining the period of a function. Although this reduction is not directly related to quantum computing, we will discuss it here for completeness, and because the prospect of using a quantum computer as a factoring engine has generated so much excitement.

Suppose we want to find a factor of the n-bit number N. Select pseudo-randomly $a < N$, and compute the greatest common divisor $\mathrm{GCD}(a, N)$, which can be done efficiently (in a time of order $(\log N)^3$) using the Euclidean algorithm. If $\mathrm{GCD}(a, N) \neq 1$ then the GCD is a nontrivial factor of N, and we are done. So suppose $\mathrm{GCD}(a, N) = 1$.

[**Aside: The Euclidean algorithm.** To compute $\mathrm{GCD}(N_1, N_2)$ (for $N_1 > N_2$) first divide N_1 by N_2 obtaining remainder R_1. Then divide N_2 by R_1, obtaining remainder R_2. Divide R_1 by R_2, *etc.* until the remainder is 0. The last nonzero remainder is $R = \mathrm{GCD}(N_1, N_2)$. To see that the algorithm works, just note that (1) R divides all previous remainders and hence also N_1 and N_2, and (2) *any* number that divides N_1 and N_2 will also divide all remainders, including R. A number that divides both N_1 and N_2, and also is divided by any number that divides both N_1 and N_2 must be $\mathrm{GCD}(N_1, N_2)$. To see how long the Euclidean algorithm takes, note that

$$R_j = qR_{j+1} + R_{j+2}, \tag{6.224}$$

where $q \geq 1$ and $R_{j+2} < R_{j+1}$; therefore $R_{j+2} < \frac{1}{2}R_j$. Two divisions reduce the remainder by at least a factor of 2, so no more than $2 \log N_1$ divisions are required, with each division using $O((\log N)^2)$ elementary operations; the total number of operations is $O((\log N)^3)$.]

The numbers $a < N$ coprime to N (having no common factor with N) form a finite group under multiplication mod N. [Why? We need to establish that each element a has an inverse. But for given $a < N$ coprime to N, each $ab \pmod N$ is distinct, as b ranges over all $b < N$ coprime to N.[16] Therefore, for some b, we must have $ab \equiv 1 \pmod N$; hence the inverse of a exists.] Each element a of this finite group has a finite *order* r, the smallest positive integer such that

$$a^r \equiv 1 \pmod N. \tag{6.225}$$

[16] If N divides $ab - ab'$, it must divide $b - b'$.

The order of a mod N is the period of the function

$$f_{N,a}(x) = a^x \pmod N. \tag{6.226}$$

We know there is an efficient quantum algorithm that can find the period of a function; therefore, if we can compute $f_{N,a}$ efficiently, we can find the order of a efficiently.

Computing $f_{N,a}$ may look difficult at first, since the exponent x can be very large. But if $x < 2^m$ and we express x as a binary expansion

$$x = x_{m-1} \cdot 2^{m-1} + x_{m-2} \cdot 2^{m-2} + \ldots + x_0, \tag{6.227}$$

we have

$$a^x (\mathrm{mod}\ N) = (a^{2^{m-1}})^{x_{m-1}} (a^{2^{m-2}})^{x_{m-2}} \ldots (a)^{x_0} \pmod N. \tag{6.228}$$

Each a^{2^j} has a large exponent, but can be computed efficiently by a *classical* computer, using repeated squaring

$$a^{2^j} \pmod N = (a^{2^{j-1}})^2 \pmod N. \tag{6.229}$$

So only $m - 1$ (classical) mod N multiplications are needed to assemble a table of all a^{2^j}'s.

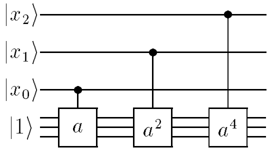
The computation of $a^x (\mathrm{mod}\ N)$ is carried out by executing a routine:

INPUT 1

For $j = 0$ to $m - 1$, if $x_j = 1$, MULTIPLY a^{2^j}.

This routine requires at most m mod N multiplications, each requiring of order $(\log N)^2$ elementary operations.[17] Since $r < N$, we will have a reasonable chance of success at extracting the period if we choose $m \sim 2 \log N$. Hence, the computation of $f_{N,a}$ can be carried out by a circuit family of size $O((\log N)^3)$. Schematically, the circuit has the structure:

[17] Using tricks for performing efficient multiplication of very large numbers, the number of elementary operations can be reduced to $O(\log N \log\log N \log\log\log N)$; thus, asymptotically for large N, a circuit family with size $O(\log^2 N \log\log N \log\log\log N)$ can compute $f_{N,a}$.



Multiplication by a^{2^j} is performed if the control qubit x_j has the value 1.

Suppose we have found the period r of a mod N. Then *if* r is even, we have

$$N \text{ divides } \left(a^{\frac{r}{2}}+1\right)\left(a^{\frac{r}{2}}-1\right). \tag{6.230}$$

We know that N does not divide $a^{r/2}-1$; if it did, the order of a would be $\leq r/2$. Thus, *if* it is also the case that N does not divide $a^{r/2}+1$, or

$$a^{r/2} \neq -1 \ (\mathrm{mod}\ N), \tag{6.231}$$

then N must have a nontrivial common factor with each of $a^{r/2}\pm 1$. Therefore, $\mathrm{GCD}(N, a^{r/2}+1) \neq 1$ is a factor (that we can find efficiently by a classical computation), and we are done.

We see that, once we have found r, we succeed in factoring N *unless* either (1) r is odd or (2) r is even and $a^{r/2} \equiv -1$ (mod N). How likely is success?

Let's suppose that N is a product of two prime factors $p_1 \neq p_2$,

$$N = p_1 p_2 \tag{6.232}$$

(this is actually the least favorable case). For each $a < p_1p_2$, there exist unique $a_1 < p_1$ and $a_2 < p_2$ such that

$$\begin{aligned} a &\equiv a_1 \quad (\mathrm{mod}\ p_1) \\ a &\equiv a_2 \quad (\mathrm{mod}\ p_2). \end{aligned} \tag{6.233}$$

Choosing a random $a < N$ is, therefore, equivalent to choosing random $a, < p_1$ and $a_2 < p_2$.

[**Aside**: We're using the **Chinese Remainder Theorem.** The a solving eq. (6.233) is unique because if a and b are both solutions, then both

p_1 and p_2 must divide $a-b$. The solution exists because every $a < p_1p_2$ solves eq. (6.233) for *some* a_1 and a_2. Since there are exactly p_1p_2 ways to choose a_1 and a_2, and exactly p_1p_2 ways to choose a, uniqueness implies that there is an a corresponding to each pair a_1, a_2.]

Now let r_1 denote the order of a_1 mod p_1 and r_2 denote the order of a_2 mod p_2. The Chinese remainder theorem tells us that $a^r \equiv 1$ (mod p_1p_2) is equivalent to

$$\begin{aligned} a_1^r &\equiv 1 \ (\mathrm{mod}\ p_1) \\ a_2^r &\equiv 1 \ (\mathrm{mod}\ p_2). \end{aligned} \tag{6.234}$$

Therefore $r = \mathrm{LCM}(r_1, r_2)$. If r_1 and r_2 are both odd, then so is r, and we lose.

But if *either* r_1 or r_2 is even, then so is r, and we are still in the game. If

$$\begin{aligned} a^{r/2} &\equiv -1 \quad (\mathrm{mod}\ p_1) \\ a^{r/2} &\equiv -1 \quad (\mathrm{mod}\ p_2). \end{aligned} \tag{6.235}$$

Then we have $a^{r/2} \equiv -1$ (mod p_1p_2) and we still lose. But if either

$$\begin{aligned} a^{r/2} &\equiv -1 \quad (\mathrm{mod}\ p_1) \\ a^{r/2} &\equiv 1 \quad (\mathrm{mod}\ p_2), \end{aligned} \tag{6.236}$$

or

$$\begin{aligned} a^{r/2} &\equiv 1 \quad (\mathrm{mod}\ p_1) \\ a^{r/2} &\equiv -1 \quad (\mathrm{mod}\ p_2), \end{aligned} \tag{6.237}$$

then $a^{r/2} \not\equiv -1$(mod p_1p_2) and we win. (Of course, $a^{r/2} \equiv 1$ (mod p_1) and $a^{r/2} \equiv 1$ (mod p_2) is not possible, for that would imply $a^{r/2} \equiv 1$ (mod p_1p_2), and r could not be the order of a.)

Suppose that

$$\begin{aligned} r_1 &= 2^{c_1} \cdot \mathrm{odd} \\ r_2 &= 2^{c_2} \cdot \mathrm{odd}, \end{aligned} \tag{6.238}$$

where $c_1 > c_2$. Then $r = \mathrm{LCM}(r_1, r_2) = 2r_2\cdot$ integer, so that $a^{r/2} \equiv 1$ (mod p_2) and eq. (6.236) is satisfied – we win! Similarly $c_2 > c_1$ implies eq. (6.237) – again we win. But for $c_1 = c_2$, $r = r_1 \cdot (\mathrm{odd}) = r_2 \cdot (\mathrm{odd}')$ so that eq. (6.235) is satisfied – in that case we lose.

Okay – it comes down to: for $c_1 = c_2$ we lose, for $c_1 \neq c_2$ we win. How likely is $c_1 \neq c_2$?

It helps to know that the multiplicative group mod p is cyclic – it contains a primitive element of order $p-1$, so that all elements are powers of the primitive element. [Why? The integers mod p are a finite *field*. If the group were not cyclic, the maximum order of the elements would be $q < p-1$, so that $x^q \equiv 1 \pmod{p}$ would have $p-1$ solutions. But that can't be: in a finite field there are no more than q qth roots of unity.]

Suppose that $p - 1 = 2^k \cdot s$, where s is odd, and consider the orders of all the elements of the cyclic group of order $p-1$. For brevity, we'll discuss only the case $k = 1$, which is the least favorable case for us. Then if b is a primitive (order $2s$) element, the even powers of b have odd order, and the odd powers of b have order $2\cdot$ (odd). In this case, then, $r = 2^c\cdot$ (odd) where $c \in \{0,1\}$, each occuring equiprobably. Therefore, if p_1 and p_2 are both of this (unfavorable) type, and a_1, a_2 are chosen randomly, the probability that $c_1 \neq c_2$ is $\frac{1}{2}$. Hence, once we have found r, our probability of successfully finding a factor is at least $\frac{1}{2}$, if N is a product of two distinct primes. If N has more than two distinct prime factors, our odds are even better. The method fails if N is a prime power, $N = p^\alpha$, but prime powers can be efficiently factored by other methods.

6.10.2 RSA

Does anyone care whether factoring is easy or hard? Well, yes, some people do.

The presumed difficulty of factoring is the basis of the security of the widely used RSA[18] scheme for public key cryptography, which you may have used yourself if you have ever sent your credit card number over the internet.

The idea behind public key cryptography is to avoid the need to exchange a secret key (which might be intercepted and copied) between the parties that want to communicate. The enciphering key is public knowledge. But using the enciphering key to infer the deciphering key involves a prohibitively difficult computation. Therefore, Bob can send the enciphering key to Alice and everyone else, but only Bob will be able to decode the message that Alice (or anyone else) encodes using the key. Encoding is a "one-way function" that is easy to compute but very hard to invert.

[18] For Rivest, Shamir, and Adleman

(Of course, Alice and Bob could have avoided the need to exchange the public key if they had decided on a private key in their previous clandestine meeting. For example, they could have agreed to use a long random string as a one-time pad for encoding and decoding. But perhaps Alice and Bob never anticipated that they would someday need to communicate privately. Or perhaps they did agree in advance to use a one-time pad, but they have now used up their private key, and they are loath to reuse it for fear that an eavesdropper might then be able to break their code. Now they are two far apart to safely exchange a new private key; public key cryptography appears to be their most secure option.)

To construct the public key Bob chooses two large prime numbers p and q. But he does not publicly reveal their values. Instead he computes the product

$$N = pq. \tag{6.239}$$

Since Bob knows the prime factorization of N, he also knows the value of the Euler function $\varphi(N)$ – the number of number less than N that are coprime with N. In the case of a product of two primes it is

$$\varphi(N) = N - p - q + 1 = (p-1)(q-1), \tag{6.240}$$

(only multiples of p and q share a factor with N). It is easy to find $\varphi(N)$ if you know the prime factorization of N, but it is hard if you know only N.

Bob then pseudo-randomly selects $e < \varphi(N)$ that is coprime with $\varphi(N)$. He reveals to Alice (and anyone else who is listening) the value of N and e, but nothing else.

Alice converts her message to ASCII, a number $a < N$. She encodes the message by computing

$$b = f(a) = a^e (\mathrm{mod}\ N), \tag{6.241}$$

which she can do quickly by repeated squaring. How does Bob decode the message?

Suppose that a is coprime to N (which is overwhelmingly likely if p and q are very large – anyway Alice can check before she encodes). Then

$$a^{\varphi(N)} \equiv 1 \pmod{N} \tag{6.242}$$

(Euler's theorem). This is so because the numbers less than N and coprime to N form a group (of order $\varphi(N)$) under mod N multiplication. The order of any group element must divide the order of the group (the powers of a form a subgroup). Since GCD$(e, \varphi(N) = 1$, we know that e has a multiplicative inverse $d = e^{-1} \bmod \varphi(N)$:

$$ed \equiv 1 \pmod{\varphi(N)}. \tag{6.243}$$

The value of d is Bob's closely guarded secret; he uses it to decode by computing:

$$\begin{aligned} f^{-1}(b) &= b^d \pmod N \\ &= a^{ed} \pmod N \\ &= a \cdot (a^{\varphi(N)})^{\text{integer}} \pmod N \\ &= a \pmod N. \end{aligned} \tag{6.244}$$

[**Aside:** How does Bob compute $d = e^{-1}$? The multiplicative inverse is a byproduct of carrying out the Euclidean algorithm to compute GCD$(e, \varphi(N)) = 1$. Tracing the chain of remainders from the bottom up, starting with $R_n = 1$:

$$\begin{aligned} 1 = R_n &= R_{n-2} - q_{n-1}R_{n-1} \\ R_{n-1} &= R_{n-3} - q_{n-2}R_{n-2} \\ R_{n-2} &= R_{n-4} - q_{n-3}R_{n-3} \\ &\quad etc. \ldots \end{aligned} \tag{6.245}$$

(where the q_j's are the quotients), so that

$$\begin{aligned} 1 &= (1 + q_{n-1}q_{n-2})R_{n-2} - q_{n-1}R_{n-3} \\ 1 &= (-q_{n-1} - q_{n-3}(1 + q_{n-1}q_{n-2}))R_{n-3} \\ &\quad + (1 + q_{n-1}q_{n-2})R_{n-4}, \\ &etc. \ldots . \end{aligned} \tag{6.246}$$

Continuing, we can express 1 as a linear combination of any two successive remainders; eventually we work our way up to

$$1 = d \cdot e + q \cdot \varphi(N), \tag{6.247}$$

and identify d as $e^{-1} \pmod{\varphi(N)}$.]

Of course, if Eve has a superfast factoring engine, the RSA scheme is insecure. She factors N, finds $\varphi(N)$, and quickly computes d. In fact, she does not really need to factor N; it is sufficient to compute the order modulo N of the encoded message $a^e \pmod N$. Since e is coprime with $\varphi(N)$, the order of $a^e \pmod N$ is the same as the order of a (both elements generate the same *orbit*, or cyclic subgroup). Once the order Ord(a) is known, Eve computes $\tilde d$ such that

$$\tilde d e \equiv 1 \pmod{\mathrm{Ord}(a)} \tag{6.248}$$

so that

$$(a^e)^{\tilde d} \equiv a \cdot (a^{\mathrm{Ord}(a)})^{\text{integer}} \pmod N \equiv a \pmod N, \tag{6.249}$$

and Eve can decipher the message. If our only concern is to defeat RSA, we run the Shor algorithm to find $r = \mathrm{Ord}(a^e)$, and we needn't worry about whether we can use r to extract a factor of N or not.

How important are such prospective cryptographic applications of quantum computing? When fast quantum computers are readily available, concerned parties can stop using RSA, or can use longer keys to stay a step ahead of contemporary technology. However, people with secrets sometimes want their messages to remain confidential for a while (30 years?). They may not be satisfied by longer keys if they are not confident about the pace of future technological advances.

And if they shun RSA, what will they use instead? Not so many suitable one-way functions are known, and others besides RSA are (or may be) vulnerable to a quantum attack. So there really is a lot at stake. If fast large scale quantum computers become available, the cryptographic implications may be far reaching.

But while quantum theory taketh away, quantum theory also giveth; quantum computers may compromise public key schemes, but also offer an alternative: secure quantum key distribution, as discussed in Chapter 4.

6.11 Phase Estimation

There is an alternative way to view the factoring algorithm (due to Kitaev) that deepens our insight into how it works: we can factor because we can measure efficiently and accurately the eigenvalue of a certain unitary operator.

Consider $a < N$ coprime to N, let x take values in $\{0, 1, 2, \ldots, N-1\}$, and let $\boldsymbol{U}_a$ denote the unitary operator

$$\boldsymbol{U}_a : |x\rangle \to |ax \ (\mathrm{mod}\ N)\rangle. \tag{6.250}$$

This operator is unitary (a permutation of the computational basis) because multiplication by a mod N is invertible.

If the order of a mod N is r, then

$$\boldsymbol{U}_a^r = \mathbf{1}. \tag{6.251}$$

It follows that all eigenvalues of $\boldsymbol{U}_a$ are rth roots of unity:

$$\lambda_k = e^{2\pi i k/r}, \quad k \in \{0, 1, 2, \ldots, r-1\}. \tag{6.252}$$

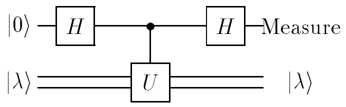
The corresponding eigenstates are

$$|\lambda_k\rangle = \frac{1}{\sqrt{r}} \sum_{j=0}^{r-1} e^{-2\pi i k j/r} |a^j x_0 (\mathrm{mod}\ N)\rangle; \tag{6.253}$$

associated with each orbit of length r generated by multiplication by a, there are r mutually orthogonal eigenstates.

$\boldsymbol{U}_a$ is not hermitian, but its *phase* (the Hermitian operator that generates $\boldsymbol{U}_a$) is an observable quantity. Suppose that we can perform a measurement that projects onto the basis of $\boldsymbol{U}_a$ eigenstates, and determines a value λ_k selected equiprobably from the possible eigenvalues. Hence the measurement determines a value of k/r, as does Shor's procedure, and we can proceed to factor N with a reasonably high success probability. But how do we measure the eigenvalues of a unitary operator?

Suppose that we can execute the unitary $\boldsymbol{U}$ conditioned on a control bit, and consider the circuit:



Here $|\lambda\rangle$ denotes an eigenstate of $\boldsymbol{U}$ with eigenvalue λ ($\boldsymbol{U}|\lambda\rangle = \lambda|\lambda\rangle$). Then the action of the circuit on the control bit is

$$\begin{aligned} |0\rangle &\to \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \to \frac{1}{\sqrt{2}}(|0\rangle + \lambda|1\rangle) \\ &\to \frac{1}{2}(1+\lambda)|0\rangle + \frac{1}{2}(1-\lambda)|1\rangle. \end{aligned} \tag{6.254}$$

Then the outcome of the measurement of the control qubit has probability distribution

$$\begin{aligned} \mathrm{Prob}(0) &= \left|\frac{1}{2}(1+\lambda)\right|^2 = \cos^2(\pi\phi) \\ \mathrm{Prob}(1) &= \left|\frac{1}{2}(1-\lambda)\right) |^2 = \sin^2(\pi\phi), \end{aligned} \tag{6.255}$$

where $\lambda = e^{2\pi i\phi}$.

As we have discussed previously (for example in connection with Deutsch's problem), this procedure distinguishes with certainty between the eigenvalues $\lambda = 1$ ($\phi = 0$) and $\lambda = -1$ ($\phi = 1/2$). But other possible values of λ can also be distinguished, albeit with less statistical confidence. For example, suppose the state on which $\boldsymbol{U}$ acts is a superposition of $\boldsymbol{U}$ eigenstates

$$\alpha_1|\lambda_1\rangle + \alpha_2|\lambda_2\rangle. \tag{6.256}$$

And suppose we execute the above circuit n times, with n distinct control bits. We thus prepare the state

$$\begin{aligned} &\alpha_1|\lambda_1\rangle \left(\frac{1+\lambda_1}{2}|0\rangle + \frac{1-\lambda_1}{2}|1\rangle\right)^{\otimes n} \\ &+\alpha_2|\lambda_2\rangle \left(\frac{1+\lambda_2}{2}|0\rangle + \frac{1-\lambda_2}{2}|1\rangle\right)^{\otimes n}. \end{aligned} \tag{6.257}$$

If $\lambda_1 \neq \lambda_2$, the overlap between the two states of the n control bits is exponentially small for large n; by measuring the control bits, we can perform the orthogonal projection onto the $\{|\lambda_1\rangle, |\lambda_2\rangle\}$ basis, at least to an excellent approximation.

If we use enough control bits, we have a large enough sample to measure Prob (0)$= \frac{1}{2}(1 + \cos 2\pi\phi)$ with reasonable statistical confidence. By executing a controlled-$(i\boldsymbol{U})$, we can also measure $\frac{1}{2}(1 + \sin 2\pi\phi)$ which suffices to determine ϕ modulo an integer.

However, in the factoring algorithm, we need to measure the phase of $e^{2\pi i k/r}$ to exponential accuracy, which seems to require an exponential number of trials. Suppose, though, that we can efficiently compute high powers of $\boldsymbol{U}$ (as is the case for $\boldsymbol{U}_a$) such as

$$\boldsymbol{U}^{2^j}. \tag{6.258}$$

By applying the above procedure to measurement of $\boldsymbol{U}^{2^j}$, we determine

$$\exp(2\pi i 2^j \phi), \tag{6.259}$$

where $e^{2\pi i\phi}$ is an eigenvalue of $\boldsymbol{U}$. Hence, measuring $\boldsymbol{U}^{2^j}$ to one bit of accuracy is equivalent to measuring the jth bit of the eigenvalue of $\boldsymbol{U}$.

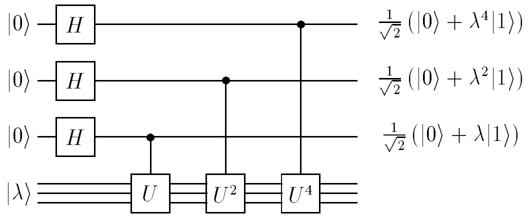
We can use this phase estimation procedure for order finding, and hence factorization. We invert eq. (6.253) to obtain

$$|x_0\rangle = \frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} |\lambda_k\rangle; \tag{6.260}$$

each computational basis state (for $x_0 \neq 0$) is an equally weighted superposition of r eigenstates of $\boldsymbol{U}_a$.

Measuring the eigenvalue, we obtain $\lambda_k = e^{2\pi i k/r}$, with k selected from $\{0, 1\ldots, r-1\}$ equiprobably. If $r < 2^n$, we measure to $2n$ bits of precision to determine k/r. In principle, we can carry out this procedure in a computer that stores fewer qubits than we would need to evaluate the QFT, because we can attack just one bit of k/r at a time.

But it is instructive to imagine that we incorporate the QFT into this phase estimation procedure. Suppose the circuit



acts on the eigenstate $|\lambda\rangle$ of the unitary transformation $\boldsymbol{U}$. The conditional $\boldsymbol{U}$ prepares $\frac{1}{\sqrt{2}}(|0\rangle + \lambda|1\rangle)$, the conditional $\boldsymbol{U}^2$ prepares $\frac{1}{\sqrt{2}}(|0\rangle + \lambda^2|1\rangle)$, the conditional $\boldsymbol{U}^4$ prepares $\frac{1}{\sqrt{2}}(|0\rangle + \lambda^4|1\rangle)$, and so on. We could perform a Hadamard and measure each of these qubits to sample the probability distribution governed by the jth bit of ϕ, where $\lambda = e^{2\pi i\phi}$. But a more efficient method is to note that the state prepared by the circuit is

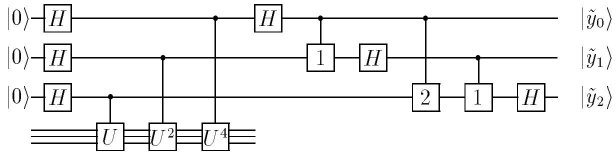
$$\frac{1}{\sqrt{2^m}} \sum_{y=0}^{2^m-1} e^{2\pi i\phi y}|y\rangle. \tag{6.261}$$

A better way to learn the value of ϕ is to perform the QFT$^{(m)}$, not the Hadamard $\boldsymbol{H}^{(m)}$, before we measure.

Considering the case $m = 3$ for clarity, the circuit that prepares and then Fourier analyzes the state

$$\frac{1}{\sqrt{8}} \sum_{y=0}^{7} e^{2\pi i\phi y}|y\rangle \tag{6.262}$$

is



This circuit very nearly carries out our strategy for phase estimation outlined above, but with a significant modification. Before we execute the final Hadamard transformation and measurement of $\tilde{y}_1$ and $\tilde{y}_2$, some conditional phase rotations are performed. It is those phase rotations that distinguish the QFT$^{(3)}$ from Hadamard transform $\boldsymbol{H}^{(3)}$, and they strongly enhance the reliability with which we can extract the value of ϕ.

We can understand better what the conditional rotations are doing if we suppose that $\phi = k/8$, for $k \in \{0, 1, 2\ldots, 7\}$; in that case, we know that the Fourier transform will generate the output $\tilde{y} = k$ with probability one. We may express k as the binary expansion

$$k = k_2k_1k_0 \equiv k_2 \cdot 4 + k_1 \cdot 2 + k_0. \tag{6.263}$$

In fact, the circuit for the least significant bit $\tilde{y}_0$ of the Fourier transform is precisely Kitaev's measurement circuit applied to the unitary $\boldsymbol{U}^4$, whose eigenvalue is

$$(e^{2\pi i\phi})^4 = e^{i\pi k} = e^{i\pi k_0} = \pm 1. \tag{6.264}$$

The measurement circuit distinguishes eigenvalues ± 1 perfectly, so that $\tilde{y}_0 = k_0$.

The circuit for the next bit $\tilde{y}_1$ is almost the measurement circuit for $\boldsymbol{U}^2$, with eigenvalue

$$(e^{2\pi i\phi})^2 = e^{i\pi k/2} = e^{i\pi(k_1 \cdot k_0)}. \tag{6.265}$$

Except that the conditional phase rotation has been inserted, which multiplies the phase by $\exp[i\pi(\cdot k_0)]$, resulting in $e^{i\pi k_1}$. Again, applying a Hadamard followed by measurement, we obtain the outcome $\tilde{y}_1 = k_1$ with certainty. Similarly, the circuit for $\tilde{y}_2$ measures the eigenvalue

$$e^{2\pi i\phi} = e^{i\pi k/4} = e^{i\pi(k_2 \cdot k_1 k_0)}, \tag{6.266}$$

except that the conditional rotation removes $e^{i\pi(\cdot k_1 k_0)}$, so that the outcome is $\tilde{y}_2 = k_2$ with certainty.

Thus, the QFT implements the phase estimation routine with maximal cleverness. We measure the less significant bits of ϕ first, and we exploit the information gained in the measurements to improve the reliability of our estimate of the more significant bits. Keeping this interpretation in mind, you will find it easy to remember the circuit for the $\mathrm{QFT}^{(n)}$!

6.12 Discrete Log

Sorry, I didn't have time for this.

6.13 Simulation of Quantum Systems

Ditto.

6.14 Summary

Classical circuits. The complexity of a problem can be characterized by the size of a uniform family of logic circuits that solve the problem: The problem is hard if the size of the circuit is a superpolynomial function of the size of the input. One classical universal computer can simulate another efficiently, so the classification of complexity is machine independent. The 3-bit Toffoli gate is universal for classical reversible computation. A reversible computer can simulate an irreversible computer without a significant slowdown and without unreasonable memory resources.

Quantum Circuits. Although there is no proof, it seems likely that polynomial-size quantum circuits cannot be simulated by polynomial-size probabilistic classical circuits ($BQP \neq BPP$); however, polynomial space is sufficient ($BQP \subseteq PSPACE$). A noisy quantum circuit can simulate an ideal quantum circuit of size T to acceptable accuracy if each quantum gate has an accuracy of order $1/T$. One universal quantum computer can simulate another efficiently, so that the complexity class BQP is machine independent. A generic two-qubit quantum gate, if it can act on any two qubits in a device, is adequate for universal quantum computation. A controlled-NOT gate plus a generic one-qubit gate is also adequate.

Fast Quantum Searching. Exhaustive search for a marked item in an unsorted database of N items can be carried out by a quantum computer in a time of order $\sqrt{N}$, but no faster. Quadratic quantum speedups can be achieved for some structured search problems, too, but some oracle problems admit no significant quantum speedup. Two parties, each in possession of a table with N entries, can locate a "collision" between their tables by exchanging $O(\sqrt{N})$ qubits, in apparent violation of the spirit (but not the letter) of the Holevo bound.

Period Finding. Exploiting quantum parallelism, the Quantum Fourier Transform in an N-dimensional space can be computed in time of order $(\log N)^2$ (compared to time $N \log N$ for the classical fast Fourier transform); if we are to measure immediately afterward, one qubit gates are sufficient to compute the QFT. Thus quantum computers can efficiently solve certain problems with a periodic structure, such as factoring and the discrete log problem.

6.15 Exercises

6.1 Linear simulation of Toffoli gate.

In class we constructed the n-bit Toffoli gate ($\theta^{(n)}$) from 3-bit Toffoli gates ($\theta^{(3)}$'s). The circuit required only one bit of scratch space, but the number of gates was exponential in n. With more scratch, we can substantially reduce the number of gates.

a) Find a circuit family with $2n-5$ $\theta^{(3)}$'s that evaluates $\theta^{(n)}$. (Here $n-3$ scratch bits are used, which are set to 0 at the beginning of the computation and return to the value 0 at the end.)

b) Find a circuit family with $4n-12$ $\theta^{(3)}$'s that evaluates $\theta^{(n)}$, which works irrespective of the initial values of the scratch bits. (Again the $n-3$ scratch bits return to their initial values, but they don't need to be set to zero at the beginning.)

6.2 A universal quantum gate set.

The purpose of this exercise is to complete the demonstration that the controlled-NOT and arbitrary one-qubit gates constitute a universal set.

a) If $\boldsymbol{U}$ is any unitary 2×2 matrix with determinant one, find unitary $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ such that

$$\boldsymbol{ABC} = \mathbf{1} \tag{6.267}$$

$$\boldsymbol{A\sigma}_x\boldsymbol{B\sigma}_x\boldsymbol{C} = \boldsymbol{U}. \tag{6.268}$$

Hint: From the Euler angle construction, we know that

$$\boldsymbol{U} = \boldsymbol{R}_z(\psi)\boldsymbol{R}_y(\theta)\boldsymbol{R}_z(\phi), \tag{6.269}$$

where, *e.g.*, $\boldsymbol{R}_z(\phi)$ denotes a rotation about the z-axis by the angle ϕ. We also know that, *e.g.*,

$$\boldsymbol{\sigma}_x\boldsymbol{R}_z(\phi)\boldsymbol{\sigma}_x = \boldsymbol{R}_z(-\phi). \tag{6.270}$$

b) Consider a two-qubit *controlled phase gate*: it applies $\boldsymbol{U} = e^{i\alpha}\mathbf{1}$ to the second qubit if the first qubit has value $|1\rangle$, and acts trivially otherwise. Show that it is actually a one-qubit gate.

c) Draw a circuit using controlled-NOT gates and single-qubit gates that implements controlled-$\boldsymbol{U}$, where $\boldsymbol{U}$ is an arbitrary 2×2 unitary transformation.

6.3 Precision.

The purpose of this exercise is to connect the accuracy of a quantum state with the accuracy of the corresponding probability distribution.

a) Let $\|\ \boldsymbol{A}\ \|_{\text{sup}}$ denote the sup norm of the operator $\boldsymbol{A}$, and let

$$\|\ \boldsymbol{A}\ \|_{\text{tr}} = \text{tr}\left[(\boldsymbol{A}^\dagger\boldsymbol{A})^{1/2}\right], \tag{6.271}$$

denote its *trace norm*. Show that

$$\|\ \boldsymbol{AB}\ \|_{\text{tr}} \le \|\ \boldsymbol{B}\ \|_{\text{sup}} \cdot \|\ \boldsymbol{A}\ \|_{\text{tr}} \quad \text{and} \quad |\ \text{tr}\ \boldsymbol{A}\ |\ \le \|\ \boldsymbol{A}\ \|_{\text{tr}}\ . \tag{6.272}$$

b) Suppose $\boldsymbol{\rho}$ and $\tilde{\boldsymbol{\rho}}$ are two density matrices, and $\{|a\rangle\}$ is a complete orthonormal basis, so that

$$P_a = \langle a|\boldsymbol{\rho}|a\rangle,$$

$$\tilde{P}_a = \langle a|\tilde{\boldsymbol{\rho}}|a\rangle, \tag{6.273}$$

are the corresponding probability distributions. Use (a) to show that

$$\sum_a |P_a - \tilde{P}_a|\ \le \|\ \boldsymbol{\rho} - \tilde{\boldsymbol{\rho}}\ \|_{\text{tr}}\ . \tag{6.274}$$

c) Suppose that $\boldsymbol{\rho} = |\psi\rangle\langle\psi|$ and $\tilde{\boldsymbol{\rho}} = |\tilde{\psi}\rangle\langle\tilde{\psi}|$ are pure states. Use (b) to show that

$$\sum_a |P_a - \tilde{P}_a|\ \le\ 2\ \|\ |\psi\rangle - |\tilde{\psi}\rangle\ \|\ . \tag{6.275}$$

6.4 Continuous-time database search

A quantum system with an n-qubit Hilbert space has the Hamiltonian

$$\boldsymbol{H}_\omega = E|\omega\rangle\langle\omega|, \tag{6.276}$$

where $|\omega\rangle$ is an unknown computational-basis state. You are to find the value of ω by the following procedure. Turn on a time-independent perturbation $\boldsymbol{H}'$ of the Hamiltonian, so that the total Hamiltonian becomes

$$\boldsymbol{H} = \boldsymbol{H}_\omega + \boldsymbol{H}'. \tag{6.277}$$

Prepare an initial state $|\psi_0\rangle$, and allow the state to evolve, as governed by $\boldsymbol{H}$, for a time T. Then measure the state. From the measurement result you are to infer ω.

a) Suppose the initial state is chosen to be

$$|s\rangle = \frac{1}{2^{n/2}} \sum_{x=0}^{2^n-1} |x\rangle, \tag{6.278}$$

and the perturbation is

$$\boldsymbol{H}' = E|s\rangle\langle s|. \tag{6.279}$$

Solve the time-independent Schrödinger equation

$$i\frac{d}{dt}|\psi\rangle = \boldsymbol{H}|\psi\rangle \tag{6.280}$$

to find the state at time T. How should T be chosen to optimize the likelihood of successfully determining ω?

b) Now suppose that we may choose $|\psi_0\rangle$ and $\boldsymbol{H}'$ however we please, but we demand that the state of the system after time T is $|\omega\rangle$, so that the measurement determines ω with success probability one. Derive a lower bound that T must satisfy, and compare to your result in (**a**). (**Hint:** As in our analysis in class, compare evolution governed by $\boldsymbol{H}$ with evolution governed by $\boldsymbol{H}'$ (the case of the "empty oracle"), and use the Schrödinger equation to bound how rapidly the state evolving according to $\boldsymbol{H}$ deviates from the state evolving according to $\boldsymbol{H}'$.)