Course Notes for CSCE 790-002 Quantum Computing and Information Fall 2021

Stephen A. Fenner*
Computer Science and Engineering Department
University of South Carolina

September 7, 2021

Abstract

These notes are mainly for me to lecture with, but you may find them useful to see what was covered when. All exercises are due one week from when they are assigned. These notes are subject to change during the semester. The date shown above is the date of the latest version.

*Columbia, SC 29208 USA. E-mail: fenner@cse.sc.edu. This material is based upon work supported by the National Science Foundation under Grant Nos. CCF-0515269 and CCF-0915948. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF).

Contents

1 Week 1: Overview

Brief, vague history of quantum mechanics, informatics, and the combination of the two.

Quantum Theory The foundations of quantum mechanics were established "by committee": Niels Bohr, Albert Einstein, Werner Heisenberg, Erwin Schrödinger, Max Planck, Louis de Broglie, Max Born, John von Neumann, Paul A.M. Dirac, Wolfgang Pauli, and others over the first half of the 20th century. The theory provides extremely accurate descriptions of the world at the atomic and subatomic levels, where "classical" (*i.e.*, Newtonian) physics and electrodynamics break down. Examples: stability of atoms, black body radiation, sharp spectral absorption lines, etc.

Informatics Broadly, this is the study of all aspects of information—its storage, transmission, and manipulation (*i.e.*, computation). It includes what is commonly called Computer Science in the US, as well as Information Theory. Foundations of Computer Science were laid at about the same time as quantum mechanics by Gottlob Frege, David Hilbert, Alonzo Church, Haskell Curry, Kurt Gödel, John Barkley Rosser, Alan Turing, Jacques Herbrand, Emil Post, Stephen Kleene and others, who were developing a formal notion of "algorithm" or "effective procedure" to understand problems in the foundations of mathematics. Foundations of computability culminated in the *Church-Turing thesis*. Largly independently, the field of Information Theory started in 1948 with Claude Shannon's paper, "A Mathematical Theory of Communication." Information theory deals with quantifying information and understanding how it can be stored and transmitted, both securely and otherwise. Shannon defined the notion of information entropy, somewhat analogously to physical entropy, and proved engineering-related results about compression and noisy transmission that are in common use today.

Quantum Information and Computation The physicist Richard Feynman first suggested the idea of a quantum computer and what it could be used for. Charles Bennett (80s?) showed that reversible computation (with no heat dissipation or entropy increase) was possible at least in principle. Paul Benioff (80s) showed how quantum dynamics could be used to simulate classical (reversible) computation, David Deutsch (80s) defined the Quantum Turing Machine (QTM) and quantum circuits as theoretical models of a quantum computer. Further foundational work was done by Bernstein & Vazirani, Yao, and others (quantum complexity theory). Bennett and Gilles Brassard (1984) proposed a scheme for unconditionally secure cryptographic key exchange based on quantum mechanical principles, using polarized photons. Deutsch & Jozsa and Simon (early 90s) gave "toy" problems on which quantum computers performed provably better than classical ones. A big breakthrough came in the mid 1990s when Peter Shor showed how a quantum computer can factor large integers quickly (1994), as well as compute discrete logarithms (these would break the security of most public key encryption schemes in use today). Grover (1996?) proposed a completely different quantum algorithm to quadratically speed up list search. Calderbank & Shor and Steane (1996?) showed that good quantum error-correcting codes exist and that fault-tolerant quantum computation is possible. This led to the *threshold theorem* (D. Aharonov, A. Yu. Kitaev(?)), which states that there is a constant $\varepsilon_0 > 0$ (current rough estimates are around 10^{-4}) such that if the noise associated with each gate can be kept below ε_0, then any quantum

computation can be carried out with arbitrarily small probability of error. This theorem shows that noise is not a fundamental impediment to quantum computation.

Implementations of Quantum Computers (the Bad News). There are several proposals for physical devices implementing the elements of quantum computation. Each has its own strengths and weaknesses. In recent years, ion traps look the most promising. We're still far off from a viable, scalable, robust prototype.

Nuclear Magnetic Resonance (NMR) Quantum bits are nuclei of atoms (hydrogen?) arranged on an organic molecule. The value of the bit is given by the spin of the nucleus. Nuclear spins can be controlled by electromagnetic pulses of the right frequency and duration. Main advantage: spins are well shielded from the outside by the electron clouds surrounding them, so they stay coherent for a long time. Main disadvantage: since the nuclei need to be on same molecule to control the distances between them, NMR does not scale well. Homay Valafar will talk about NMR toward the end of the course.

Ions in traps Qubits are ions kept equally spaced in a row (a couple of inches apart) by an oscillating electric field. Laser pulses can control the states of the ions.

Quantum dots Qubits are particles (electrons?) kept in nanoscopic wells on the surface of a silicon chip. Main advantage: easy to control and fabricate (solid state). Main disadvantage: short decoherence times.

Optical schemes Qubits are polarized photons traveling through mirrors, lenses, crystals, and the vacuum. Main advantages: photons don't decay and their polarizations are easy to measure; computation is at the speed of light. Main disadvantage: hard to get photons to interact with each other.

Superconducting/Josephson junctions I don't know much about this, except that it presumably needs temperatures close to absolute zero.

Implementations of Quantum Cryptography (the Good News). Quantum crypto not only works in the real world, but works just fine on fiber optic networks already in place. British Telecomm (mid 1990s?) demonstrated the BB84 quantum key exchange protocol using cable laid across Lake Geneva in Switzerland. I believe the scheme has also been demonstrated to work with photons through the air over modest distances (a few kilometers?). It is now feasible to use the fiber optic cable already in place to implement quantum crypto in the network of a major city (New York banks are already using it(?)). It still won't work over really large distances without classical repeaters ("quantum amplification" is theoretically impossible).

2 Week 1: Preliminaries

Just Enough Linear Algebra to Understand Just Enough Quantum Mechanics. We let $\mathbb{Z}$ denote the set of integers, $\mathbb{Q}$ denote the set of rational numbers, $\mathbb{R}$ denote the set of real numbers, and $\mathbb{C}$ denote the set of complex numbers.

The Complex Numbers. $\mathbb{C}$ is the set of all numbers of the form $z = x + iy$, where $x, y \in \mathbb{R}$ and $i^2 = -1$. We often represent z as the point (x, y) in the plane. The *complex conjugate* (or *adjoint*) of z is

$$z^* = \bar{z} = x - iy.$$

Note that $x = (z + z^*)/2$ and is the real part of z ($\Re(z)$). Similarly, $y = (z - z^*)/2i$ is the imaginary part of z ($\Im(z)$). The *norm* or *absolute value* of z is

$$|z| = \sqrt{z^* z} = \sqrt{x^2 + y^2} \geqslant 0,$$

with equality holding iff $z = 0$. If $z_1, z_2 \in \mathbb{C}$, it's easy to check that $|z_1 z_2| = |z_1| \cdot |z_2|$. It's not quite so easy to check that

$$|z_1 + z_2| \leqslant |z_1| + |z_2|\,, \tag{1}$$

but see Corollary B.2 in Section B.1 for a proof. (1) is an example of a *triangle inequality*.

Exercise 2.1 Let $z := 3 - 7i$ and $w := -1 + 2i$. Find (a) $z + w$, (b) zw, (c) $|z|$, (d) z^*, and (e) $1/w$.

Exercise 2.2 Check that $(z_1 z_2)^* = z_1^* z_2^*$ and $(z_1 + z_2)^* = z_1^* + z_2^*$ and $(-z_1)^* = -z_1^*$ for all $z_1, z_2 \in \mathbb{C}$. Express each answer in the form $x + iy$ for real x, y.

If $z \neq 0$, then the *argument* of z ($\arg(z)$) is defined as the angle that z makes with the positive real axis. Our convention will be that $0 \leqslant \arg(z) < 2\pi$. It is known that $\arg(z_1 z_2) = \arg(z_1) + \arg(z_2)$ up to a multiple of 2π.

The real numbers $\mathbb{R}$ forms a subset of $\mathbb{C}$ consisting of those complex numbers with 0 imaginary part, namely,

$$\mathbb{R} = \{z \in \mathbb{C} : z = z^*\}.$$

The *unit circle* in $\mathbb{C}$ is the set of all z of unit norm, *i.e.*, $\{z \in \mathbb{C} : |z| = 1\}$.

$\mathbb{C}$ is an *algebraically closed* field. That is, every polynomial of positive degree with coefficients in $\mathbb{C}$ has a root in $\mathbb{C}$, in fact n of them, where n is the degree of the polynomial. This is equivalent to saying that every polynomial over $\mathbb{C}$ is a product of linear (*i.e.*, degree 1) factors. This fact is known as the Fundamental Theorem of Algebra.

Every polynomial over $\mathbb{R}$ can be factored into real polynomial factors of degrees 1 and 2. This implies that any odd-degree real polynomial has at least one real root.

The Exponential Map. For any z, we can define $e^z = \exp(z)$ by the usual power series:

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots + \frac{z^k}{k!} + \cdots, \tag{2}$$

which converges for all z.

Here are some essential properties of the exponential map on $\mathbb{C}$:

- For all $z, w \in \mathbb{C}$, $e^{z+w} = e^z e^w$.

- $e^0 = 1$.
- $e^{-z} = 1/e^z$.
- $e^z \neq 0$.

Exercise 2.3 Show that for any real θ,

$$e^{i\theta} = \cos\theta + i\sin\theta\,. \tag{3}$$

This very important identity is known as *Euler's formula*, named after the Swiss mathematician Leonhard Euler. It connects the exponential and trigonometric functions. [Hint: Compare the power series for $e^{i\theta}$ with those for $\sin\theta$ and $\cos\theta$. Other proofs exist.]

Exercise 2.4 Using Euler's formula (from the previous exercise), find $e^{i\pi/2}$ and $e^{-i\pi/3}$. Express each answer in the form $x + iy$ for real x, y.

Exercise 2.5 Show that for all $\theta \in \mathbb{R}$,

$$\begin{aligned} \cos\theta &= \frac{e^{i\theta} + e^{-i\theta}}{2}\,, \\ \sin\theta &= \frac{e^{i\theta} - e^{-i\theta}}{2i}\,. \end{aligned}$$

Thus we can express the basic trigonometric functions in terms of exponentials. [Hint: Use Euler's formula and the fact that $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$.]

By Exercise 2.3, we have $e^z = e^x(\cos y + i\sin y)$. The unit circle is the set $\{e^{i\theta} : \theta \in \mathbb{R}\}$.

Vector Spaces. We'll deal with finite dimensional vector spaces *only*. Much of quantum mechanics requires infinite dimensional spaces, but thankfully, the QM that relates to information and computation only requires finite dimensions. So all our vector spaces are finite dimensional.

Our vector spaces will usually be over $\mathbb{C}$, the field of complex numbers, but sometimes they will be over $\mathbb{R}$ (*i.e.*, real vector spaces), and when we do information theory, will need to look at bit vectors (vectors in spaces over the two-element field $\mathbb{Z}_2 = \{0, 1\}$).

In a vector space, vectors can be added to each other and multiplied by scalars, obeying the usual rules. If V is an n-dimensional vector space and $\mathcal{B} = \{b_1, \ldots, b_n\}$ is a basis for V, then every $v \in V$ is written as a linear combination of basis vectors:

$$v = a_1 b_1 + \cdots + a_n b_n,$$

where $a_1, \ldots, a_n$ are unique scalars. Thus we can identify the vector v with the n-tuple

$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix},$$

which we may also write as $(a_1, \ldots, a_n)$. Under this identification, vector addition and scalar multiplication are componentwise.

The vector $(0, \ldots, 0)$ is the *zero vector*, denoted by 0.

Matrices. For integers $m, n > 0$, an $m \times n$ matrix is a rectangular array of scalars with m rows and n columns. If A is such a matrix and $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$, we denote the (i, j)th entry of A (*i.e.*, the scalar in the ith row and jth column) as $[A]_{ij}$ or $A[i, j]$. The former notation is useful if the matrix is given by a more complicated expression.

A matrix A is *upper triangular* if $[A]_{ij} = 0$ whenever $i > j$. A is *lower triangular* if $[A]_{ij} = 0$ whenever $i < j$. A is *triangular* if A is either upper or lower triangular. If A is both upper and lower triangular, then we can say that A is a *diagonal matrix*. In this case, all nonzero entries of A must lie on the main diagonal. Triangular matrices have some nice properties that make them simple to work with in some cases.

Adding and Multiplying Matrices. Given positive integers m and n and $m \times n$ matrices A and B, we can define the *matrix sum* $A + B$ to be the unique $m \times n$ matrix satisfying

$$[A + B]_{ij} = [A]_{ij} + [B]_{ij}$$

for all $1 \leqslant i \leqslant m$ and $1 \leqslant i \leqslant n$. That is, one just adds corresponding entries in A and B for the corresponding entry in the sum. For this to be well defined, A and B must have the same dimensions, in which case we say that A and B are *conformant* (for matrix addition); otherwise, $A + B$ is undefined. If k is a scalar, we can define the *scalar multiplication* of k with A as the unique $m \times n$ matrix kA satisfying

$$[kA]_{ij} = k[A]_{ij}$$

for all i and j as above. One just multiplies each entry of A by k to get the corresponding entry of kA. One can also write Ak for the same matrix. As you may expect, we write $-A$ for $(-1)A$ and write $A - B$ for $A + (-B)$. If $k \neq 0$, we can also write A/k for $(1/k)A$.

For positive integers m, n, and s, suppose A is an $m \times s$ matrix and B is an $s \times n$ matrix. Then we define the *matrix product* of A and B as the unique $m \times n$ matrix AB satisfying

$$[AB]_{ij} = \sum_{k=1}^{s} [A]_{ik}[B]_{kj}$$

for all $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. Note that the number of columns of A must equal the number of rows of B for the product to be well-defined, in which case we say that A and B are *conformant* (for matrix multiplication).

Most of the usual laws of addition and multiplication of scalars extend to matrices. In each identity below, we use A, B, C to stand for arbitrary matrices, I for a unit matrix, and k and ℓ for any scalars. For each identity, one side is well-defined if and only if the other side is well-defined.

Commutativity of matrix $+$: $A + B = B + A$.

Associativity of matrix $+$: $(A + B) + C = A + (B + C)$.

Identity for matrix $+$: $0 + A = A$.

Matrix negation: $A - A = 0$.

Associativity of scalar $\times$: $(k\ell)A = k(\ell A)$.

Distributivity of scalar $\times$ over matrix $+$: $k(A + B) = kA + kB$.

Distributivity of scalar $\times$ over scalar $+$: $(k + \ell)A = kA + \ell A$.

Identity for scalar $\times$: $1A = A$.

Associativity of matrix $\times$: $(AB)C = A(BC)$.

Distributivity of matrix $\times$ over matrix $+$: $A(B + C) = AB + AC$.

Distributivity of matrix $\times$ over matrix $+$: $(B + C)A = BA + CA$.

Commutativity and associativity of scalar $\times$: $k(AB) = (kA)B = A(kB)$.

There is no commutative law for matrix multiplication; that is, it is not generally true that $AB = BA$ for matrices A and B, even if both sides are well-defined. If this equation does hold, then we say that A and B *commute*.

Exercise 2.6 Find two 2×2 matrices A and B such that $AB = 0$ (the zero matrix), but $BA \neq 0$.

The Identity Matrix. For any n, the $n \times n$ *identity matrix* or *unit matrix* I_n has 1's on its main diagonal and 0's everywhere off the diagonal, so

$$[I_n]_{ij} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

This equation also defines the expression δ_{ij}, which is called the *Kronecker delta*. For example,

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

Unit matrices have the property that for any matrix A (say, $p \times q$),

$$I_p A = A I_q = A .$$

We may drop the subscript (I instead of I_n) if the dimension is clear from the context.

Nonsingular Matrices. A square matrix A is *nonsingular* or *invertible* iff there exists a matrix B of the same dimensions such that $AB = BA = I$. Such a B, if it exists, is uniquely determined by A and is denoted A^{-1}. In this case, it is of course true that B is also nonsingular and that $B^{-1} = A$.

Determinant. For an $n \times n$ matrix A, the *determinant* of A, denoted $\det A$, is a scalar value that depends on the entries and their positions inside A. A compact expression for the determinant is beyond the scope of this course, and besides, we won't deal with it very much, except to define eigenvalues and eigenvectors. But at least for the record we can say (without proof) that the map det mapping $n \times n$ matrices to scalars is the unique map satisfying the following two properties:

1. $\det(AB) = (\det A)(\det B)$ for all $n \times n$ matrices A and B.
2. For any $n \times n$ triangular matrix A, $\det A$ is the product of all the main diagonal elements of A, *i.e.*, $\det A = \prod_{i=1}^{n} [A]_{ii}$.

One fundamental fact about the determinant is that a matrix A is nonsingular if and only if $\det A \neq 0$.

Trace. If A is an $n \times n$ matrix, the *trace* of A (denoted $\operatorname{tr} A$) is defined as the sum of all the diagonal elements of A, *i.e.*,

$$\operatorname{tr} A = \sum_{i=1}^{n} [A]_{ii}.$$

The trace has three fundamental properties:

1. $\operatorname{tr} I = n$, where I is the $n \times n$ identity matrix.
2. $\operatorname{tr}(A + aB) = \operatorname{tr} A + a \operatorname{tr} B$, for $n \times n$ matrices A and B and scalar a. (The trace is linear.)
3. $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for any $n \times n$ matrices A and B.

In fact, tr is the *only* function from $n \times n$ matrices to scalars that satisfies (1)–(3) above.

Exercise 2.7 (Challenging) Prove this last statement.

Exercise 2.8 Show that for any integers $m, n \geqslant 1$, if A is an $n \times m$ matrix and B is an $m \times n$ matrix, then

$$\operatorname{tr}(AB) = \operatorname{tr}(BA)\ . \tag{4}$$

This verifies item (3) above about the trace. We will use this fact frequently.

Hilbert Spaces. A vector space $\mathcal{H}$ over $\mathbb{C}$ is a *Hilbert space* if it has a scalar product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ that behaves as follows for all $u, v, w \in \mathcal{H}$ and $a \in \mathbb{C}$:

1. $\langle u, v + aw \rangle = \langle u, v \rangle + a\langle u, w \rangle$ ($\langle \cdot, \cdot \rangle$ is linear in the second argument).
2. $\langle u, v \rangle = \langle v, u \rangle^*$ ($\langle \cdot, \cdot \rangle$ is conjugate symmetric).
3. $\langle u, u \rangle \geqslant 0$, and if $u \neq 0$ then $\langle u, u \rangle > 0$.

Note that (2) implies that $\langle u, u \rangle \in \mathbb{R}$, so (3) merely asserts that it can't be negative. Also note that (1) and (2) imply that $\langle v + aw, u \rangle = \langle v, u \rangle + a^* \langle w, u \rangle$, *i.e.*, $\langle \cdot, \cdot \rangle$ is *conjugate linear* in the first argument. Such a scalar product is called a *Hermitean form* or a *Hermitean inner product*.

The *norm* of a vector $u \in \mathcal{H}$ is defined as $\|u\| = \sqrt{\langle u, u \rangle}$. Note that by (3), $\|0\| = 0$ and $\|u\| > 0$ if $u \neq 0$.

Exercise 2.9 Show that for any $u \in \mathcal{H}$ and any $a \in \mathbb{C}$, $\|au\| = |a|\|u\|$.

Example. We consider the vector space $\mathbb{C}^n$ of all n-tuples of complex numbers (for some $n > 0$), where vector addition and scalar multiplication are componentwise, *i.e.*,

$$\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} u_1 + v_1 \\ \vdots \\ u_n + v_n \end{bmatrix} \qquad \text{and} \qquad a \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} au_1 \\ \vdots \\ au_n \end{bmatrix}.$$

We define the Hermitean inner product for all vectors $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ as

$$\langle u, v \rangle = u_1^* v_1 + \cdots + u_n^* v_n = \sum_{i=1}^{n} u_i^* v_i.$$

In this example, u and v can be expressed as linear combinations over the "standard" basis $\{e_1, \ldots, e_n\}$, where

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{5}$$

where the 1 occurs in the ith row.

Exercise 2.10 Check that the three properties of a Hermitean form are satisfied in this example.

Note that if we restrict the u_i and v_i to be real numbers, then this is just the familiar dot product of two real vectors. Also note that in this example,

$$\|u\| = \sqrt{\langle u, u \rangle} = \sqrt{u_1^* u_1 + \cdots + u_n^* u_n} = \sqrt{|u_1|^2 + \cdots + |u_n|^2}.$$

Orthogonality and Normality. In a genuine sense, the example above is the *only* example that really matters. First some more definitions. Two vectors u, v in a Hilbert space $\mathcal{H}$ are *orthogonal* or *perpendicular* if $\langle u, v \rangle = 0$. A vector u is a *normal* or a *unit vector* if $\|u\| = 1$. A set of vectors $v_1, \ldots, v_k \in \mathcal{H}$ is an *orthogonal set* if different vectors are orthogonal. The set is an *orthonormal set* if, in addition, each vector is a unit vector. That is, for all $1 \leqslant i, j \leqslant k$, we have

$$\langle v_i, v_j \rangle = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

A basis for $\mathcal{H}$ is an *orthonormal basis* if it is an orthonormal set. Orthonormal bases are special and have nice properties that make them preferable to other bases. From now on we will assume that all our bases (for Hilbert spaces) are orthonormal unless I say otherwise, and I won't.

In the example above, $e_1, \ldots, e_n$ clearly form an orthonormal basis. We'll see later that every Hilbert space has an orthonormal basis—lots of them, in fact. But let's get back to our example. If we fix an orthonormal basis $\mathcal{B} = \{\beta_1, \ldots, \beta_n\}$ for a Hilbert space $\mathcal{H}$, then we can write two vectors $u, v \in \mathcal{H}$ in terms of $\mathcal{B}$ as

$$u = \sum_{i=1}^{n} u_i \beta_i \qquad \text{and} \qquad v = \sum_{j=1}^{n} v_j \beta_j,$$

for some unique scalars $u_1, \ldots, u_n, v_1, \ldots, v_n \in \mathbb{C}$. Let's see what $\langle u, v \rangle$ is.

$$\begin{aligned}
\langle u, v \rangle &= \langle u_1 \beta_1 + \cdots + u_n \beta_n, v \rangle \\
&= \sum_{i=1}^{n} u_i^* \langle \beta_i, v \rangle \text{ (conjugate linearity in the first argument)} \\
&= \sum_{i} u_i^* \langle \beta_i, v_1 \beta_1 + \cdots + v_n \beta_n \rangle \\
&= \sum_{i} u_i^* \sum_{j=1}^{n} v_j \langle \beta_i, \beta_j \rangle \text{ (linearity in the second argument)} \\
&= \sum_{i,j} u_i^* v_j \langle \beta_i, \beta_j \rangle \\
&= \sum_{i,j} u_i^* v_j \delta_{ij} \text{ (the basis is orthonormal)} \\
&= \sum_{i=1}^{n} u_i^* v_i.
\end{aligned}$$

In other words, $\langle u, v \rangle$ is exactly the quantity of our example above, if we identify u with the tuple $(u_1, \ldots, u_n) \in \mathbb{C}^n$ and v with the tuple $(v_1, \ldots, v_n) \in \mathbb{C}^n$.

Exercise 2.11 Show that any orthogonal set of nonzero vectors is linearly independent. [Hint: Let v be any linear combination of such vectors, and consider $\langle v, v \rangle$. You'll need the fact that $\langle \cdot, \cdot \rangle$ is positive definite.]

3 Week 2: Preliminaries

Linear Transformations and Matrices. Let U and V be vector spaces. A *linear map* is a function $T : U \to V$ such that, for all vectors $u, v \in U$ and scalar a,

$$T(u + av) = Tu + aTv.$$

The vector addition and scalar multiplication on the left-hand side is in U, and the right-hand side is in V. If $\{\alpha_1, \ldots, \alpha_n\}$ is a basis for U and $\{\beta_1, \ldots, \beta_m\}$ is a basis for V, then T can be expressed uniquely in matrix form with respect to these bases: For each $1 \leqslant j \leqslant n$, we write $T\alpha_j$ uniquely as a linear combination of the β_i:

$$T\alpha_j = \sum_{i=1}^{m} a_{ij}\beta_i, \tag{6}$$

where each a_{ij} is a scalar. Now let A be the $m \times n$ matrix whose (i, j)th entry is a_{ij}. Expressing any $u \in U$ with respect to the first basis (of U) as

$$u = \sum_{j=1}^{n} u_j \alpha_j = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$

we get

$$\begin{aligned}
Tu &= T\left(\sum_{j=1}^{n} u_j \alpha_j\right) \\
&= \sum_{j=1}^{n} u_j T\alpha_j \text{ (by linearity)} \\
&= \sum_{j} u_j \left(\sum_{i=1}^{m} a_{ij}\beta_i\right) \text{ (by (6))} \\
&= \sum_{i} \left(\sum_{j} a_{ij}u_j\right) \beta_i \\
&= \begin{bmatrix} \sum_j a_{1j}u_j \\ \vdots \\ \sum_j a_{mj}u_j \end{bmatrix} \\
&= \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \\
&= A \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},
\end{aligned}$$

expressed with respect to the second basis (of V). Thus applying T to a vector u amounts to multiplying the corresponding matrix on the left with the corresponding column vector on the right.

Conversely, given bases for U and for V, an $m \times n$ matrix defines a unique linear map T whose action on a vector u is given above.

Thus, **linear maps and matrices are interchangeable**, given bases for the requisite spaces.

Linear maps (with the same domain and codomain) can be added and multiplied by scalars thus:

$$\begin{aligned}(T_1 + T_2)u &= T_1u + T_2u,\\ (aT)u &= a(Tu).\end{aligned}$$

The two equations above define $T_1 + T_2$ and aT respectively (a a scalar) by showing how they map an arbitrary vector u. This makes the set of all such linear maps a vector space in its own right.

If U and V are Hilbert spaces and the $\{\alpha_j\}$ and $\{\beta_i\}$ are orthonormal bases, then each entry a_{ij} can be expressed as a scalar product in V:

$$\langle \beta_i, T\alpha_j \rangle = \langle \beta_i, a_{1j}\beta_1 + \cdots + a_{mj}\beta_m \rangle = \sum_{k=1}^{m} a_{kj}\langle \beta_i, \beta_k \rangle = a_{ij}.$$

One upshot of this is that a linear map T is completely determined by the quantities $\langle \beta_i, T\alpha_j \rangle$ for all i and j.

Adjoints. If A is any $m \times n$ matrix over $\mathbb{C}$, the *adjoint* of A (denoted A^* or $A^\dagger$) is the $n \times m$ matrix obtained by taking the transpose of A and then taking the complex conjugate of each entry. That is,

$$[A^*]_{ij} = ([A]_{ji})^*,$$

for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m$. The star on the left denotes the adjoint operator on matrices while the star on the right denotes complex conjugation in $\mathbb{C}$.

Note the following:

1. $(A^*)^* = A$.
2. $(A + aB)^* = A^* + a^*B^*$. (Here, A and B have the same dimensions, and $a \in \mathbb{C}$.)
3. $(AB)^* = B^*A^*$.

An important special case is u^* where $u = (u_1, \ldots, u_n)$ is a column vector (*i.e.*, an $n \times 1$ matrix). We have,

$$u^* = \begin{bmatrix} u_1^* & \cdots & u_n^* \end{bmatrix}.$$

That is, u^* is a *row vector* (*i.e.*, a $1 \times n$ matrix), called the *dual vector* of u. If $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ are vectors in some Hilbert space, expressed with respect to an orthonormal basis $\{\alpha_1, \ldots, \alpha_n\}$, then by our previous example we have

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i^* v_i = u^* v \,. \tag{7}$$

Here, we identify the 1×1 matrix $u^* v$ with the scalar comprising its sole entry.

If $\mathcal{H}$ and $\mathcal{J}$ are Hilbert spaces and $T : \mathcal{H} \to \mathcal{J}$ is linear, then there exists a unique linear map $T^* : \mathcal{J} \to \mathcal{H}$ such that for all $u \in \mathcal{H}$ and $v \in \mathcal{J}$,

$$\langle v, Tu \rangle = \langle T^* v, u \rangle.$$

Note that the left-hand side is the scalar product in $\mathcal{J}$, and the right-hand side is the scalar product in $\mathcal{H}$.

If we pick any orthonormal bases for $\mathcal{H}$ and $\mathcal{J}$, then these two definitions of the adjoint coincide. That is, if T is represented by the matrix A, then T^* is represented by the matrix A^*.

Exercise 3.1 (Challenging) Prove this fact.

Polarization Identities. Let $\mathcal{H}$ and cJ be Hilbert spaces and $A, B : \mathcal{H} \to \mathcal{J}$ linear maps. We have the following easily verifiable *polarization identity*: For every $x, y \in \mathcal{H}$,

$$\langle Ax, By \rangle = \frac{1}{4} \sum_{k=0}^{3} (-i)^k \langle A(x + i^k y), B(x + i^k y) \rangle \, . \tag{8}$$

Exercise 3.2 Verify Equation (8).

Equation (8) has a number of interesting special cases. Here are two: If $A = B$, then we have

$$\langle Ax, Ay \rangle = \frac{1}{4} \sum_{k=0}^{3} (-i)^k \langle A(x + i^k y), A(x + i^k y) \rangle = \frac{1}{4} \sum_{k=0}^{3} (-i)^k \left\| A(x + i^k y) \right\|^2 , \tag{9}$$

and if $\mathcal{H} = \mathcal{J}$ and $A = B = I_{\mathcal{H}}$, then we have

$$\langle x, y \rangle = \frac{1}{4} \sum_{k=0}^{3} (-i)^k \langle x + i^k y, x + i^k y \rangle = \frac{1}{4} \sum_{k=0}^{3} (-i)^k \left\| x + i^k y \right\|^2 . \tag{10}$$

Equation (10) is significant because it shows that the inner product on $\mathcal{H}$ is completely determined by the norm itself. Equation (9) implies that if an operator A preserves norms, it must also preserve inner products.

Gram-Schmidt Orthonormalization. We prefer orthonormal bases for our Hilbert spaces. Here we show that they actually exist, and in abundance. Let $\mathcal{H}$ be an n-dimensional Hilbert space and let $\{b_1, \ldots, b_n\}$ be any basis (not necessarily orthonormal) for $\mathcal{H}$. For $i = 1$ to n in order, define

$$\begin{aligned} x_i &= b_i - \sum_{k=1}^{i-1} \langle y_k, b_i \rangle y_k \\ y_i &= \frac{x_i}{\|x_i\|}. \end{aligned}$$

This is known as the *Gram-Schmidt procedure*. We'll see that $\{y_1, \ldots, y_n\}$ is an orthonormal basis. It's not obvious that the y_i are even well-defined, since we need to establish that $\|x_i\|$ in the denominator is nonzero. We can prove the following facts simultaneously by induction on i for $1 \leqslant i \leqslant n$, that is, assuming that all the facts are true for all $j < i$, we prove all the facts for i:

1. $x_i \neq 0$ (and thus $\|x_i\| > 0$).
2. $\|y_i\| = 1$.
3. $\{b_1, \ldots, b_i\}$, $\{x_1, \ldots, x_i\}$, and $\{y_1, \ldots, y_i\}$ are each linearly independent sets of vectors which span the same subspace of $\mathcal{H}$.
4. $\langle y_i, b_i \rangle = \langle b_i, y_i \rangle > 0$.
5. $\langle y_j, y_i \rangle = \langle y_i, y_j \rangle = 0$ for all $j < i$.

For the last item, we compute

$$\langle y_j, y_i \rangle = \frac{\langle y_j, x_i \rangle}{\|x_i\|} = \frac{1}{\|x_i\|} \left(\langle y_j, b_i \rangle - \sum_{k<i} \langle y_k, b_i \rangle \langle y_j, y_k \rangle \right) = \frac{\langle y_j, b_i \rangle - \langle y_j, b_i \rangle}{\|x_i\|} = 0.$$

The second to last equation comes from the fact that $\langle y_j, y_k \rangle = \delta_{jk}$ for all $j, k < i$, which is part of the inductive hypothesis.

It turns out (we won't prove this) that given a basis $b_1, \ldots, b_n$ there can only be one unique list $y_1, \ldots, y_n$ satisfying all the items (2)–(5) above.

Exercise 3.3 (Challenging) Prove this fact.

Exercise 3.4 Prove that applying the Gram-Schmidt procedure to a basis that is already orthonormal just results in the same basis.

Hermitean and Unitary Operators.

Definition 3.5 If $\mathcal{H}$ and $\mathcal{J}$ are Hilbert spaces, we let $\mathcal{L}(\mathcal{H}, \mathcal{J})$ denote the space of all linear maps from $\mathcal{H}$ to $\mathcal{J}$. We abbreviate $\mathcal{L}(\mathcal{H}, \mathcal{H})$ by $\mathcal{L}(\mathcal{H})$, the space of all linear operators on $\mathcal{H}$, with identity element I. Note that $\mathcal{L}(\mathcal{H}, \mathcal{J})$ is a vector space over $\mathbb{C}$.

A map $A \in \mathcal{J}(\mathcal{H})$ is *Hermitean* (or *self-adjoint*) if $A^* = A$. A map A is *unitary* if $AA^* = I$ (equivalently, $A^*A = I$).

For any $u, v \in \mathcal{H}$, we have the following easy facts:

- If A is Hermitean, then $\langle u, Av \rangle = \langle Au, v \rangle$. This follows immediately from the fact that $\langle u, Av \rangle = \langle A^*u, v \rangle$.
- If A and B are Hermitean then so is $A + B$.

- If A is Hermitean and a is real, then aA is Hermitean.
- If A is Hermitean, then so is A^*.
- If A is unitary, then $\langle Au, Av\rangle = \langle u, v\rangle$, that is, A preserves the scalar product. To see this, we just compute
$$\langle Au, Av\rangle = \langle A^*Au, v\rangle = \langle Iu, v\rangle = \langle u, v\rangle.$$
- If A and B are unitary, then so is AB.
- If A is unitary, then so is A^*. Note that $A^* = A^{-1}$ in this case.
- I is both Hermitean and unitary.

More on all this later.

$\mathcal{L}(\mathcal{H})$ is a Hilbert space. In Definition 3.5, we mentioned that $\mathcal{L}(\mathcal{H}, \mathcal{J})$ is a vector space over $\mathbb{C}$. In fact, its dimension is the product of the dimensions of $\mathcal{H}$ and of $\mathcal{J}$: Suppose $\mathcal{H}$ has dimension n and $\mathcal{J}$ has dimension m. Given orthonormal bases for each space, an element of $\mathcal{L}(\mathcal{H}, \mathcal{J})$ corresponds an $m \times n$ matrix. You can think of this matrix as a vector with mn components which just happen to be arranged in a 2-dimensional array rather than a single column. The vector addition and scalar multiplication operations on these matrices are componentwise, just as with vectors, so $\mathcal{L}(\mathcal{H}, \mathcal{J})$ has dimension mn.

There is a natural inner product that one can define on $\mathcal{L}(\mathcal{H}, \mathcal{J})$ that makes it into an mn-dimensional Hilbert space. For all $A, B \in \mathcal{L}(\mathcal{H}, \mathcal{J})$, define

$$\langle A, B\rangle := \operatorname{tr}(A^*B)\ . \tag{11}$$

This is known as the *Hilbert-Schmidt inner product* on $\mathcal{L}(\mathcal{H}, \mathcal{J})$. It looks similar to the expression u^*v for the inner product of vectors u and v (Equation (7)), except that A^*B is not a scalar but an operator in $\mathcal{L}(\mathcal{H})$, and so we take the trace to get a scalar result.

Exercise 3.6 Show that $\mathcal{L}(\mathcal{H}, \mathcal{J})$, together with its Hilbert-Schmidt inner product, satisfies all the axioms of a Hilbert space. [Hint: You can certainly just verify the axioms directly. Alternatively, represent operators as matrices with respect to some fixed orthonormal bases of $\mathcal{H}$ and $\mathcal{J}$, respectively, then show that if A and B are $m \times n$ matrices, then $\operatorname{tr}(A^*B)$ is the usual inner product of A and B on $\mathbb{C}^{mn}$, where we identify each matrix with the mn-dimensional vector of all its entries.]

The Hilbert-Schmidt inner product interacts nicely with composition of linear maps (or equivalently, matrix multiplication).

Exercise 3.7 Let $\mathcal{H}$, $\mathcal{J}$, and $\mathcal{K}$ be Hilbert spaces. Verify directly that for any $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, $B \in \mathcal{L}(\mathcal{J}, \mathcal{K})$, and $C \in \mathcal{L}(\mathcal{H}, \mathcal{J})$,

$$\langle A, BC\rangle = \langle B^*A, C\rangle = \langle AC^*, B\rangle\ . \tag{12}$$

This means that you can move a right or left factor from one side of the inner product to the other, provided you take its adjoint. [Hint: Use Exercise 2.8 along with basic properties of adjoints. You may assume that A, B, and C are all matrices of appropriate dimensions.]

4 Week 2: Preliminaries

Exercise 4.1 Let $b_1 = (-3, 0, 4)$, $b_2 = (3, -1, 2)$, and $b_3 = (0, 1, -1)$. Perform the Gram-Schmidt procedure above on $\{b_1, b_2, b_3\}$ to find the corresponding $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$.

Dirac Notation. In what follows, we fix an n-dimensional Hilbert space $\mathcal{H}$ and some orthonormal basis for it, so we can identify vectors with column vectors in the usual way. Recall that for column vectors $u, v \in \mathcal{H}$, we have

$$\langle u, v\rangle = u^* v\,.$$

Paul Dirac suggested a notation which somewhat reconciles the two sides of this equation: if we let $|\psi\rangle$ denote the column vector v and we let $\langle\varphi|$ denote the row vector u^*, then $\langle u, v\rangle = u^* v = \langle\varphi|\psi\rangle$ is just the usual multiplication of a row vector and a column vector (the two vertical bars overlap). Note how the product $\langle\varphi|\psi\rangle$ looks like $\langle u, v\rangle$ with the comma replaced by a vertigule. This notation has become standard in quantum mechanics. We denote a (column) vector by $|\psi\rangle$, where ψ is some label identifying it, and we denote its corresponding dual (row) vector by $\langle\psi|$ (thus $\langle\psi| = |\psi\rangle^*$ and vice versa: $|\psi\rangle = \langle\psi|^*$). The choice of delimiters tells us whether we are talking about a column vector or a row vector. A vector of the form $|\psi\rangle$ (*i.e.*, a column vector) is called a *ket vector*. If $|\varphi\rangle$ is another ket vector, its dual (a row vector) $\langle\varphi| = |\varphi\rangle^*$ is called a *bra vector*, so that the scalar $\langle\varphi|\psi\rangle$ can be called the *bracket* ("bra-ket") of $|\varphi\rangle$ and $|\psi\rangle$.

We'll start using Dirac notation because the book uses it, although there are some times when the notation just gets too clunky, and so then we will go back to using the "standard" notation.

We can combine kets and bras in other ways. For example, $|\psi\rangle\langle\varphi|$ is a column vector on the left multiplied by a row vector on the right (in standard notation, vu^*, where u and v are as above). This is then an $n \times n$ matrix, or considered another way, a linear operator $\mathcal{H} \to \mathcal{H}$ that takes a vector $|\chi\rangle$ and maps it to the vector $|\psi\rangle\langle\varphi|\chi\rangle = (\langle\varphi|\chi\rangle)|\psi\rangle$ (that is, the vector $|\psi\rangle$ multiplied by the scalar $\langle\varphi|\chi\rangle$). In any case, combining bras and kets just amounts to the usual vector or matrix multiplication.

As a special case, if $\{e_1, \ldots, e_n\}$ is the orthonormal basis for $\mathcal{H}$ that we have fixed, then, letting $|i\rangle := e_i$ for all $1 \leqslant i \leqslant n$, we have, for all $1 \leqslant i, j \leqslant n$,

$$|i\rangle\langle j| = e_i e_j^* = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

Where the 1 is in the ith row and jth column. This matrix is usually denoted E_{ij}. Notice that if A is a linear map $\mathcal{H} \to \mathcal{H}$ whose corresponding matrix has entries a_{ij}, then by the equation above we must have,

$$A = \sum_{i,j} a_{ij} E_{ij} = \sum_{i,j} a_{ij} |i\rangle\langle j|,$$

where both indices in the summation run from 1 to n. In particular, the identity operator is given by

$$I = \sum_i |i\rangle\langle i|.$$

Change of (Orthonormal) Basis. Let $\mathcal{H}$ be as before, and let $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ be two orthonormal bases for $\mathcal{H}$. There is a unique linear map $U \in \mathcal{L}(\mathcal{H})$ mapping the first basis to the second, *i.e.*, $Ue_i = f_i$ for all $1 \leqslant i \leqslant n$. Now for each $1 \leqslant i, j \leqslant n$ we have

$$\langle e_i, U^*Ue_j\rangle = \langle Ue_i, Ue_j\rangle = \langle f_i, f_j\rangle = \delta_{ij} = \langle e_i, e_j\rangle = \langle e_i, Ie_j\rangle.$$

Since the linear map U^*U is uniquely determined by the quantities above, we must therefore have $U^*U = I$, and thus U is unitary.

Conversely, if U is unitary and $\{e_1, \ldots, e_n\}$ is an orthonormal basis, then $\{Ue_1, \ldots, Ue_n\}$ is also an orthonormal basis, because U preserves the scalar product.

We conclude that the operators needed to change orthonormal bases in a Hilbert space are exactly the unitary operators.

Unitary Conjugation. If A and B are two linear operators in $\mathcal{L}(\mathcal{H})$ (equivalently, two $n \times n$-matrices), then we say that A is *unitarily conjugate* to B if there exists a unitary U such that $B = UAU^*$. The relation "is unitarily conjugate to" is an equivalence relation on $\mathcal{L}(\mathcal{H})$, that is, it is reflexive, symmetric, and transitive.

Exercise 4.2 Prove this. I.e., prove that if $A, B, C \in \mathcal{L}(\mathcal{H})$, then:

- A is unitarily conjugate to itself.
- If A is unitarily conjugate to B, then B is unitarily conjugate to A.
- If A is unitarily conjugate to B and B is unitarily conjugate to C, then A is unitarily conjugate to C.

Unitary conjugation allows us to change orthonormal bases. Suppose $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ are two orthonormal bases for $\mathcal{H}$ and let U be the unique unitary operator such that $Ue_i = f_i$ for all $1 \leqslant i \leqslant n$. Suppose that A is some linear operator on $\mathcal{H}$. We want to compare the matrix entries of A with respect to the two different bases. With respect to the first basis (the e-basis), the (i, j)th entry of the matrix A is given by $\langle e_i, Ae_j\rangle = e_i^*Ae_j$ (or $\langle i|A|j\rangle$ using Dirac notation). With respect to the second basis (the f-basis), the same entry is $\langle f_i, Af_j\rangle$. Starting with this, we get

$$\langle f_i, Af_j\rangle = \langle Ue_i, AUe_j\rangle = \langle e_i, U^*AUe_j\rangle.$$

The right-hand side is the (i, j)th entry of the matrix representing the operator U^*AU with respect to the e-basis.

To summarize, if M_A and M'_A are the matrices representing the operator A with respect to the e- and f-bases respectively, then

$$M'_A = M_U^* M_A M_U,$$

where M_U is the matrix representing the operator U with respect to the e-basis.

Thus, **changing orthonormal basis amounts to unitary conjugation of the corresponding matrices**.

Back to Quantum Physics: The Double Slit Experiment. It's been known since early in the 20th century that light comes in discrete packets (particles) called photons. People have observed individual photons hitting a photoelectric detector (or a photographic plate) at specific times and pinpoint locations, causing local electric currents in the detector (or dots to appear on the plate).

On the other hand, light also exhibits wavelike properties. In the double slit experiment, light from a laser beam is shined on an opaque barrier with two small openings close to each other (on the order of the wavelength of the light). A screen is placed on the other side of the barier. What you see on the screen are alternating bands of light and dark—a standard interference pattern caused by the light waves from the two slits interfering constructively and destructively with each other. This is easily visible to the naked eye. If you block one of the slits, then the interference pattern goes away and you just see a smoothly contoured, glowing blob on the screen (that depends on the width of the slit).

Here is a plausible (though ultimately wrong) explanation in terms of photons: the photons somehow are changing phase in time, and the photons that go through the top slit are interfering with the photons going through the bottom slit.

Let's see why this is wrong. Now alter the experiment as follows: Make the light source *extremely* dim, so that it emits on average only one photon per second, and replace the screen with a photographic plate (or photodetector) that will register where each photon hits. The photons appear to hit the plate at random places, but if you run the experiment a long time (thousands or millions of photons), you see that, statistically, the distribution of photon hits resembles the same wavy interference pattern as before. That is, the probability of a photon hitting any given location is proportional to the intensity of the light at that location in the original experiment.

We can't say the photons are interfering with each other, since one photon goes through long before the next one comes. The only explanation is that each photon is somehow passing through *both* slits at the same time and interfering with *itself* on the other side. This cannot be explained at all by classical physics, which asserts that the photon, being a particle, must travel through either the upper slit or the lower slit, but not both. Indeed, if you put detectors at both slits, the photon will only be detected at (at most) one slit or the other, not both.

Another thing that classical physics cannot explain is the random behavior of the photons at the plate. You can send two identical photons, of exactly the same frequency and moving in exactly the same direction, and they will wind up at different locations at the plate. So the behavior of the photons is not deterministic but *inherently random*.

Quantum mechanics is needed to explain both these phenomena as follows: Each photon does indeed correspond to a wave that goes through both slits, but the amplitude of this wave at any location is related to the *probability* of the photon being at that location. These waves interfere with each other and cause the interference pattern in the statistical distribution of photons at the plate.

So the two hallmarks of quantum mechanics are: (i) nondeterminism (inherent randomness) and (ii) interference of probabilities. More later.

5 Week 3: Unitary conjugation

Invariance under Unitary Conjugation: Trace and Determinant. If A and U are $n \times n$ matrices and U is unitary, then by Equation (4) of Exercise 2.8,

$$\operatorname{tr}(UAU^*) = \operatorname{tr}(U^*UA) = \operatorname{tr}(IA) = \operatorname{tr} A.$$

In other words, the tr function is invariant under unitary conjugation, *i.e.*, if matrices A and B are unitarily conjugate, then their traces are equal. This means that the tr function is really a function of the underlying operator and does not depend on which orthonormal basis you use to represent the operator as a matrix. (In fact, it does not depend on any basis, orthonormal or otherwise.)

It's worth looking at what the trace looks like in Dirac notation. If A is an operator and $\{e_1, \ldots, e_n\}$ is an orthonormal basis, then we know (letting $|i\rangle = e_i$ for all i, as before) that $[A]_{ij} = \langle e_i, Ae_j \rangle = \langle i|A|j\rangle$ for the matrix of A with respect to this basis. So,

$$\operatorname{tr} A = \sum_{i=1}^{n} [A]_{ii} = \sum_i \langle e_i, Ae_i \rangle = \sum_i e_i^* A e_i = \sum_i \langle i|A|i\rangle \,, \tag{13}$$

and this quantity does not depend on the particular orthonormal basis we choose.

Similarly, the determinant function det is also invariant under unitary conjugation. This follows from the fact that $\det(AB) = \det A \det B$ and $\det(A^{-1}) = (\det A)^{-1}$ for any nonsingular A. For A and U as above, we have

$$\det(UAU^*) = \det(UAU^{-1}) = (\det U)(\det A)(\det U)^{-1} = \det A.$$

So like the trace, det is really a function of the operator and does not depend on the basis used to represent the operator as a matrix.

Here are some other invariants under unitary conjugation. In each case, U is an arbitrary unitary operator.

The adjoint. For any A, clearly $(UAU^*)^* = UA^*U^*$. (The adjoint of a conjugate is the conjugate of the adjoint.)

Being Hermitean. If A is Hermitean, then $(UAU^*)^* = UA^*U^* = UAU^*$, so UAU^* is also Hermitean.

Being unitary. If A is unitary, then $(UAU^*)(UAU^*)^* = UAU^*UA^*U^* = UAA^*U^* = UU^* = I$, so UAU^* is also unitary.

Orthogonal Subspaces, Projection Operators. Projection operators are important for understanding how measurements are made in quantum mechanics. They are also important because they can serve as building blocks for more general operators. We will spend some quality time with them.

Again, let $\mathcal{H}$ be an n-dimensional Hilbert space, and let $V, W \subseteq \mathcal{H}$ be subspaces of $\mathcal{H}$. V and W are *mutually orthogonal* if $\langle v, w\rangle = 0$ for every $v \in V$ and $w \in W$.

Exercise 5.1 Show that if V and W are mutually orthogonal, then no nonzero vector can be in $V \cap W$.

There is a natural one-to-one correspondence between the subspaces of $\mathcal{H}$ and certain linear operators on $\mathcal{H}$ known as *projection operators*.

Definition 5.2 An *(orthogonal) projection operator* or *projector* on $\mathcal{H}$ is a linear map $P \in \mathcal{L}(\mathcal{H})$ such that

1. $P = P^*$, *i.e.*, P is Hermitean, and
2. $P^2 = P$, *i.e.*, P is "idempotent."

There are two trivial projection operators on $\mathcal{H}$, namely, I (the identity) and 0 (the zero operator, which maps every vector to 0). There are many nontrivial projection operators as well.

Exercise 5.3 Prove that an operator $P \in \mathcal{L}(\mathcal{H})$ is a projector if and only if $P = P^*P$.

Exercise 5.4 Show that if P and Q are projection operators and $PQ = 0$, then $QP = 0$ as well, and $P + Q$ is a projection operator. [Hint: To show that $QP = 0$, take the adjoint of both sides of the equation $PQ = 0$.]

Given a projection operator P on $\mathcal{H}$, let V be the image of P, that is, $V = \operatorname{img} P := \{Pv : v \in \mathcal{H}\}$. Then it is easy to check that V is a subspace of $\mathcal{H}$, and we say that "P projects onto V." Notice that if $u \in V$ then there is a v such that $Pv = u$, and so

$$Pu = PPv = Pv = u.$$

That is, P fixes every vector in V, and so clearly we also have $V = \{u \in \mathcal{H} : Pu = u\}$.

Not only does P project onto V but it does so *orthogonally*. This means that P moves any vector v *perpendicularly* onto V, or more precisely, $\langle u, Pv - v\rangle = 0$ for any $u \in V$, where $Pv - v$ is the vector representing the net movement from v to Pv. To see that $\langle u, Pv - v\rangle = 0$, we write $u = Pw$ for some w and just calculate:

$$\langle u, Pv - v\rangle = \langle Pw, Pv - v\rangle = \langle Pw, Pv\rangle - \langle Pw, v\rangle = \langle P^*Pw, v\rangle - \langle Pw, v\rangle = \langle Pw, v\rangle - \langle Pw, v\rangle = 0.$$

Conversely, if V is any subspace of $\mathcal{H}$, then there is a unique projection operator P that projects orthogonally onto V as above. First I'll show uniqueness: If P and Q are projectors that both orthogonally project onto V, then for any $v, w \in \mathcal{H}$ we have

$$\langle Pw, Pv\rangle = \langle P^*Pw, v\rangle = \langle P^2w, v\rangle = \langle Pw, v\rangle,$$

and

$$\langle Pw, Qv\rangle = \langle Q^*Pw, v\rangle = \langle QPw, v\rangle = \langle Pw, v\rangle.$$

The last equation follows from the fact that Q fixes every vector in V, in particular, Q fixes Pw. Putting these two facts together, we have

$$\langle Pw, Pv - Qv\rangle = \langle Pw, Pv\rangle - \langle Pw, Qv\rangle = \langle Pw, v\rangle - \langle Pw, v\rangle = 0.$$

Since w was chosen arbitrarily, this means that $Pv - Qv$ is orthogonal to every vector of the form Pw, *i.e.*, every vector in V. But $Pv - Qv$ is itself in V because both Pv and Qv are in V. Thus $Pv - Qv$ is orthogonal to *itself*, and this means that

$$0 = \langle Pv - Qv, Pv - Qv\rangle = \|Pv - Qv\|^2,$$

and so $Pv - Qv = 0$, hence $Pv = Qv$. Since v was chosen arbitrarily, we must have $P = Q$.

Now for existence. Let V be given. Choose some basis $\{b_1, \ldots, b_k\}$ for V, which, by Gram-Schmidt, we can assume is orthonormal. Here, k is the dimension of V, and $0 \leqslant k \leqslant n$. Extend this basis for V to a basis $\mathcal{B} = \{b_1, \ldots, b_n\}$ for $\mathcal{H}$, which (again by Gram-Schmidt) we can assume is orthonormal. Now let $P \in \mathcal{L}(\mathcal{H})$ be the linear operator whose matrix (with respect to $\mathcal{B}$) is given by

- $[P]_{ii} = 1$ for $1 \leqslant i \leqslant k$,
- $[P]_{ii} = 0$ for $k + 1 \leqslant i \leqslant n$, and
- $[P]_{ij} = 0$ for $i \neq j$.

Thus P is given by a diagonal matrix where the first k diagonal entries are 1 and the rest are 0. Clearly, $P = P^*$ and $P^2 = P$, so P is a projector. Furthermore, P fixes each of the basis vectors $b_1, \ldots, b_k$ and so it fixes each vector in V. P annihilates all the other $b_{k+1}, \ldots, b_n$, and so $Pv \in V$ for all $v \in \mathcal{H}$. Thus P projects orthogonally onto V.

Exercise 5.5 Let V be a subspace of $\mathcal{H}$ and let P be its corresponding projection operator. Show that $\dim V = \operatorname{tr} P$. [Hint: Consider the matrix construction just above.]

Exercise 5.6 Suppose $|\psi\rangle$ is a unit vector in $\mathcal{H}$ (*i.e.*, $\langle\psi|\psi\rangle = 1$). Show that $|\psi\rangle\langle\psi|$ is a projection operator. What subspace does it project onto? What is $\operatorname{tr}|\psi\rangle\langle\psi|$?

Exercise 5.7 Find the 3×3 matrix for the projector P that projects orthogonally onto the two-dimensional subspace of $\mathbb{C}^3$ spanned by $v_1 = (1, -1, 0)$ and $v_2 = (2, 0, i)$. P is the unique operator satisfying: (i) $P^2 = P = P^*$, (ii) $Pv_1 = v_1$, (iii) $Pv_2 = v_2$, and (iv) $\operatorname{tr} P = 2$. [Hint: If y_1 and y_2 are orthogonal unit vectors, then $y_1 y_1^* + y_2 y_2^*$ projects onto the subspace spanned by y_1 and y_2. Use Gram-Schmidt to find y_1 and y_2 given v_1 and v_2. When you find P, check items (i)–(iv) above.]

Exercise 5.8 Let V and W be subspaces of $\mathcal{H}$ with corresponding projection operators P and Q, respectively. Prove that V and W are mutually orthogonal if and only if $PQ = 0$. [Hint: For the forward direction, consider $\|PQv\|^2$ for any vector $v \in \mathcal{H}$. For the reverse direction, consider $\langle Pv, Qw\rangle$ for any vectors $v, w \in \mathcal{H}$, and move the P to the right-hand side of the bracket.]

If V is a subspace of $\mathcal{H}$, we define the *orthogonal complement* of V (denoted $V^\perp$) to be

$$V^\perp = \{u \in \mathcal{H} : (\forall v \in V)[\langle u, v\rangle = 0]\}.$$

$V^\perp$ is clearly a subspace of $\mathcal{H}$.

Exercise 5.9 Show that if V is a subspace of $\mathcal{H}$ with corresponding projection operator P, then $I - P$ is the projection operator corresponding to $V^\perp$.

Exercise 5.10 Show that if V is a subspace of $\mathcal{H}$, then $\mathcal{H}$ is the *direct sum* of V and $V^\perp$ (written $\mathcal{H} = V \oplus V^\perp$), that is, every vector in $\mathcal{H}$ is the *unique* sum of a vector in V with a vector in $V^\perp$. [Hint: use the previous exercise and the fact that $V \cap V^\perp = \{0\}$.]

Definition 5.11 A *complete set of orthogonal projectors*, also called a *decomposition of* I, is a collection $\{P_i : i \in \mathcal{I}\}$ of nonzero projectors on $\mathcal{H}$ such that

1. $P_i P_j = 0$ for all $i, j \in \mathcal{I}$ with $i \neq j$, and
2. $\sum_{i\in\mathcal{I}} P_i = I$ (the identity map).

Here, $\mathcal{I}$ is any finite set of distinct labels. We may have $\mathcal{I} = \{1, \ldots, k\}$ for some k, but there are other possibilities, including real numbers, or labels that are not numbers at all.

We will see later (Exercise 9.34) that condition (1) is actually redundant; it follows from condition (2).

Taking the trace of both sides of item (2), we get

$$\sum_{i\in\mathcal{I}} \operatorname{tr} P_i = \operatorname{tr} \sum_{i\in\mathcal{I}} P_i = \operatorname{tr} I = n.$$

Since each $P_i \neq 0$, its trace is a positive integer (Exercise 5.5), so there can be at most n many projection operators in any complete set, where $n = \dim \mathcal{H}$.

For each $i \in \mathcal{I}$, let V_i be the subspace that P_i projects onto. By Exercise 5.8, the V_i are all pairwise mutually orthogonal. Furthermore, the V_i together span all of $\mathcal{H}$: for any $v \in \mathcal{H}$,

$$v = Iv = \sum_{i\in\mathcal{I}} P_i v, \tag{14}$$

but $P_i v \in V_i$ for each i, so v is the sum of vectors from the V_i. Generalizing Exercise 5.10, one can show that $\mathcal{H} = \bigoplus_{i\in\mathcal{I}} V_i$ is the direct sum of the V_i. That means that every $v \in \mathcal{H}$ is the sum of *unique* vectors in the respective spaces V_i, and this sum is given by (14) above.

As a special case, if P projects onto a proper, nonzero subspace V of $\mathcal{H}$, then $\{P, I - P\}$ is a complete set of projectors corresponding to the two subspaces V and $V^\perp$.

Exercise 5.12 Let $\{P_i : i \in \mathcal{I}\}$ be a complete set of orthogonal projectors over $\mathcal{H}$, and let $v \in \mathcal{H}$ be any vector. Show by direct calculation that

$$\|v\|^2 = \sum_{a\in\mathcal{I}} \|P_a v\|^2.$$

Exercise 5.13 Suppose P and Q are projection operators on $\mathcal{H}$ projecting onto subspaces $V \subseteq \mathcal{H}$ and $W \subseteq \mathcal{H}$, respectively. Show that if P and Q commute, that is, if $PQ = QP$, then PQ is a projection operator projecting onto $V \cap W$.

Fundamentals of Quantum Mechanics. We now know enough math to present the fundamental principles of quantum mechanics. For now, I will abide by the Copenhagen interpretation of quantum mechanics first put forward by Niels Bohr. This is the best-known interpretation and is easy to work with, albeit somewhat unsatisfying philosophically. Another well-known interpretation is the Everett interpretation, a.k.a. the many-worlds interpretation or the unitary interpretation. More on that later. There are still other interpretations, but there are no conflicts between any of these interpretations; they all use the same math and lead to the same predictions. The differences are merely philosophical.

Physical Systems and States. A *physical system* is some part of nature, for example, the position of an electron orbiting an atom, the electric field surrounding the earth, the speed of a train, etc. The last two are "macroscopic," dealing with big objects with lots of mass, momentum, and energy. Although in principle quantum mechanics covers all these systems, it is most conveniently applied to microscopic systems like the first.

The most basic principle of quantum mechanics relevant to us is that to every physical system S there corresponds a Hilbert space $\mathcal{H} = \mathcal{H}_S$, called the *state space* of S.[1] At any given point in time, the system is in some *state*, which for now we can define as a unit vector $|\psi\rangle \in \mathcal{H}$. (We will revise this definition later on, but our revision does not invalidate anything we discuss until then.) The state of the system may change with time, depending on the forces (internal and external) applied to the system. We may write the state of the system at time t as $|\psi(t)\rangle$.

Time Evolution of an Isolated System. Let's assume that our system S is isolated, *i.e.*, it is not interacting with any other systems. The state of S evolves in time, but this evolution is *linear* in the following sense: For any two times $t_1, t_2 \in \mathbb{R}$, there is a linear operator $U = U(t_2, t_1) \in \mathcal{L}(\mathcal{H})$ such that if the system is in the state $|\psi(t_1)\rangle$ at time t_1 then at time t_2 the system will be in the state

$$|\psi(t_2)\rangle = U|\psi(t_1)\rangle.$$

The operator U only depends on the system (its internal forces) and on the times t_1 and t_2, but *not* on the particular state the system happens to be in. That is, the single operator U describes how the system evolves from *any* state at t_1 to the resulting state at t_2. Note that t_1 and t_2 are arbitrary; t_2 does not necessarily have to come after t_1.

Since U maps states to states, it must be norm-preserving. From this one can show that it must preserve the scalar product. That is, U must be unitary. Here are some other basic, intuitive facts:

1. $U(t, t) = I$ for any time t. (If no time elapses, then the state has no time to change.)

[1] In general, $\mathcal{H}$ may be infinite dimensional. The systems we care about, however, are all *bounded*, which means they correspond to finite dimensional spaces.

2. $U(t_1, t_2) = U(t_2, t_1)^{-1} = U(t_2, t_1)^*$ for all times t_1, t_2. (Tracing the evolution of the system backward in time should undo the changes made by running the system forward in time.)

3. $U(t_3, t_1) = U(t_3, t_2)U(t_2, t_1)$ for all times t_1, t_2, t_3. (Running the system from t_1 to t_2 and then from t_2 to t_3 has the same effect on the state as running the system from t_1 to t_3. Recall that operator composition reads from right to left.)

(Item (2) actually follows from items (1) and (3).) If the system S is known, then $U(t_2, t_1)$ can be computed with arbitrary accuracy, at least in principle. In many simple cases, $U(t_2, t_1)$ is known exactly, and can even be controlled precisely by manipulating the system S. Controlling U is crucial to quantum computation. We'll see specific examples a bit later.

Projective Measurement. Now and then, we'd like to get information about the state of our system S. It turns out that quantum mechanics puts severe limitations on how much information we can extract, and disallows us from extracting this information in a purely passive way.

The standard way of getting information about the state of a system is by making an *observation*, also called a *measurement*. These are terms of art which unfortunately don't bear much intuitive resemblance to their every-day meanings. A typical (and very general) type of measurement is a *projective measurement*.[2] If $\mathcal{H}$ is the Hilbert space of system S, then a projective measurement on S corresponds to a complete set $\{P_k : k \in \mathcal{I}\}$ of orthogonal projectors on $\mathcal{H}$. The elements of $\mathcal{I}$ are the *possible outcomes* of the measurement. If the system is in state $|\psi\rangle$ when the measurement is performed, then the measurement will produce exactly one of the possible outcomes *randomly* such that each outcome $k \in \mathcal{I}$ is produced with probability

$$\Pr[k] = \||P_k|\psi\rangle\|^2 = (P_k|\psi\rangle)^* P_k|\psi\rangle = \langle\psi|P_k P_k|\psi\rangle = \langle\psi|P_k|\psi\rangle \,. \tag{15}$$

Furthermore, immediately after the measurement, the state of the system will be

$$|\psi_k\rangle = \frac{P_k|\psi\rangle}{\|P_k|\psi\rangle\|} = \frac{P_k|\psi\rangle}{\sqrt{\langle\psi|P_k|\psi\rangle}} = \frac{P_k|\psi\rangle}{\sqrt{\Pr[k]}} \,, \tag{16}$$

where k is the outcome of the measurement.

If each projector is 1-dimensional (i.e., $\operatorname{tr} P_k = 1$ for every k), then such a measurement is called a *von Neumann measurement*. Von Neumann measurements allow for the largest possible number of outcomes: $|\mathcal{I}| = \dim(\mathcal{H})$, the dimension of $\mathcal{H}$.

6 Week 3: Projective measurements (cont.)

A number of points need to be emphasized and clarified.

- The outcome of the projective measurement is intrinsically random. You can prepare the system S in the exact same state $|\psi\rangle$ twice, perform the exact same projective measurement

[2] There are other, more "general" types of measurement, but these can actually be implemented using projective measurements on larger systems, so these other measurements really aren't more general than projective measurements.

both times, and get different outcomes. The only things that we can predict from our experiments are the *statistics* of the outcomes. If we know the state $|\psi\rangle$ of the system when the measurement is performed, then in principle we can compute $\Pr[k]$ for each outcome k, and then if we run the same experiment many times (say a million times), then we can expect to see outcome k occur about a $\Pr[k]$ fraction of the time. This is indeed what happens.

- There can be at most a finite, discrete number of possible outcomes associated with any projective measurement—no more than $\dim(\mathcal{H})$ (at least for bounded systems).
- The probabilities defined by (15) are certainly nonnegative, but we need to check that they sum to 1. We have

$$\sum_{k\in\mathcal{I}} \Pr[k] = \sum_k \langle\psi|P_k|\psi\rangle = \langle\psi| \left(\sum_k P_k\right) |\psi\rangle = \langle\psi|I|\psi\rangle = \langle\psi|\psi\rangle = \||\psi\rangle\|^2 = 1,$$

  since $|\psi\rangle$ is a unit vector.
- Performing a projective measurement in general disturbs the system being measured. The measurement actually consists of an interaction between the system and the measuring apparatus, and one cannot be affected without affecting the other. This disturbance of the system being measured is not just a practical matter of us not building our instruments delicate enough; it is a fundamental and unavoidable physical reality, sometimes referred to "collapse of the wavefunction."
- Suppose that we perform the measurement above on S in state $|\psi\rangle$ and get outcome k, so that the state becomes $|\psi_k\rangle = P_k|\psi\rangle/\|P_k|\psi\rangle\|$ as in (16), then we immediately repeat the same measurement. The probability of getting any outcome $j \in \mathcal{I}$ from the second measurement is

$$\Pr[j] = \|P_j|\psi_k\rangle\|^2 = \left\| \frac{P_j P_k|\psi\rangle}{\|P_k|\psi\rangle\|} \right\|^2 = \delta_{kj}.$$

  That is, we see the outcome k again with certainty, and the state immediately after the second measurement is

$$\frac{P_k|\psi_k\rangle}{\|P_k|\psi_k\rangle\|} = \frac{|\psi_k\rangle}{\||\psi_k\rangle\|} = |\psi_k\rangle,$$

  unchanged from after the first measurement. So the first measurement changes the state to be consistent with whatever the outcome is, so that repetitions of the same measurement will always yield the same outcome (provided, of course, that the state does not evolve between measurements).
- If $|\psi\rangle$ is a state and $\theta \in \mathbb{R}$, then $e^{i\theta}|\psi\rangle$ is also a state. The unit norm scalar $e^{i\theta}$ is known as a "phase factor." Note that
  1. if U is unitary, then obviously $Ue^{i\theta}|\psi\rangle = e^{i\theta}U|\psi\rangle$, and
  2. for the projective measurement $\{P_k\}_{k\in\mathcal{I}}$ above, the probability of seeing k when the system is in state $e^{i\theta}|\psi\rangle$ is

$$\|P_k e^{i\theta}|\psi\rangle\|^2 = |e^{i\theta}|^2 \|P_k|\psi\rangle\|^2 = \|P_k|\psi\rangle\|^2,$$

     that is, the same for the state $|\psi\rangle$, and finally,

3. if outcome k occurs, then the state after the measurement is

$$\frac{P_k e^{i\theta}|\psi\rangle}{\|P_k e^{i\theta}|\psi\rangle\|} = e^{i\theta}\frac{P_k|\psi\rangle}{\|P_k|\psi\rangle\|} = e^{i\theta}|\psi_k\rangle.$$

This means that the phase factor just "goes along for the ride" and does not affect the statistics of any projective measurment (or any other type of measurement, either). The state $|\psi\rangle$ and $e^{i\theta}|\psi\rangle$ are *physically indistinguishable*, and so we can choose overall phase factors arbitrarily in defining a state, or we are free to ignore them as we wish. More on this later.

We'll now see how this all plays out for a two-dimensional system.

A Perfect Example: Electron Spin. Rotating objects possess *angular momentum*. The angular momentum of an object is a vector in $\mathbb{R}^3$ that depends on the distribution of mass in the object and how the object is rotating. For any given object, the length of its angular momentum vector is proportional to the speed of the rotation (in revolutions per minute, say), and the vector's direction is pointing (roughly) along the axis of rotation in the direction given by the "right hand rule": a disk rotating counterclockwise in the x, y-plane has its angular momentum vector pointing in the positive z-direction. A Frisbee thrown by a right-handed person (using the usual backhand flip) rotates clockwise when viewed from above, so its angular momentum vector points down toward the ground.

If a rotating object carries a net electric charge, then it has a *magnetic moment* vector that is proportional to the angular momentum times the net charge. Shooting an object with a magnetic moment through a nonuniform electric field imparts a force to the object, causing it to deflect and change direction. The deflection force is along the axis given by the gradient of the electric field and is proportional to the component of the magnetic moment along that gradent axis. You can measure the component of the magnetic moment along the gradient axis this way by seeing the amount of deflection.

Electrons deflect when shot through a nonuniform magnetic field as well, so they possess magnetic moment. This can only mean that they have angular momentum as well, even though, being elementary particles, they have no constituent parts that can rotate around one another. This is just one of the many bizarre aspects of the microscopic world.

In the *Stern-Gerlach experiment*, randomly oriented electrons are shot through a nonuniform electric field whose gradient is oriented in the $+z$-direction (vertically). According to classical physics, we would expect the electrons to deflect by random amounts, causing a smooth up-down spread in the beam. Instead, what we actually observe is the beam split into two sharp beams of roughly equal intensity: one going up, the other going down (see Figure 1). So each electron only goes up the same amount or down the same amount. This experiment amounts to a projective measurement of the spin of an electron, at least in the z-direction. There are two possible outcomes: spin-up and spin-down. It is natural then to model the physical system of electron spin as a two-dimensional Hilbert space, with an orthonormal basis $\{|\uparrow\rangle, |\downarrow\rangle\}$, where $|\uparrow\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is the spin-up state and $|\downarrow\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is the spin-down state. We may also write $|\uparrow\rangle$ and $|\downarrow\rangle$ as $|+z\rangle$ and $|-z\rangle$,

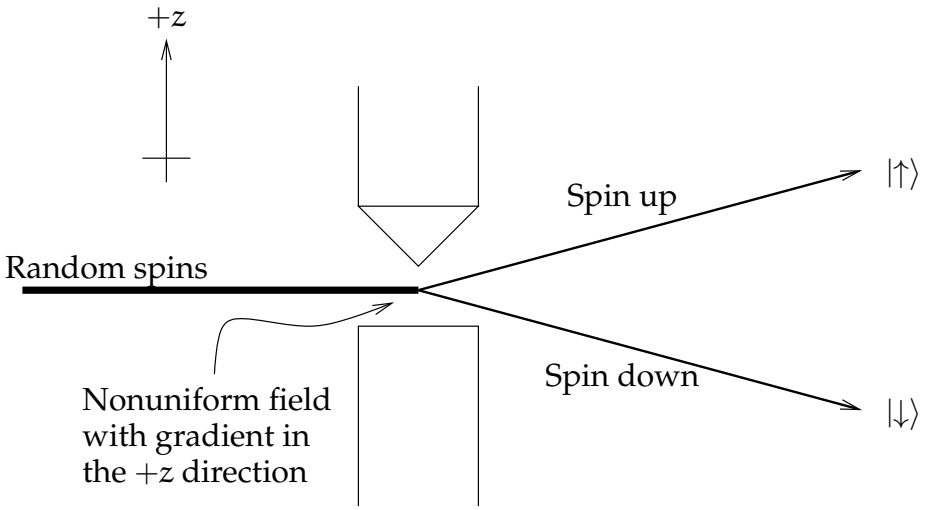


Figure 1: Stern-Gerlach exeriment: The electron beam comes in from the left, passes through a nonuniform field between the two probes, and splits into two beams. The field gradient is oriented along the axis of the probes, which is here given by the $+z$-direction.

respectively, to make clear along what axis the spin is aligned. The projectors in the projective measurement are then

$$\mathrm{P}_{\uparrow} = \mathrm{P}_{+z} = |\uparrow\rangle\langle\uparrow| = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

which projects onto the space spanned by $|\uparrow\rangle$ and corresponds to the spin-up outcome, and

$$\mathrm{P}_{\downarrow} = \mathrm{P}_{-z} = |\downarrow\rangle\langle\downarrow| = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

which projects onto the space spanned by $|\downarrow\rangle$ and corresponds to the spin-down outcome. As we'll see in a little bit, a two-dimensional Hilbert space actually suffices for modeling electron spin.

7 Week 4: Qubits

Qubits. In digital information processing, the basic unit of information is the *bit*, short for *binary digit*. Each bit has two distinct states that we care about: 0 and 1. In quantum information processing, we use bits as well, but we regard them as quantum systems that have two states $|0\rangle$ and $|1\rangle$ that form an orthonormal basis for a two-dimensional Hilbert space. Such systems are called *quantum bits*, or *qubits* for short. Any two-dimensional Hilbert space will do to model a qubit. This is why it is useful to consider the electron spin example. In fact, electron spin is one proposed way to implement a qubit: $|\uparrow\rangle$ is identified with $|0\rangle$ and $|\downarrow\rangle$ with $|1\rangle$.[3] We'll return to the electron spin example, but what we say applies generally to any system with a two-dimensional Hilbert space (sometimes called a "two-level system"), which can then in principle be used to implement a qubit. To emphasize this point, we'll use $|0\rangle$ and $|1\rangle$ to stand for $|\uparrow\rangle$ and $|\downarrow\rangle$, respectively, and we'll let the projectors P_0 and P_1 stand for $P_\uparrow$ and $P_\downarrow$, respectively.

Back to Electron Spin. Using the Stern-Gerlach apparatus oriented in a particular direction, we can prepare electrons to have spins in that direction. We simply retain one emerging beam and discard the other. Figure 2 shows electrons being prepared to spin in one direction in the x, z-plane, then measured in the $+z$-direction.

The general state of an electron spin is

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle.$$

That is, it is some linear combination of spin-up and spin-down. We would now like to determine which linear combinations correspond to which spin directions (in 3-space). Since $|\psi\rangle$ is a unit vector, we have

$$\begin{aligned} 1 &= \langle\psi|\psi\rangle \\ &= (\alpha^*\langle 0| + \beta^*\langle 1|)(\alpha|0\rangle + \beta|1\rangle) \\ &= \alpha^*\alpha\langle 0|0\rangle + \alpha^*\beta\langle 0|1\rangle + \beta^*\alpha\langle 1|0\rangle + \beta^*\beta\langle 1|1\rangle \\ &= |\alpha|^2 + |\beta|^2. \end{aligned}$$

Indeed, the probability of seeing $|0\rangle$ (spin-up) is

$$\langle\psi|P_0|\psi\rangle = \langle\psi|0\rangle\langle 0|\psi\rangle = \alpha^*\alpha = |\alpha|^2.$$

And similarly, the probability of seeing $|1\rangle$ (spin-down) is $|\beta|^2$. Since phase factors don't matter, we can assume from now on that $\alpha \in \mathbb{R}$ and $\alpha \geqslant 0$, because we can multiply $|\psi\rangle$ by the right phase factor, namely $e^{-i\arg(\alpha)}$.

Now consider the state $|\uparrow_\theta\rangle = \alpha|0\rangle + \beta|1\rangle$ prepared by the apparatus on the left of Figure 2, corresponding to a spin pointing at angle θ from the $+z$-axis in the $+x$ direction (Cartesian coordinates $(\sin\theta, 0, \cos\theta)$, which has unit length). Here $0 \leqslant \theta \leqslant \pi$. When it passes through the vertical apparatus on the right, the beam splits into two beams whose intensities are proportional

[3] Another system with a two-dimensional Hilbert space is photon polarization, where we can take as our basis the state $|\leftrightarrow\rangle$ (horizontal polarization) and the state $|\updownarrow\rangle$ (vertical polarization). All other polarization states (e.g., slanted or circular) are linear combinations of these two.

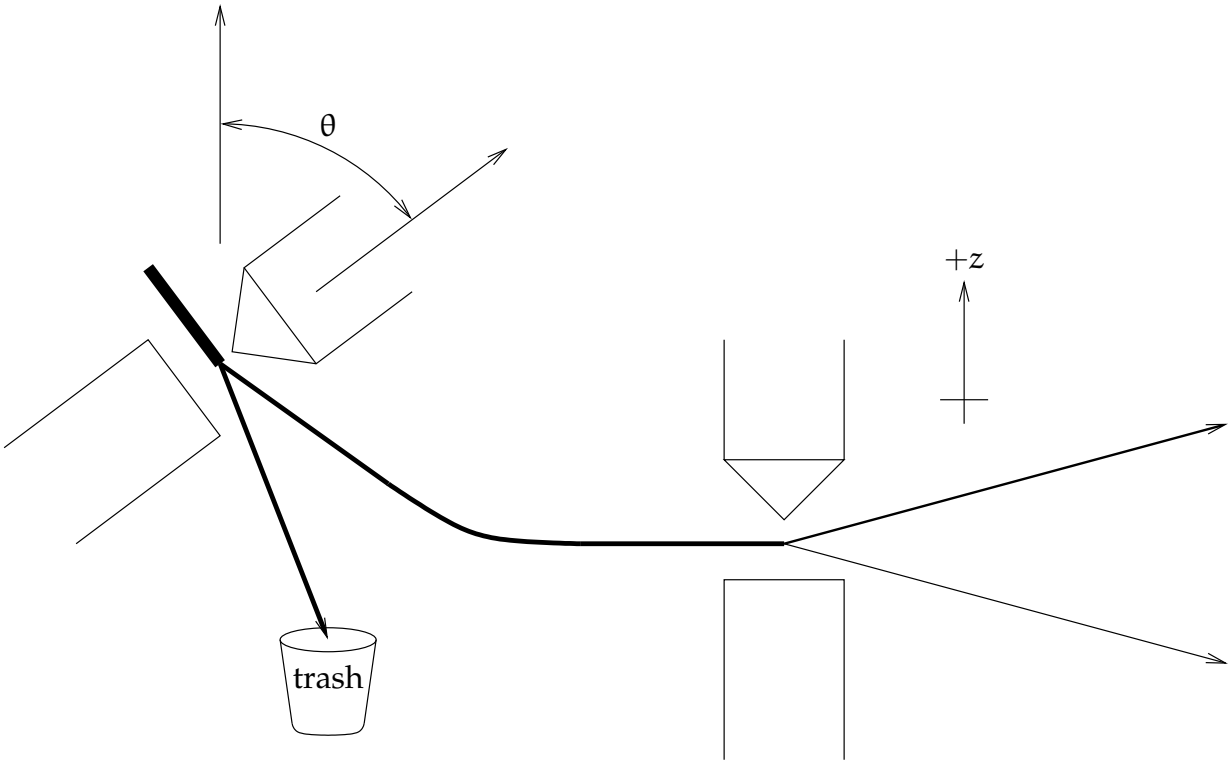


Figure 2: Electrons are prepared by the tilted apparatus on the left to spin at an angle θ from the $+z$-axis. These are then fed into a vertical apparatus.

to their probabilities. According to classical mechanics, the average deflection is proportional to the vertical component of the spin vector, *i.e.*, $\cos\theta$. If quantum mechanics is to agree with classical mechanics in the macroscopic limit, then the average deflection of the two beams must also be $\cos\theta$. The deflection of the spin-up beam is $+1$, and the deflection of the spin-down beam is -1, so the average deflection is

$$(+1)\Pr[\uparrow] + (-1)\Pr[\downarrow] = 2\Pr[\uparrow] - 1 = 2\langle\uparrow_\theta |P_{+z}|\uparrow_\theta\rangle - 1 = 2\alpha^2 - 1. \tag{17}$$

This must be $\cos\theta$, so solving for α in terms of θ and remembering that $\alpha \geqslant 0$, we get

$$\alpha = \sqrt{\frac{1+\cos\theta}{2}} = \cos\frac{\theta}{2}.$$

Since $0 \leqslant |\beta|^2 = 1 - \alpha^2$, we have $|\beta| = \sin(\theta/2)$. Thus,

$$\beta = e^{i\varphi}\sin\frac{\theta}{2},$$

for some real φ with $0 \leqslant \varphi < 2\pi$. In experiments, these relative intensities are actually observed.

It is worth mentioning at this point that for *any* $\alpha \geqslant 0$ and $\beta \in \mathbb{C}$ such that $\alpha^2 + |\beta|^2 = 1$, there are $0 \leqslant \theta \leqslant \pi$ and $0 \leqslant \varphi < 2\pi$ such that

$$\begin{aligned} \alpha &= \cos\frac{\theta}{2}, \\ \beta &= e^{i\varphi}\sin\frac{\theta}{2}, \end{aligned}$$

giving the general spin state as

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

Furthermore, θ and φ are uniquely determined by $|\psi\rangle$ except when $\alpha = 0$ or $\beta = 0$, in which case $\theta = \pi$ or $\theta = 0$, respectively, but φ is completely undetermined.

Now look at the case where $\theta = \pi/2$, that is, the spin is pointing in the $+x$ direction (to the right). We get

$$|{+x}\rangle = \left|\uparrow_{\pi/2}\right\rangle = \cos\frac{\pi}{4}|0\rangle + e^{i\varphi}\sin\frac{\pi}{4}|1\rangle = \frac{|0\rangle + e^{i\varphi}|1\rangle}{\sqrt{2}}.$$

We are free to adjust the phase factor of $|1\rangle$ to absorb the $e^{i\varphi}$ above. That is, without changing the physics, we redefine[4]

$$|1\rangle := e^{i\varphi}|1\rangle.$$

By the phase-adjustment we now get the "spin-right" state

$$|{+x}\rangle = \left|\uparrow_{\pi/2}\right\rangle = |{\rightarrow}\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}.$$

[4]Mathematicians may not like doing this, but physicists and computer scientists aren't bothered by it.

The corresponding one-dimensional projector is

$$P_{+x} = P_{\rightarrow} = |+x\rangle\langle+x| = \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)\left(\frac{\langle 0| + \langle 1|}{\sqrt{2}}\right) = \frac{1}{2}(|0\rangle\langle 0| + |0\rangle\langle 1| + |1\rangle\langle 0| + |1\rangle\langle 1|),$$

which has matrix form $(1/2)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

Now we consider the state $|+y\rangle$ representing spin in the $+y$ direction. A $+y$ spin has no $+z$-component, so if $|+y\rangle$ is measured along the z-axis, we get $\Pr[\uparrow] = \Pr[\downarrow] = 1/2$, as with $|+x\rangle$. Thus,

$$|+y\rangle = \frac{|0\rangle + e^{i\varphi}|1\rangle}{\sqrt{2}} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ e^{i\varphi} \end{bmatrix},$$

for some $0 \leqslant \varphi < 2\pi$. If we now measure a $+y$ spin in the $+x$ direction, we should again get equal probabilities of spin-left and spin-right, since the spin is perpendicular to x. Thus we should have

$$\frac{1}{2} = \Pr[\rightarrow] = \langle +y|P_{\rightarrow}|+y\rangle = \frac{1}{4}\begin{bmatrix} 1 & e^{-i\varphi} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ e^{i\varphi} \end{bmatrix} = \frac{1 + \cos\varphi}{2}.$$

So $\cos\varphi = 0$, and it follows that $\varphi \in \{\pi/2, 3\pi/2\}$ and so $e^{i\varphi} = \pm i$. It does not matter which value we choose; the math and physics is equivalent either way. So we'll arbitrarily set $\varphi := \pi/2$, whence we get

$$|+y\rangle = \frac{|0\rangle + i|1\rangle}{\sqrt{2}}.$$

The corresponding projector is

$$P_{+y} = |+y\rangle\langle+y| = \left(\frac{|0\rangle + i|1\rangle}{\sqrt{2}}\right)\left(\frac{\langle 0| - i\langle 1|}{\sqrt{2}}\right) = \frac{1}{2}(|0\rangle\langle 0| - i|0\rangle\langle 1| + i|1\rangle\langle 0| + |1\rangle\langle 1|),$$

which has matrix form $(1/2)\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$.

Let's review:

$$|+x\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{18}$$

$$|+y\rangle = \frac{|0\rangle + i|1\rangle}{\sqrt{2}} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}, \tag{19}$$

$$|+z\rangle = |0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{20}$$

The corresponding projectors are

$$\begin{aligned} P_{+x} = |+x\rangle\langle+x| &= \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{1}{2}(I + X), \\ P_{+y} = |+y\rangle\langle+y| &= \frac{1}{2}\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix} = \frac{1}{2}(I + Y), \\ P_{+z} = |+z\rangle\langle+z| &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{2}(I + Z), \end{aligned}$$

where

$$X = \sigma_x = \sigma_1 = 2P_{+x} - I = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{21}$$

$$Y = \sigma_y = \sigma_2 = 2P_{+y} - I = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \tag{22}$$

$$Z = \sigma_z = \sigma_3 = 2P_{+z} - I = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{23}$$

X, Y, and Z are known as the *Pauli spin matrices*. More on them later.

Now consider a general spin state, written in terms of θ and φ:

$$|\psi\rangle = |\uparrow_{\theta,\varphi}\rangle = \cos(\theta/2)|0\rangle + e^{i\varphi}\sin(\theta/2)|1\rangle = \begin{bmatrix} \cos(\theta/2) \\ e^{i\varphi}\sin(\theta/2) \end{bmatrix}.$$

(Recall that $0 \leqslant \theta \leqslant \pi$ and $0 \leqslant \varphi < 2\pi$ are arbitrary.) The direction of this spin is given by a vector $s = (x_s, y_s, z_s)$ in 3-space with Cartesian coordinates $x_s, y_s, z_s \in \mathbb{R}$. How do we find x_s, y_s, z_s? We know that these values are the average deflections observed when the spin is measured in the $+x$, $+y$, or $+z$ axes, respectively. So generalizing Equation (17), we must have

$$x_s = 2\langle\psi|P_{+x}|\psi\rangle - 1 = \langle\psi|X|\psi\rangle = \cos(\theta/2)\sin(\theta/2)(e^{i\varphi} + e^{-i\varphi}) = \sin\theta\cos\varphi, \tag{24}$$

$$y_s = 2\langle\psi|P_{+y}|\psi\rangle - 1 = \langle\psi|Y|\psi\rangle = \cos(\theta/2)\sin(\theta/2)(-i)(e^{i\varphi} - e^{-i\varphi}) = \sin\theta\sin\varphi, \tag{25}$$

$$z_s = 2\langle\psi|P_{+z}|\psi\rangle - 1 = \langle\psi|Z|\psi\rangle = \cos^2(\theta/2) - \sin^2(\theta/2) = \cos\theta. \tag{26}$$

Thus s is exactly the point on the unit sphere whose spherical coordinates are (θ, φ).[5]

Exercise 7.1 Verify Equations (24–26) using matrix multiplication and trig.

Exercise 7.2 What is the spin direction corresponding to the state $(\sqrt{3}|0\rangle - |1\rangle)/2$? Express your answer as simply as possible.

Exercise 7.3 What spin state corresponds to the direction $s = (-2/3, 2/3, 1/3)$? Express your answer as simply as possible.

Exercise 7.4 (Very useful!) Show that if $|\psi\rangle$ is a general spin state corresponding to the direction $s = (x_s, y_s, z_s)$ as described above, then

$$|\psi\rangle\langle\psi| = \frac{1}{2}(I + x_s X + y_s Y + z_s Z). \tag{27}$$

The right-hand side is sometimes written as $(1/2)(I + s \cdot \sigma)$, abusing the dot product notation.

[5]Each vector s on the unit sphere can be described using spherical coordinates, *i.e.*, two angles θ and φ, where $0 \leqslant \theta \leqslant \pi$ is the angle between s and the $+z$ axis (the "latitude" of s, measured down from the North Pole), and $0 \leqslant \varphi < 2\pi$ is the angle one would have to swivel the x, z-plane counterclockwise around the $+z$ axis until it hits s (the "longitude" of s, measured east of Greenwich, *i.e.*, east of the x, z-plane). If s has spherical coordinates (θ, φ), then its Cartesian coordinates are $(\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$.

8 Week 4: Density operators

Density Operators. One problem with using a vector $|\psi\rangle$ to represent a physical state is that the vector carries more information than is physically relevant, namely, an overall phase factor. The physically relevant portion of $|\psi\rangle$ is really just the one-dimensional subspace that it spans (which does not depend on any phase factors), or equivalently, the projector $|\psi\rangle\langle\psi|$ that orthogonally projects onto that subspace. For this and other reasons, one may *define* the state of a system to be a one-dimensional projection operator $\rho = |\psi\rangle\langle\psi|$ instead of a vector $|\psi\rangle$. This alternate view of states is known as the *density operator formalism*, and ρ is known as a *density operator* or *density matrix*. Besides the advantage of discarding the physically irrelevant phase information, this formalism has other advantages that we will see later when we discuss quantum information theory. For many of the tasks at hand, however, either formalism will suffice, and we will use both as is convenient.

We need to describe the two basic physical processes that we have discussed—time evolution and projective measurement—in terms of the density operator formalism.

Time evolution of an isolated system. In the original formalism, time evolution is described by a unitary operator U such that any state $|\psi\rangle$ evolves to a state $U|\psi\rangle$ in the given interval of time. In the new density operator formalism, the state $\rho = |\psi\rangle\langle\psi|$ would evolve under U to the new state

$$\rho' = U\rho U^*. \tag{28}$$

To see why this is so, we merely observe that the new state should be $|\varphi\rangle\langle\varphi|$, where $|\varphi\rangle = U|\psi\rangle$. We get

$$\rho' = |\varphi\rangle\langle\varphi| = |\varphi\rangle|\varphi\rangle^* = U|\psi\rangle(U|\psi\rangle)^* = U|\psi\rangle|\psi\rangle^*U^* = U|\psi\rangle\langle\psi|U^* = U\rho U^*.$$

Projective measurement. Suppose we are given a complete set $\{P_k : k \in \mathcal{I}\}$ of projectors corresponding to a projective measurement. In the original formalism, if the system is in state $|\psi\rangle$ before the measurement, then the probability of outcome k is $\langle\psi|P_k|\psi\rangle$. Since this probability is physically relevant (we can collect statistics over many identical experiments), we had better get the same probability in the new formalism: when the state of the system is $\rho = |\psi\rangle\langle\psi|$ before the measurement, the probability of outcome k is given by

$$\Pr[k] = \mathrm{tr}(P_k\rho) = \langle P_k, \rho\rangle\ , \tag{29}$$

where the right-hand side refers to the Hilbert-Schmidt inner product on $\mathcal{L}(\mathcal{H})$ (see Equation (11)). To see that this is the same as in the original formulation, we can use the form of the trace given by Equation (13), where we choose an orthonormal basis $\{e_1, \ldots, e_n\}$ such that $e_1 = |\psi\rangle$. Letting $|i\rangle = e_i$ for all i as before (and so $|\psi\rangle = |1\rangle$), we then get

$$\begin{aligned}\mathrm{tr}(P_k\rho) &= \sum_{i=1}^{n}\langle i|P_k\rho|i\rangle = \sum_i \langle i|P_k|\psi\rangle\langle\psi|i\rangle \\ &= \sum_i \langle i|P_k|1\rangle\langle 1|i\rangle = \langle 1|P_k|1\rangle = \langle\psi|P_k|\psi\rangle\ ,\end{aligned}$$

which is the same as originally defined. Alternatively, we can use the commuting property of the trace (Equation (4)) to get the same thing:

$$\mathrm{tr}(P_k\rho) = \mathrm{tr}(P_k|\psi\rangle\langle\psi|) = \mathrm{tr}(\langle\psi|P_k|\psi\rangle) = \langle\psi|P_k|\psi\rangle \ .$$

That last equation holds because $\langle\psi|P_k|\psi\rangle$ is just a scalar (a 1×1 matrix). Assuming the outcome is k, the state after the measurement should be $\rho_k = |\psi_k\rangle\langle\psi_k|$, where $|\psi_k\rangle = P_k|\psi\rangle/\||P_k|\psi\rangle\|$. This simplifies:

$$|\psi_k\rangle\langle\psi_k| = \frac{P_k|\psi\rangle\langle\psi|P_k^*}{\||P_k|\psi\rangle\|^2} = \frac{P_k|\psi\rangle\langle\psi|P_k}{\Pr[k]} = \frac{P_k\rho P_k}{\Pr[k]} = \frac{P_k\rho P_k}{\mathrm{tr}(P_k\rho)} = \frac{P_k\rho P_k}{\langle P_k, \rho\rangle} \ .$$

Thus, the post-measurement state after outcome k is given by

$$\rho_k = \frac{P_k\rho P_k}{\Pr[k]} = \frac{P_k\rho P_k}{\mathrm{tr}(P_k\rho)} = \frac{P_k\rho P_k}{\langle P_k, \rho\rangle} \ . \tag{30}$$

Note that $\mathrm{tr}(P_k\rho) = \mathrm{tr}(P_k^2\rho) = \mathrm{tr}(P_k\rho P_k)$, so the denominator in (30), *i.e.*, the probability of getting the outcome k, is the trace of the numerator. (Obviously, ρ_k is undefined if $\Pr[k] = 0$, but if that's the case, we'd never see outcome k.)

Exercise 8.1 Show that if $|\psi_1\rangle$ and $|\psi_2\rangle$ are unit vectors, and $\rho_1 = |\psi_1\rangle\langle\psi_1|$ and $\rho_2 = |\psi_2\rangle\langle\psi_2|$, then

$$\langle\rho_1, \rho_2\rangle = \mathrm{tr}(\rho_1\rho_2) = |\langle\psi_1|\psi_2\rangle|^2 \ .$$

Properties of the Pauli Operators. The operators X, Y, Z defined in (21–23) play a prominent role in quantum mechanics and quantum informatics. Here we'll present their most important properties in one place for ease of reference. All of these facts are easy to verify, and we leave that for the exercises.

1. $X^2 = Y^2 = Z^2 = I$.
2. (a) $XY = -YX = iZ$.
   (b) $YZ = -ZY = iX$.
   (c) $ZX = -XZ = iY$.
3. X, Y, and Z are all both Hermitean and unitary.
4. $\mathrm{tr}\, X = \mathrm{tr}\, Y = \mathrm{tr}\, Z = 0$.
5. $\det X = \det Y = \det Z = -1$.

Exercise 8.2 Verify all the above equations.

Note that there is a cyclic symmetry among the Pauli matrices. If we simultaneously substitute $X \mapsto Y$, $Y \mapsto Z$, and $Z \mapsto X$ everywhere in the equations above, we get the same equations. We won't pursue it here, but you can use the Pauli operators to represent the quaternions $\mathbb{H}$.

The four 2×2 matrices I, X, Y, Z (also denoted $\sigma_0, \sigma_1, \sigma_2, \sigma_3$, respectively) form a basis for the space $\mathcal{L}(\mathbb{C}^2)$ of all operators over $\mathbb{C}^2$ (i.e., 2×2 matrices over $\mathbb{C}$). That is, for any 2×2 matrix A, there are unique coefficients $a_0, a_1, a_2, a_3 \in \mathbb{C}$ such that

$$A = a_0 I + a_1 X + a_2 Y + a_3 Z = \sum_{i=0}^{3} a_i \sigma_i. \tag{31}$$

The coefficients can often be found by inspection, but there is a brute force method to find them:

Exercise 8.3

1. Verify that the set
$$\left\{ \frac{I}{\sqrt{2}}, \frac{X}{\sqrt{2}}, \frac{Y}{\sqrt{2}}, \frac{Z}{\sqrt{2}} \right\}$$
is an orthonormal basis for $\mathcal{L}(\mathbb{C}^2)$ (with the Hilbert-Schmidt norm, of course).

2. Show that if A is given by Equation (31), then
$$a_i = \frac{1}{2} \operatorname{tr}(\sigma_i A) = \frac{1}{2} \langle \sigma_i, A \rangle$$
for all $0 \leqslant i \leqslant 3$. [Hint: Use the previous item.] This implies that
$$\langle \sigma_i, \sigma_j \rangle = 2\delta_{ij} \tag{32}$$
for all $i, j \in \{0, 1, 2, 3\}$.

Exercise 8.4 Show that if $A = xX + yY + zZ$ for real numbers x, y, z such that $x^2 + y^2 + z^2 = 1$, then $A^2 = I$. Thus A is both Hermitean and unitary.

Single-Qubit Unitary Operators. In this topic, we show that applying any unitary operator to a one-qubit system amounts to a rigid rotation in $\mathbb{R}^3$, and conversely, any rigid rotation in $\mathbb{R}^3$ corresponds to a unitary operator. We've seen that a general one-qubit state can be written, up to an overall phase factor, as

$$|\psi\rangle = |\uparrow_{\theta,\varphi}\rangle = \cos(\theta/2)|0\rangle + e^{i\varphi} \sin(\theta/2)|1\rangle,$$

for some $0 \leqslant \theta \leqslant \pi$ and some $0 \leqslant \varphi < 2\pi$, and that this state corresponds uniquely (and vice versa) to the point s on the unit sphere in $\mathbb{R}^3$ with spherical coordinates (θ, φ) (and thus with Cartesian coordinates $(x_s, y_s, z_s) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$). (Think of s as the spin direction of an electron, for example.) The unit sphere in question here is known as the *Bloch sphere*. We'll now show that the action of a unitary operator U on one-qubit states amounts to a rigid rotation S_U of the Bloch sphere.

It's slightly more convenient to work in the density operator formalism, using Equation (27). Given any one-qubit unitary operator U, we define the map S_U from the Bloch sphere onto itself as follows: For any point $s = (x_s, y_s, z_s)$ on the Bloch sphere (s is a vector in $\mathbb{R}^3$ of length 1), let

$$\rho_s = \frac{1}{2}(I + x_s X + y_s Y + z_s Z)$$

be the corresponding one-qubit state, according to Equation (27). Then let

$$\rho_t = U\rho_s U^*$$

be the state obtained by evolving the system in state ρ_s by U. The state ρ_t can be written as

$$\rho_t = \frac{1}{2}(I + x_t X + y_t Y + z_t Z),$$

for some unique $t = (x_t, y_t, z_t)$ on the Bloch sphere. We now define $S_U(s)$ to be this t.[6]

It is immediate from the definition that for unitaries U and V we have $S_{UV} = S_U S_V$.

To show that S_U rotates the sphere rigidly, we first show that S_U preserves dot products of vectors on the Bloch sphere, that is, $S_U(r) \cdot S_U(s) = r \cdot s$ for any r and s on the Bloch sphere. This implies that S_U is a rigid map of the Bloch sphere onto itself, but it does not imply that S_U is a rotation, because S_U might be orientation-reversing, *e.g.*, a reflection. We'll see that S_U preserves orientation (aka chirality, aka "handedness"), so that it must be a rotation.[7]

Let $r = (r_1, r_2, r_3)$ and $s = (s_1, s_2, s_3)$ be any two points on the Bloch sphere, with corresponding states $\rho_r = (1/2)\sum_{i=0}^{3} r_i\sigma_i$ and $\rho_s = (1/2)\sum_{j=0}^{3} s_j\sigma_j$ as above, where we define $r_0 = s_0 = 1$. Recall that the dot product of r and s is $r \cdot s = r_1s_1 + r_2s_2 + r_3s_3$. Let's compute $\langle \rho_r, \rho_s \rangle$ using Equation (32):

$$\begin{aligned}
\langle \rho_r, \rho_s \rangle &= \left\langle \frac{1}{2}\sum_{i=0}^{3} r_i\sigma_i \,,\, \frac{1}{2}\sum_{j=0}^{3} s_j\sigma_j \right\rangle \\
&= \frac{1}{4}\sum_{i,j} r_i s_j \langle \sigma_i, \sigma_j \rangle = \frac{1}{4}\sum_{i,j} r_i s_j (2\delta_{ij}) \\
&= \frac{1}{2}\sum_{i=0}^{3} r_i s_i = \frac{1}{2}\left(1 + \sum_{i=1}^{3} r_i s_i\right) \\
&= \frac{1 + r \cdot s}{2} \,,
\end{aligned}$$

so

$$r \cdot s = 2\langle \rho_r, \rho_s \rangle - 1 \,. \tag{33}$$

Since r and s were arbitrary, we should also have

$$S_U(r) \cdot S_U(s) = 2\langle \rho_{S_U(r)} \,,\, \rho_{S_U(s)} \rangle - 1 \,,$$

but now,

$$\langle \rho_{S_U(r)} \,,\, \rho_{S_U(s)} \rangle = \mathrm{tr}((U\rho_r U^*)(U\rho_s U^*)) = \mathrm{tr}(U\rho_r\rho_s U^*) = \mathrm{tr}(\rho_r\rho_s) = \langle \rho_r, \rho_s \rangle \,,$$

and so $S_U(r) \cdot S_U(s) = r \cdot s$ as we wanted.

[6] There's no reason that we have to restrict the vector s to be on the Bloch sphere. We can define S_U in precisely the same way for any $s \in \mathbb{R}^3$, giving a map from all of $\mathbb{R}^3$ to $\mathbb{R}^3$. From the sequel, it will be evident that this is a linear map.

[7] A linear map A from $\mathbb{R}^n$ to $\mathbb{R}^n$ preserves orientation iff $\det A > 0$, and it reverses orientation iff $\det A < 0$.

Now is perhaps a good time to clear up some confusion that may arise about points on the Bloch sphere. Letting $|\psi_r\rangle$ and $|\psi_s\rangle$ be such that $\rho_r = |\psi_r\rangle\langle\psi_r|$ and $\rho_s = |\psi_s\rangle\langle\psi_s|$, then combining Equation (33) with Exercise 8.1 above, we get

$$r \cdot s = 2|\langle\psi_r|\psi_s\rangle|^2 - 1.$$

Thus $\langle\psi_r|\psi_s\rangle = 0$ iff $r \cdot s = -1$. In other words, qubit states that are *orthogonal* in the Hilbert space correspond to *antipodal*—or *opposite*—points on the Bloch sphere. We kind of knew this already, since the two possible outcomes of the Stern-Gerlach spin measurement (in any direction) are opposite spins (e.g., $|\uparrow\rangle$ and $|\downarrow\rangle$), and must (as with any projective measurement) correspond to orthogonal states.

Before showing that S_U must preserve orientation, we'll show that for any rigid rotation S of the Bloch sphere, there is a unitary U such that $S = S_U$. Geometrically, any rotation S of the unit sphere can be decomposed into a sequence of three simple rotations:

1. a counterclockwise rotation $S_z(\psi)$ about the $+z$ axis through an angle ψ where $0 \leqslant \psi < 2\pi$,
2. a counterclockwise rotation $S_y(\theta)$ about the $+y$ axis through an angle θ where $0 \leqslant \theta \leqslant \pi$, and
3. another counterclockwise rotation $S_z(\varphi)$, about the $+z$ axis, this time through an angle φ where $0 \leqslant \varphi < 2\pi$.

The last two rotations have the effect of moving the North Pole (*i.e.*, the point $(0,0,1)$) to an arbitrary point on the sphere (with spherical coordinates (θ,φ)) in a standard way. The only remaining freedom left in choosing S is an initial rotation that fixes the North Pole, *i.e.*, the first rotation above. The three angles φ,θ,ψ are uniquely determined by S (except when $\theta = 0$ or $\theta = \pi$), and are called the *Euler angles* of S.

So $S = S_z(\varphi)S_y(\theta)S_z(\psi)$, and so to implement S, we only need to find unitaries for rotations around the $+z$ and $+y$ axes. For any angle φ, define

$$R_z(\varphi) = \begin{bmatrix} e^{-i\varphi/2} & 0 \\ 0 & e^{i\varphi/2} \end{bmatrix}. \tag{34}$$

$R_z(\varphi)$ is obviously unitary, and

$$R_z(\varphi)(\alpha|0\rangle + \beta|1\rangle) = e^{-i\varphi/2}\alpha|0\rangle + e^{i\varphi/2}\beta|1\rangle \propto \alpha|0\rangle + e^{i\varphi}\beta|1\rangle,$$

and so if $U = R_z(\varphi)$, then $S_U = S_z(\varphi)$. (Here and elsewhere, we use the expression $A \propto B$ to mean that A and B may differ only by an overall phase factor, *i.e.*, there exists an angle $\omega \in \mathbb{R}$ such that $A = e^{i\omega}B$.) For any angle θ, define

$$R_y(\theta) = \begin{bmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{bmatrix}. \tag{35}$$

$R_y(\theta)$ is unitary, and it is straightforward to show that if $U = R_y(\theta)$, then $S_U = S_y(\theta)$. Thus any rotation S can be realized as S_U for some unitary U. Later, we will see a direct way of translating

between a 1-qubit unitary U and its corresponding rotation S_U. For completeness, we define a unitary corresponding to rotation of φ counterclockwise about the x-axis:

$$R_x(\varphi) = R_y(\pi/2)R_z(\varphi)R_y(-\pi/2) = \begin{bmatrix} \cos(\varphi/2) & -i\sin(\varphi/2) \\ -i\sin(\varphi/2) & \cos(\varphi/2) \end{bmatrix}. \tag{36}$$

Now to show that S_U must preserve orientation, we show that the orientation-reversing map M that maps each point (x, y, z) on the Bloch sphere to its antipodal point $(-x, -y, -z)$ is not of the form S_U for any unitary U. This suffices, because if S is any orientation-reversing rigid map of the Bloch sphere, then S^{-1} is also rigid and orientation-reversing, which means that the map $T = MS^{-1}$ is orientation-preserving and hence a rotation. Therefore, $T = S_V$ for some unitary V, as we just showed. But then $M = TS$, and so if we assume that $S = S_W$ for some unitary W, we then have $M = S_V S_W = S_{VW}$, a contradiction.

Suppose $M = S_U$ for some unitary U. Then, since U must reverse the directions of all spins, we must have, for example, $U|0\rangle = U|+z\rangle \propto |-z\rangle = |1\rangle$ and $U|1\rangle = U|-z\rangle \propto |+z\rangle = |0\rangle$. Expressing U as a matrix in the $\{|+z\rangle, |-z\rangle\}$ basis, we must then have

$$U = \begin{bmatrix} 0 & e^{i\sigma} \\ e^{i\tau} & 0 \end{bmatrix}$$

for some $\sigma, \tau \in \mathbb{R}$. Now consider the state

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{i(\tau-\sigma)/2}|1\rangle) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ e^{i(\tau-\sigma)/2} \end{bmatrix},$$

which corresponds to some point p on the equator of the Bloch sphere. We have

$$U|\psi\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & e^{i\sigma} \\ e^{i\tau} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ e^{i(\tau-\sigma)/2} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{i(\tau+\sigma)/2} \\ e^{i\tau} \end{bmatrix} = \frac{e^{i(\tau+\sigma)/2}}{\sqrt{2}} \begin{bmatrix} 1 \\ e^{i(\tau-\sigma)/2} \end{bmatrix} \propto |\psi\rangle.$$

So U does not change the state $|\psi\rangle$ more than by a phase factor, and thus S_U leaves the point p fixed, which means that $S_U \neq M$, a contradiction.[8]

Example. X, Y, and Z are unitary, so what are S_X, S_Y, and S_Z? It's easy to check that $X = iR_x(\pi)$, $Y = iR_y(\pi)$, and $Z = iR_z(\pi)$, and so (since phase factors don't matter) S_X, S_Y, and S_Z are rotations about the x-, y-, and z-axes, respectively, through the angle π (half a revolution).

Exercise 8.5 Prove the claim that if $U = R_y(\theta)$, then $S_U = S_y(\theta)$. [Hint: First check that $R_y(\theta)|+y\rangle \propto |+y\rangle$, and thus S_U fixes the point $(0, 1, 0)$ where the Bloch sphere intersects the $+y$-axis. Then check that $R_y(\theta)|+z\rangle = R_y(\theta)|0\rangle \propto \cos(\theta/2)|0\rangle + \sin(\theta/2)|1\rangle = |\uparrow_{\theta,0}\rangle$, so S_U moves the point $(0, 0, 1)$ to the point $(\sin\theta, 0, \cos\theta)$. Finally, check that $R_y(\theta)|+x\rangle = R_y(\theta)\left|\uparrow_{\pi/2,0}\right\rangle \propto \cos(\theta/2 + \pi/4)|0\rangle + \sin(\theta/2 + \pi/4)|1\rangle = \left|\uparrow_{\theta+\pi/2,0}\right\rangle$, so S_U moves the point $(1, 0, 0)$ to $(\cos\theta, 0, -\sin\theta)$.]

[8] This result has physical significance. It says that there is no single physical process that can reverse the spin of any isolated electron.

A direct translation. Here we give (without proof) a direct way to pass between the 2×2 unitary matrix U and the corresponding rotation S_U (a 3×3 real matrix), and vice versa, using the Pauli matrices. First we give some basic facts about rotations of $\mathbb{R}^n$ in general and of $\mathbb{R}^3$ in particular (some of which we've seen before). You can skip this list if you want.

1. An $n \times n$ matrix S with real entries gives a rigid (length- and angle-preserving) transformation of $\mathbb{R}^n$ iff
$$S^T S = I \,, \tag{37}$$
or equivalently, $SS^T = I$. (Here S^T denotes the transpose of S, and I denotes the $n \times n$ identity matrix.) In this case, we also say that S is an *orthogonal matrix*. Notice that, since S is real, $S^T = S^*$, and so S is orthogonal iff S is unitary.

2. Any orthogonal matrix has determinant ± 1. If the determinant is $+1$, then the transformation is orientation-preserving (i.e., a rotation); otherwise it is orientation-reversing.

3. Let S be a real $n \times n$ matrix. If S is a rotation of $\mathbb{R}^n$, then $S = S^c \neq 0$. (Here, S^c denotes the cofactor matrix of S.) The converse also holds if $n > 2$.

4. If n is odd, then every rotation S of $\mathbb{R}^n$ has 1 as an eigenvalue. That is, S fixes some nonzero vector $\hat{n} \in \mathbb{R}^n$. Thus if $n = 3$, then S moves points around some fixed axis (through the vector $\hat{n} \in \mathbb{R}^3$) counterclockwise through some angle $\psi \in [0, \pi]$—when viewing the origin from $\hat{n}$. (If you want $\pi < \psi < 2\pi$, then this is just the same as a counterclockwise rotation around $-\hat{n}$ through angle $2\pi - \psi$, which is in the interval $[0, \pi]$.)

5. If S is a rotation of $\mathbb{R}^3$, then $-1 \leqslant \operatorname{tr} S \leqslant 3$, and the angle of rotation (described above) is given by
$$\psi = \cos^{-1}\left(\frac{\operatorname{tr} S - 1}{2}\right) \,. \tag{38}$$
In particular, $\operatorname{tr} S = 3$ just when $\psi = 0$, that is, when $S = I$. Also, $\operatorname{tr} S = -1$ just when $\psi = \pi$. In either of these two special cases, $S^2 = I$, which implies that S is a symmetric matrix, because $S = S^{-1} = S^T$. If $0 < \psi < \pi$ (the general case), then S is not symmetric.

6. If S is as above and the angle of rotation ψ satisfies $0 < \psi < \pi$, then the vector $\hat{n}$ is unique up to multiplication by a positive scalar. It can be chosen to be
$$\hat{n} = ([S]_{32} - [S]_{23},\ [S]_{13} - [S]_{31},\ [S]_{21} - [S]_{12}) \,. \tag{39}$$
The norm of this particular vector is $\|\hat{n}\| = \sqrt{(1 + \operatorname{tr} S)(3 - \operatorname{tr} S)} = 2 \sin \psi$.

From unitaries to rotations. First, given a 2×2 unitary U, the corresponding rotation is given by the matrix
$$S_U = \frac{1}{2}\begin{bmatrix} \langle X, UXU^* \rangle & \langle X, UYU^* \rangle & \langle X, UZU^* \rangle \\ \langle Y, UXU^* \rangle & \langle Y, UYU^* \rangle & \langle Y, UZU^* \rangle \\ \langle Z, UXU^* \rangle & \langle Z, UYU^* \rangle & \langle Z, UZU^* \rangle \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \langle XU, UX \rangle & \langle XU, UY \rangle & \langle XU, UZ \rangle \\ \langle YU, UX \rangle & \langle YU, UY \rangle & \langle YU, UZ \rangle \\ \langle ZU, UX \rangle & \langle ZU, UY \rangle & \langle ZU, UZ \rangle \end{bmatrix} ,$$
recalling that $\langle A, B \rangle = \operatorname{tr}(A^* B)$ is the Hilbert-Schmidt inner product on $\mathcal{L}(\mathbb{C}^2)$. That is, for all $i, j \in \{1, 2, 3\}$,
$$[S_U]_{ij} = \frac{1}{2}\langle \sigma_i, U\sigma_j U^* \rangle = \frac{1}{2}\langle \sigma_i U, U\sigma_j \rangle \,.$$

Exercise 8.6 Find the 3×3 matrix S_U where

$$U = \frac{1}{5}\begin{bmatrix} 3 & -4i \\ -4i & 3 \end{bmatrix}.$$

Also find $\hat{n}$ (exact expression) and ψ (decimal approximation to three significant digits).

From rotations to unitaries. Now given some rotation S of $\mathbb{R}^3$ (S is a 3×3 real matrix), we find a U such that $S = S_U$. There are two cases:

- If $\operatorname{tr} S \neq -1$, then $S = S_U$ for exactly those U satisfying

$$\begin{aligned} U &\propto \frac{1}{2\sqrt{1+\operatorname{tr} S}}\left[(1+\operatorname{tr} S)I + i(([S]_{23}-[S]_{32})X + ([S]_{31}-[S]_{13})Y + ([S]_{12}-[S]_{21})Z)\right] \\ &= (\cos(\psi/2))I - \frac{i(\hat{n}\cdot\sigma)}{4\cos(\psi/2)} = (\cos(\psi/2))I - i\sin(\psi/2)\,(\hat{m}\cdot\sigma) = e^{-i\psi(\hat{m}\cdot\sigma)/2}, \end{aligned}$$

  where ψ and $\hat{n}$ are given by Equations (38) and (39), respectively, and $\hat{m} := \hat{n}/\|\hat{n}\|$ is the normalized version of $\hat{n}$. Here we use the fact that $\sqrt{1+\operatorname{tr} S} = 2\cos(\psi/2)$. I'll explain the last equation in the chain more fully next time. These expressions give the unique U with *positive trace* such that $S_U = S$ (and in addition, $\det U = 1$).

- If $\operatorname{tr} S = -1$, then any one of the following three alternatives is a characterization of all U such that $S = S_U$, provided it is well-defined:

$$\begin{aligned} U &\propto \frac{([S]_{11}+1)X + [S]_{21}Y + [S]_{31}Z}{\sqrt{2([S]_{11}+1)}}, \\ U &\propto \frac{[S]_{12}X + ([S]_{22}+1)Y + [S]_{32}Z}{\sqrt{2([S]_{22}+1)}}, \\ U &\propto \frac{[S]_{13}X + [S]_{23}Y + ([S]_{33}+1)Z}{\sqrt{2([S]_{33}+1)}}. \end{aligned}$$

  The i-th expression above (for $i \in \{1,2,3\}$) is well-defined iff $[S]_{ii} \neq -1$. This is true for at least one of the three for any rotation S of $\mathbb{R}^3$ with trace -1.

Exercise 8.7 Find a U such that

$$S_U = \frac{1}{169}\begin{bmatrix} -151 & 24 & 72 \\ -24 & 137 & -96 \\ -72 & -96 & -119 \end{bmatrix}.$$

Also give $\hat{n}$ and $\cos\psi$.

9 Week 5: Linear Algebra: Exponential Map, Spectral Theorem, etc.

The Exponential Map (Again). Equation (2) defines e^z via a power series for all scalars z. We can use the same power series to extend the definition to operators.

Definition 9.1 Let A be an operator in $\mathcal{L}(\mathcal{H})$ or an $n \times n$ matrix. Define

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^k}{k!} + \cdots = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \tag{40}$$

($A^0 = I$ by convention.)

If A is an operator or matrix, then so is e^A. The sum in (40) converges absolutely[9] for all A. The exponential map has may useful properties. Here's one of the most useful, which generalizes the familiar rule that $e^{z_1+z_2} = e^{z_1}e^{z_2}$ for scalars z_1, z_2.

Proposition 9.2 *If operators A and B commute (i.e., $AB = BA$), then*

$$e^{A+B} = e^A e^B.$$

Proof. This closely mirrors the standard proof for scalars. We manipulate the power series directly. Since A commutes with B, we can expand and rearrange factors in the expression $(A+B)^k$ to arrive at an operator version of the Binomial Theorem:

$$(A+B)^k = \sum_{j=0}^{k} \binom{k}{j} A^j B^{k-j},$$

for all integers $k \geqslant 0$. So,

$$\begin{aligned} e^{A+B} &= \sum_{k=0}^{\infty} \frac{(A+B)^k}{k!} = \sum_k \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} A^j B^{k-j} = \sum_k \frac{1}{k!} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} A^j B^{k-j} \\ &= \sum_k \sum_{j=0}^{k} \frac{A^j B^{k-j}}{j!(k-j)!} = \sum_k \sum_{j,\ell \geqslant 0 \ \&\ j+\ell=k} \frac{A^j B^\ell}{j!\ell!} \qquad \text{(setting } \ell := k-j\text{)} \\ &= \sum_{j=0}^{\infty} \sum_{\ell=0}^{\infty} \frac{A^j B^\ell}{j!\ell!} = \left(\sum_{j=0}^{\infty} \frac{A^j}{j!} \right) \left(\sum_{\ell=0}^{\infty} \frac{B^\ell}{\ell!} \right) = e^A e^B \ . \end{aligned}$$

□

We'll leave the other properties of e^A as exercises.

[9] We won't delve deeply into what it means for an infinite sequence of operators $A_1, A_2, \ldots$ to converge (absolutely or otherwise). One easy way to express the notion of convergence (among several equivalent ways) is to say that there exists an operator A such that for all vectors v, the sequence of vectors $A_1v, A_2v, \ldots$ converges to Av. The operator A, if it exists, must be unique, and we write $A = \lim_{n\to\infty} A_n$. Convergence of an infinite series of operators is equivalent to the convergence of the sequence of partial sums, as usual. Absolute convergence, which we don't bother to define here, implies that you can regroup and rearrange terms in the sum freely without worry.

Exercise 9.3 Verify the following for any operators or square matrices A and B and any $\theta \in \mathbb{R}$:

1. $e^0 = I$, where 0 is the zero operator. [Hint: Inspect the power series.]
2. $e^{-A} = (e^A)^{-1}$. [Hint: Use the previous item and Proposition 9.2.]
3. $e^{A^*} = (e^A)^*$. [Hint: You may use the fact that the adjoint of an infinite (convergent) sum is the sum of the adjoints. We know this already for finite sums.]
4. If A is Hermitean, then e^{iA} is unitary. [Hint: Use the previous two items and the fact that $(iA)^* = -iA$.]
5. A commutes with e^A. [Hint: Inspect the power series.]
6. If A and B commute, then so do e^A and e^B. [Hint: Use Proposition 9.2.]
7. If U is unitary, then $e^{UAU^*} = Ue^AU^*$. [Hint: Inspect the power series.]
8. If $A^2 = I$ (think Pauli matrices!), then $e^{i\theta A} = (\cos\theta)I + i(\sin\theta)A$. This is analogous to Exercise 2.3. [Hint: Inspect the power series.]
9. $$\begin{aligned} R_x(\theta) &= e^{-i\theta X/2}, \\ R_y(\theta) &= e^{-i\theta Y/2}, \\ R_z(\theta) &= e^{-i\theta Z/2}, \end{aligned}$$
   where $R_x(\theta)$, $R_y(\theta)$, and $R_z(\theta)$ are defined by Equations (34–36). It then follows from Proposition 9.2 that $R_x(\theta + \varphi) = R_x(\theta)R_x(\varphi)$ for all $\theta, \varphi \in \mathbb{R}$, and similarly for $R_y(\cdot)$ and $R_z(\cdot)$. [Hint: Use the previous item.]

Exercise 9.4 (Challenging) Let $\hat{n} = (x, y, z) \in \mathbb{R}^3$ such that $x^2+y^2+z^2 = 1$, and let $A = xX+yY+zZ$. Then $A^2 = I$ by Exercise 8.4. For angle $\omega \in \mathbb{R}$, define

$$R_{\hat{n}}(\omega) = e^{-i\omega A/2} = (\cos(\omega/2))I - i(\sin(\omega/2))A.$$

Show that if $U = R_{\hat{n}}(\omega)$, then S_U is a rotation of the Bloch sphere about the axis through $\hat{n}$ counterclockwise through angle ω. [Hint: Observe that rotating around $\hat{n}$ through angle ω is equivalent to

1. rotating the sphere so that $\hat{n}$ coincides with $(0, 0, 1)$ on the $+z$-axis, then
2. rotating around the $+z$-axis counterclockwise through angle ω, then
3. undoing the rotation in item 1 above, which moves $(0, 0, 1)$ back to $\hat{n}$.

(Let (θ, φ) be the spherical coordinates of $\hat{n}$. To achieve the first rotation, first rotate around $+z$ through angle $-\varphi$ to bring $\hat{n}$ into the x, z-plane, then rotate around $+y$ through angle $-\theta$.) Now verify via direct matrix multiplication that

$$R_{\hat{n}}(\omega) = R_z(\varphi)R_y(\theta)R_z(\omega)R_y(-\theta)R_z(-\varphi).$$

This decomposition is known as the S^3 *parameterization* of $R_{\hat{n}}(\omega)$.]

We need another linear algebraic detour.

Upper Triangular Matrices and Schur Bases. In the next topic, we'll be talking about basis-independent properties of operators, but we will occasionally need to introduce an orthonormal basis so that we can talk about matrices, and although all such bases are equivalent, some are more convenient than others. If $A \in \mathcal{L}(\mathcal{H})$ is an operator, a *Schur basis* for A is an orthonormal basis with respect to which A is represented by an *upper triangular* matrix, *i.e.*, an $n \times n$ matrix M whose entries below its diagonal are all zero: $[M]_{ij} = 0$ if $i > j$. Upper triangular matrices have many nice properties, so we'll sometimes choose a Schur basis when it is convenient. Particularly in this section, we will derive some facts about operators using a Schur basis. Theorem B.5 in Section B.2 shows that we can *always* choose a Schur basis:

Theorem 9.5 (Theorem B.5 in Section B.2) *Every $n \times n$ matrix is unitarily conjugate to an upper triangular matrix. That is, for every $n \times n$ matrix M, there is an upper triangular T and unitary U (both $n \times n$ matrices) such that $M = UTU^*$.*

Thus a Schur basis always exists for any linear operator. The proof of Theorem 2.1 uses the fact that every operator has an eigenvalue, which we'll discuss in the next topic.

One key property of an upper triangular matrix is that its determinant is just the product of its diagonal entries: if T is upper triangular, then

$$\det T = \prod_{i=1}^{n} [T]_{ii}. \tag{41}$$

Exercise 9.6 Show that if A and B are both upper triangular matrices, then so is AB, and for each $1 \leqslant i \leqslant n$, we have $[AB]_{ii} = [A]_{ii}[B]_{ii}$, that is, the diagonal entries just multiply individually.

Exercise 9.7 Show that if A is a nonsingular, upper triangular matrix, then A^{-1} is upper triangular. What are the diagonal entries of A^{-1} in terms of those of A?

Exercise 9.8 Show that if A is upper triangular, then so is e^A, and we have $[e^A]_{ii} = e^{[A]_{ii}}$ for all $1 \leqslant i \leqslant n$. [Hint: Use the results of Exercise 9.6 and Equation (40) defining e^A.]

Exercise 9.9 (One of my favorites.) Show that if A is any operator, then $\det e^A = e^{\operatorname{tr} A}$. [Hint: Pick a Schur basis for A, then use the previous exercise and Equation (41).]

Lower triangular matrices are defined analogously and have similar properties. A matrix is *diagonal* if it is both upper and lower triangular, *i.e.*, all its nondiagonal entries are zero.

Eigenvectors, Eigenvalues, and the Characteristic Polynomial. Let $A \in \mathcal{L}(\mathcal{H})$ be an operator. A nonzero vector $v \in \mathcal{H}$ such that $Av = \lambda v$, where $\lambda \in \mathbb{C}$, is called an *eigenvector* of A, and λ is its corresponding *eigenvalue*. Likewise, a scalar λ is an *eigenvalue* of A if it is the eigenvalue of some eigenvector of A. If λ is an eigenvalue of A, then we have $0 = Av - \lambda v = (A - \lambda I)v$, for some nonzero vector v. That is, the operator $A - \lambda I$ maps the nonzero vector v to 0, which means that $A - \lambda I$ is singular, which in turn means that $\det(A - \lambda I) = 0$. Conversely, if λ is a scalar such that

$\det(A-\lambda I)=0$, then $A-\lambda I$ is singular, and so it maps some nonzero vector v to 0, and so we have $(A-\lambda I)v=0$, or equivalently, $Av=\lambda v$. Thus v is an eigenvector of A with eigenvalue λ. Thus the eigenvalues of A are exactly those scalars λ such that $\det(A-\lambda I)=0$.

Let's write $A-\lambda I$ in matrix form with respect to some (any) orthonormal basis. Setting $a_{ij}=[A]_{ij}$, we get

$$A-\lambda I=\begin{bmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda \end{bmatrix},$$

where $n=\dim\mathcal{H}$. Fixing all the a_{ij} to be constant and considering λ to be a variable, one can show that $\det(A-\lambda I)$ is a polynomial in λ with degree n. This is the *characteristic polynomial* of A,[10] and we denote it $\mathrm{char}_A(\lambda)$. From our considerations above, the eigenvalues of A are precisely the roots of the polynomial char_A. Since $\mathbb{C}$ is algebraically closed (see the second lecture), char_A has n roots, and so A has exactly n eigenvalues, not necessarily all distinct. The (multi)set of eigenvalues of A is known as the *spectrum* of A.

Exercise 9.10 We know that char_A is basis-independent because it is defined in terms of basis-independent things. Show directly that

$$\mathrm{char}_A(\lambda)=\mathrm{char}_{UAU^*}(\lambda),$$

for any operator A, unitary operator U, and scalar λ.

The fact that A has at least one eigenvalue is a key ingredient in the proof that A has a Schur basis (Theorem B.5 in Section B.2), as well as in the proof of the Spectral Theorem, below. So now that we know that a Schur basis for A really exists, let's assume that we chose a Schur basis for A above, and so $a_{ij}=0$ for all $i>j$, and hence $A-\lambda I$ is also upper triangular. So taking the determinant, which is just the product of the diagonal entries, we get

$$\mathrm{char}_A(\lambda)=\det(A-\lambda I)=\prod_{i=1}^{n}(a_{ii}-\lambda). \tag{42}$$

From (42) it is clear that the eigenvalues of A—the roots of char_A—are exactly $a_{11},\ldots,a_{nn}$. This is true because we chose a basis making the matrix representing A upper triangular, but from this we get two useful, basis-independent facts: If $\lambda_1,\ldots,\lambda_n$ are the eigenvalues of A counted with multiplicities, then

- $\mathrm{tr}\,A=\sum_{i=1}^{n}\lambda_i$, and
- $\det A=\prod_{i=1}^{n}\lambda_i$.

[10] The characteristic polynomial of A is often defined instead as $\det(\lambda I-A)$, which, among other things, guarantees that its leading coefficient is 1. The two definitions coincide for even n, but for odd n, one is the negation of the other. But in any case, both polynomials have the same roots, which is the important thing.

Some of the coefficients of the polynomial char_A are familiar. If we expand (42) and group together powers of $(-\lambda)$, we get

$$\begin{aligned} \text{char}_A(\lambda) &= (-\lambda)^n + (a_{11} + a_{22} + \cdots + a_{nn})(-\lambda)^{n-1} + \cdots + a_{11}a_{22}\cdots a_{nn} \quad (43) \\ &= (-\lambda)^n + (\text{tr}\, A)(-\lambda)^{n-1} + \cdots + \det A. \quad (44) \end{aligned}$$

The constant term is $\det A$, which can also be seen by noting that this term is

$$\text{char}_A(0) = \det(A - 0I) = \det A.$$

Exercise 9.11 Find the eigenvalues of the 2×2 matrix $A = \begin{bmatrix} 3 & -1 \\ 4 & -2 \end{bmatrix}$. Find eigenvectors corresponding to each eigenvalue.

What are the eigenvalues of A^* in terms of those of A? We have,

$$\text{char}_{A^*}(\lambda) = \det(A^* - \lambda I) = \det((A - \lambda^* I)^*) = (\det(A - \lambda^* I))^* = (\text{char}_A(\lambda^*))^*,$$

and so $\text{char}_{A^*}(\lambda) = 0$ if and only if $\text{char}_A(\lambda^*) = 0$. Thus, the eigenvalues of A^* are the complex conjugates of the eigenvalues of A.

Eigenvectors and Eigenvalues of Normal Operators. Suppose A is a Hermitean operator. Then for any eigenvector v of A with eigenvalue λ, we have

$$\lambda\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Av\rangle = \langle A^*v, v\rangle = \langle Av, v\rangle = \langle \lambda v, v\rangle = \lambda^*\langle v, v\rangle.$$

Since $\langle v, v\rangle = \|v\|^2 > 0$, we get $\lambda = \lambda^*$. That is, A has only real eigenvalues. If λ_1 and λ_2 are distinct eigenvalues of A associated with eigenvectors v_1 and v_2, respectively, then

$$\lambda_2\langle v_1, v_2\rangle = \langle v_1, Av_2\rangle = \langle Av_1, v_2\rangle = \lambda_1^*\langle v_1, v_2\rangle = \lambda_1\langle v_1, v_2\rangle.$$

Thus if $\lambda_1 \neq \lambda_2$, then this can only be because $\langle v_1, v_2\rangle = 0$. In other words, eigenvectors of a Hermitean operator with distinct eigenvalues must be orthogonal.

An *eigenbasis* for an operator A is an orthonormal basis of eigenvectors of A. In such a basis, A is given by a diagonal matrix. If A is Hermitean, then A has an eigenbasis. This can be proved directly by a routine induction on the dimension of A, but it also a special case of a more general result.

An operator (or matrix) A is *normal* if it commutes with its adjoint, *i.e.*, $AA^* = A^*A$. It follows that a normal matrix must be square. Note that normality of matrices is unitarily invariant (if M is normal then any unitary conjugate of M is normal), and hence independent of the choice of orthonormal basis. This can be verified directly, or just by observing that normality is defined in terms of properties of the operator A itself, and is therefore basis-independent. Notice that all Hermitean operators, all unitary operators, and all operators represented by diagonal matrices (in some basis) are normal. The next theorem suggests that normal operators are especially important.

Theorem 9.12 (Spectral Theorem for Normal Operators) *Every normal operator has an eigenbasis. That is, if $A \in \mathcal{L}(\mathcal{H})$ is normal, then there is an orthonormal basis with respect to which A is represented by a diagonal matrix whose diagonal elements are the eigenvalues of A.*

In the rest of this section, we prove this theorem and explore some of its consequences. Section B.2 in Appendix B includes a proof of the Spectral Theorem using a Schur basis for A.[11] The proof of the Spectral Theorem we give here is independent from that and does not need a Schur basis for A. We first prove some technical facts about normal matrices and operators.

Lemma 9.13 *Let M be an $n \times n$ matrix, given in block form as*

$$M = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] ,$$

where A and D are square matrices. If M is normal, then

$$\langle B, B\rangle = \langle C, C\rangle .$$

Proof. We have, ignoring all but the top-left block,

$$\begin{aligned} M^*M &= \left[\begin{array}{c|c} A^* & C^* \\ \hline B^* & D^* \end{array}\right]\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|c} A^*A + C^*C & \cdots \\ \hline \cdots & \cdots \end{array}\right] , \\ MM^* &= \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]\left[\begin{array}{c|c} A^* & C^* \\ \hline B^* & D^* \end{array}\right] = \left[\begin{array}{c|c} AA^* + BB^* & \cdots \\ \hline \cdots & \cdots \end{array}\right] . \end{aligned}$$

M normal means $M^*M = MM^*$, so equating the top-left blocks, we have

$$A^*A + C^*C = AA^* + BB^* .$$

Taking the trace of both sides and using the properties of the trace gives

$$\operatorname{tr} A^*A + \operatorname{tr} C^*C = \operatorname{tr} AA^* + \operatorname{tr} BB^* = \operatorname{tr} A^*A + \operatorname{tr} B^*B .$$

Thus $\langle C, C\rangle = \operatorname{tr} C^*C = \operatorname{tr} B^*B = \langle B, B\rangle$ as asserted. □

Lemma 9.14 *Let $\mathcal{H}$ be a Hilbert space. Suppose that $R \in \mathcal{L}(\mathcal{H})$ is normal and there is a subspace $V \subseteq \mathcal{H}$ such that R maps V into V (i.e., $R(V) \subseteq V$). Then R maps $V^\perp$ into $V^\perp$, and R restricted to V (respectively $V^\perp$) is a normal operator in $\mathcal{L}(V)$ (respectively $\mathcal{L}(V^\perp)$).*

Proof. Let $n = \dim(\mathcal{H})$ and let $k = \dim(V)$ (whence $\dim(V^\perp) = n - k$). We can assume that $1 \leqslant k < n$, otherwise the statement is trivial. We can choose an orthonormal basis $\mathcal{B} := \{b_1, \ldots, b_n\}$ for $\mathcal{H}$ such that $\{b_1, \ldots, b_k\}$ is an orthonormal basis for V and $\{b_{k+1}, \ldots, b_n\}$ is an orthonormal basis for $V^\perp$. Letting M be the matrix of R with respect to $\mathcal{B}$, we can write M in block form as

$$M = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] ,$$

[11] In Theorem B.6, we show that if a matrix is both normal and upper triangular, then it is diagonal. Hence A is represented in this basis as a diagonal matrix. That is, any Schur basis of a normal operator A is an eigenbasis of A.This is an alternate proof of the Spectral Theorem for Normal Operators.

where A is a $k \times k$ matrix. For all $1 \leqslant j \leqslant k$, we have $Rb_j \in V$, so Rb_j (represented by the jth column of M) is a linear combination of $b_1, \ldots, b_k$ only. This means that $C = 0$, which implies $\langle B, B\rangle = \langle C, C\rangle = 0$ by the previous lemma. Thus $B = 0$, and we have

$$M = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & D \end{array}\right] .$$

From this we get that $R(V^\perp) \subseteq V^\perp$, because for $k < j \leqslant n$, the jth column of M, which represents Rb_j, is a linear combination of $b_{k+1}, \ldots, b_n$ only. Furthermore,

$$\left[\begin{array}{c|c} A^*A & 0 \\ \hline 0 & D^*D \end{array}\right] = M^*M = MM^* = \left[\begin{array}{c|c} AA^* & 0 \\ \hline 0 & DD^* \end{array}\right] ,$$

and it follows by equating blocks that A and D are both normal matrices. □

We now have the tools in place to give a short, tidy proof of the Spectral Theorem, Theorem 9.12.

Proof. [Spectral Theorem, Theorem 9.12] The proof is by induction on $\dim(\mathcal{H})$. If $\dim(\mathcal{H}) = 1$, then the statement follows from the fact that every nonzero vector is an eigenvector of A and hence forms an eigenbasis for A. Now suppose that $\dim(\mathcal{H}) = n > 1$, and assume the statement holds for proper subspaces of $\mathcal{H}$. Let λ be an eigenvalue of A with some corresponding eigenvector $b_1 \in \mathcal{H}$. Without loss of generality, we can assume $\|b_1\| = 1$. Let $V := \{ab_1 : a \in \mathbb{C}\}$ be the 1-dimensional subspace of $\mathcal{H}$ spanned by b_1. For any $a \in \mathbb{C}$, we have $Aab_1 = aAb_1 = a\lambda b_1 \in V$, that is, A maps V into V. By Lemma 9.14, A also maps $V^\perp$ into $V^\perp$, and its restriction A' to $V^\perp$ is a normal operator in $\mathcal{L}(V^\perp)$. Since $\dim(V^\perp) = n - 1 < n$, we apply the inductive hypothesis to get an eigenbasis $\{b_2, \ldots, b_n\} \subseteq V^\perp$ for A'. Then $\{b_1, b_2, \ldots, b_n\}$ is an eigenbasis for A. □

If $U \in \mathcal{L}(\mathcal{H})$ is unitary, then U is normal, and choosing an eigenbasis for U, we represent it as a diagonal matrix D. For each $1 \leqslant i \leqslant n$, let $d_i = [D]_{ii}$. Since $DD^* = I$ (D is unitary), we have

$$1 = [I]_{ii} = [DD^*]_{ii} = d_i d_i^* = |d_i|^2,$$

and thus the eigenvalues of U all lie on the unit circle in $\mathbb{C}$.

We can get more milage out of Lemma 9.14. The next lemma generalizes what we showed for Hermitean operators.

Lemma 9.15 *If $A \in \mathcal{L}(\mathcal{H})$ is normal, and v_1 and v_2 are eigenvectors of A with distinct eigenvalues, then $\langle v_1, v_2\rangle = 0$.*

Proof. Let λ_1 and λ_2 be the eigenvalues of v_1 and v_2, respectively. By assumption, $\lambda_1 \neq \lambda_2$. We can assume without loss of generality that $\|v_1\| = 1$. Let V be the 1-dimensional subspace of $\mathcal{H}$ spanned by v_1. As in the previous proof, we see that A maps V into V. In the spirit of Gram-Schmidt, let

$$y := v_2 - \langle v_1, v_2\rangle v_1 .$$

Then $\langle v_1, y\rangle = 0 = \langle y, v_1\rangle$, and so $\langle y, av_1\rangle = a\langle y, v_1\rangle = 0$ for all $a \in \mathbb{C}$. Thus $y \in V^\perp$. Lemma 9.14 states that A must then map $V^\perp$ into $V^\perp$. In particular, $\langle Ay, v_1\rangle = \langle v_1, Ay\rangle = 0$. We have

$$\begin{aligned} 0 &= \langle v_1, Ay\rangle = \langle v_1 \ , \ Av_2 - \langle v_1, v_2\rangle Av_1\rangle = \langle v_1 \ , \ \lambda_2 v_2 - \langle v_1, v_2\rangle \lambda_1 v_1\rangle \\ &= \lambda_2\langle v_1, v_2\rangle - \langle v_1, v_2\rangle\lambda_1\langle v_1, v_1\rangle = (\lambda_2 - \lambda_1)\langle v_1, v_2\rangle . \end{aligned}$$

Since $\lambda_2 - \lambda_1 \neq 0$, we must have $\langle v_1, v_2 \rangle = 0$. □

If A is an operator and λ is an eigenvalue of A, then we define the *eigenspace* of A with respect to λ as

$$\mathcal{E}_\lambda(A) = \{v \in \mathcal{H} : Av = \lambda v\}.$$

This is a subspace of $\mathcal{H}$ with positive dimension.

Corollary 9.16 *If A is normal, then its eigenspaces are mutually orthogonal and span $\mathcal{H}$. The dimension of each eigenspace is the same as the multiplicity of the corresponding eigenvalue.*

The following corollary is useful because it shows that any normal operator is a unique linear combination of projectors that form a csop, thus revealing how projectors form the building blocks of normal operators through their eigenvalues. It is essentially an alternate formulation of the Spectral Theorem.

Corollary 9.17 *If A is normal, then there is a unique set $\{(P_1, \lambda_1), \ldots, (P_k, \lambda_k)\}$, such that the $\lambda_j \in \mathbb{C}$ are all distinct, the set $\{P_1, \ldots, P_k\}$ is a complete set of orthogonal projectors, and*

$$A = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_k P_k \,. \tag{45}$$

Furthermore, $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of A, and each P_j orthogonally projects onto $\mathcal{E}_{\lambda_j}(A)$.

Proof. Suppose $\dim(\mathcal{H}) = n$. Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of A, and for all $1 \leqslant j \leqslant k$ let P_j be the orthogonal projector onto $\mathcal{E}_{\lambda_j}(A)$. Choose an eigenbasis $\{b_1, \ldots, b_n\}$ for A. For each $1 \leqslant i \leqslant n$, let $1 \leqslant j_i \leqslant k$ be such that $Ab_i = \lambda_{j_i} b_i$. Then b_i lies in the eigenspace $\mathcal{E}_{\lambda_{j_i}}(A)$. This means that $b_i = P_{j_i} b_i$, and it follows that for all $P_j b_i = P_j P_{j_i} b_i = 0$ for all $j \neq j_i$. And so we have

$$Ab_i = \lambda_{j_i} b_i = \lambda_{j_i} P_{j_i} = \left(\sum_{j=1}^{k} \lambda_j P_j \right) b_i \,.$$

Thus both sides of Equation (45) act the same way on each b_i, and so they must be equal. This proves existence.

To show uniqueness, we show that for any decomposition

$$A = \sum_{j=1}^{\ell} \mu_j Q_j$$

where the μ_j are pairwise distinct scalars and $\{Q_1, \ldots, Q_\ell\}$ is a complete set of orthogonal projectors, it must be that the μ_j are all the eigenvalues of A with the respective Q_j projecting onto the corresponding eigenspaces (and so incidentally, $\ell = k$). Fix a j such that $1 \leqslant j \leqslant \ell$, and notice that $AQ_j = \mu_j Q_j$. For any $v \in \mathcal{H}$, we then have $AQ_j v = \mu_j Q_j v$, and thus $Q_j v$ is an eigenvector of A provided $Q_j v \neq 0$. Since $Q_j \neq 0$, such a v must exist, and this shows that μ_j is an eigenvalue of A and Q_j maps $\mathcal{H}$ into the corresponding eigenspace $\mathcal{E}_{\mu_j}(A)$. This implies $\{\mu_1, \ldots, \mu_\ell\} \subseteq \{\lambda_1, \ldots, \lambda_k\}$. We'll be done if we show two things: (1) that Q_j maps $\mathcal{H}$ *onto* $\mathcal{E}_{\mu_j}(A)$; and (2) every eigenvalue of A is in $\{\mu_1, \ldots, \mu_\ell\}$. Let $u \neq 0$ be any eigenvector of A with some eigenvalue λ.

We claim that $Q_j u = 0$ for all j such that $\mu_j \neq \lambda$. We have

$$u = Iu = \sum_{j=1}^{\ell} Q_j u \,. \tag{46}$$

Fix j and suppose $\mu_j \neq \lambda$. Let $v = Q_j u$. Then $v \in \mathcal{E}_{\mu_j}(A)$ by what we showed about Q_j above, and if $v \neq 0$, then v is an eigenvector of A with eigenvalue $\mu_j \neq \lambda$. But by Lemma 9.15, we have

$$0 = \langle u, v \rangle = \langle u, Q_j u \rangle = \langle u, Q_j^* Q_j u \rangle = \langle Q_j u, Q_j u \rangle = \|v\|^2 \,,$$

which implies $0 = v = Q_j u$, and that establishes the claim. Now if $\lambda \neq \mu_j$ for *all* $1 \leqslant j \leqslant \ell$, then $u = 0$ by Equation (46); contradiction. Hence $\lambda \in \{\mu_1, \ldots, \mu_\ell\}$. Since λ is an arbitrary eigenvalue of A, we have $\{\lambda_1, \ldots, \lambda_k\} = \{\mu_1, \ldots, \mu_\ell\}$ (and since the members of each set are pairwise distinct, we must also have $k = \ell$). Now for any $1 \leqslant j \leqslant \ell$ and for any $u \in \mathcal{E}_{\mu_j}(A)$, Equation (46) and the claim give $u = Q_j u$, that is, Q_j fixes pointwise all elements of $\mathcal{E}_{\mu_j}(A)$. Thus Q_j maps $\mathcal{H}$ onto $\mathcal{E}_{\mu_j}(A)$. □

The right-hand side of Equation (45) is called the *spectral decomposition* of A.

Exercise 9.18 Show that if A is a normal operator with spectral decomposition

$$A = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_k P_k$$

as in Corollary 9.17, then for any integer $m \geqslant 0$,

$$A^m = \lambda_1^m P_1 + \lambda_2^m P_2 + \cdots + \lambda_k^m P_k.$$

(We define $A^0 = I$ by convention.) [Hint: Induction on m.]

Exercise 9.19 Show that if A is a normal operator with spectral decomposition

$$A = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_k P_k$$

as in Corollary 9.17, then

$$e^A = e^{\lambda_1} P_1 + e^{\lambda_2} P_2 + \cdots + e^{\lambda_k} P_k.$$

[Hint: Use the last exercise.]

We'll be dealing with normal operators almost exclusively from now on.

Exercise 9.20 We know that any Hermitean operator is normal with real eigenvalues. Prove the converse: any normal operator with only real eigenvalues is Hermitean. [Hint: Use an eigenbasis.]

Exercise 9.21 We know that any unitary operator is normal with eigenvalues on the unit circle in $\mathbb{C}$. Prove the converse: any normal operator with all eigenvalues on the unit circle is unitary. [Hint: Use an eigenbasis.]

Scalar Functions Applied to Operators. Let $\Omega \subseteq \mathbb{C}$ be some set and suppose $f : \Omega \to \mathbb{C}$ is some function mapping scalars to scalars. It is often natural and useful to extend the definition of f to apply to operators $A \in \mathcal{L}(\mathcal{H})$, where $\mathcal{H}$ is a Hilbert space, with the results also being operators in $\mathcal{L}(\mathcal{H})$. There are at least two situations where this can be done:

1. The value $f(x)$ is expressible as a convergent power series about some point $x_0 \in \Omega$: for every $x \in \Omega$,
$$f(x) = \sum_{i=0}^{\infty} c_i (x - x_0)^i$$
for some coefficients $c_0, c_1, \ldots \in \mathbb{C}$ independent of x. For any operator $A \in \mathcal{L}(\mathcal{H})$, we then define
$$f(A) := \sum_{i=0}^{\infty} c_i (A - x_0 I)^i \,,$$
provided the right-hand side converges.

2. The function f is arbitrary. For any *normal* operator $A \in \mathcal{L}(\mathcal{H})$ all of whose eigenvalues are contained in Ω, we define
$$f(A) := \sum_{j=1}^{k} f(\lambda_j) P_j = f(\lambda_1) P_1 + \cdots + f(\lambda_k) P_k \,,$$
where $A = \lambda_1 P_1 + \cdots + \lambda_k P_k$ is the unique spectral decomposition of A given by Equation (45) of Corollary 9.17, above.

We've seen an example of item (1) above with the natural exponential map $z \mapsto e^z$ defined on all of $\mathbb{C}$, and Exercises 9.18 and 9.19 show that this definition agrees with item (2) as well. In fact, it can be shown that if both conditions (1) and (2) hold for some f and A, then the two definitions will coincide. We will see an instance of item (2) below, when we take the square root of a positive operator. It often the case that a special property that f with respect to scalars has an analogous (but perhaps weaker) property when applied to operators. For example, $e^{z+w} = e^z e^w$ for all $z, w \in \mathbb{C}$, and for operators $A, B \in \mathcal{L}(\mathcal{H})$ we have $e^{A+B} = e^A e^B$ as well, *provided* A and B commute.

Here are two more general facts. The first says among other things that this concept is covariant under unitary conjugation. The second applies to item (2) specifically.

Proposition 9.22 *Let function $f : \Omega \to \mathbb{C}$ and operator $A \in \mathcal{L}(\mathcal{H})$ satisfy the conditions of either item (1) or item (2) above. Then A and $f(A)$ commute. Furthermore, for any unitary operator $U \in \mathcal{L}(\mathcal{H})$, we have that f and UAU^* also satisfy the same condition(s), and $f(UAU^*) = Uf(A)U^*$.*

Proposition 9.23 *Suppose $f : \Omega \to \mathbb{C}$ and $A \in \mathcal{L}(\mathcal{H})$ satisfy the conditions of item (2) above. Then $f(A)$ is the unique operator in $\mathcal{L}(\mathcal{H})$ such that, for any $v \in \mathcal{H}$ and $\lambda \in \mathbb{C}$, if v is an eigenvector of A with eigenvalue λ, then v is an eigenvector of $f(A)$ with eigenvalue $f(\lambda)$.*

Positive Operators.

Definition 9.24 An operator $A \in \mathcal{L}(\mathcal{H})$ is *positive* or *positive semidefinite* (written $A \geqslant 0$) iff $v^*Av \geqslant 0$ for all $v \in \mathcal{H}$. We say that A is *strictly positive* or *positive definite* (written $A > 0$) if $v^*Av > 0$ for all nonzero $v \in \mathcal{H}$.

Since $u^*Av = \langle u, Av\rangle$ for all vectors $u, v \in \mathcal{H}$, positivity of A is equivalent to $\langle v, Av\rangle \geqslant 0$ for all $v \in \mathcal{H}$, or in Dirac notation, $\langle\psi|A|\psi\rangle \geqslant 0$ for all $|\psi\rangle \in \mathcal{H}$. Obviously, strict positivity implies positivity.

For example, the zero operator $0 \in \mathcal{L}(\mathcal{H})$ and the identity operator $I \in \mathcal{L}(\mathcal{H})$ are clearly positive: $v^*0v = 0$ and $v^*Iv = v^*v = \|v\|^2 \geqslant 0$ for all v. In fact, $I > 0$ as well.

Exercise 9.25 Verify that if $A \geqslant 0$ and $B \geqslant 0$ are positive operators and $a \geqslant 0$ is a nonnegative real number, then $A + B \geqslant 0$ and $aA \geqslant 0$.

Positive operators play a huge role in the study of quantum information, so it is worth spending some time with them.

Exercise 9.26 (A bit challenging) Show for any operator $A \in \mathcal{L}(\mathcal{H})$ that A is Hermitean if and only if $v^*Av \in \mathbb{R}$ for all $v \in \mathcal{H}$. (Thus every positive operator is Hermitean and hence normal.) [Hint: The forward direction is easy. For the reverse direction, consider the matrix elements of A with respect to some orthonormal basis $b_1, \ldots, b_n$. Consider three types of cases:

1. $v = b_k$ for some k. What does this tell you about the diagonal elements $[A]_{kk}$?
2. $v = b_k + b_j$ for some $k \neq j$. This allows you to relate $[A]_{kj}$ and $[A]_{jk}$ in some way.
3. $v = b_k + ib_j$ for the same k, j above. This allows you to relate $[A]_{kj}$ and $[A]_{jk}$ further.]

Exercise 9.27 Show that $A \geqslant 0$ if and only if A is normal and all its eigenvalues are nonnegative real numbers. (It follows that if $A \geqslant 0$, then $\operatorname{tr} A \geqslant 0$.) [Hint: Use the previous exercise.]

Exercise 9.28 Show that if $A \geqslant 0$ and $\operatorname{tr} A = 0$, then $A = 0$. [Hint: Use the previous exercise.]

Exercise 9.29 Show that the zero operator is the only operator A satisfying $A \geqslant 0$ and $-A \geqslant 0$. [Hint: Use the previous two exercises.]

Exercise 9.30 Show that the following are equivalent for any operator A:

1. $A > 0$.
2. A is normal, and all its eigenvalues are positive reals.
3. $A \geqslant 0$ and A is nonsingular.

You may have noticed that you can determine a lot about a normal operator by its spectrum. This is not too surprising, because

- most properties we've been looking at of the underlying matrices are basis-invariant (i.e., invariant under unitary conjugation),
- every normal operator is representable by a diagonal matrix in some basis, and
- the spectrum of a diagonal matrix is just the set of diagonal elements of the matrix.

Each entry in the following table is easily checked by representing the operator as a diagonal matrix with respect to an eigenbasis.

A normal operator is ...	**... iff all its eigenvalues are ...**
nonsingular (invertible)	nonzero
Hermitean	real
unitary	on the unit circle
positive	nonnegative
strictly positive	positive
a projector	either 0 or 1
the identity	1
the zero operator	0

If $A \geqslant 0$ is a positive operator, then there exists a unique positive operator $B \geqslant 0$ such that $B^2 = A$. We denote B by $A^{1/2}$ or by $\sqrt{A}$. To see that B exists, we decompose

$$A = \lambda_1 P_1 + \cdots + \lambda_k P_k$$

uniquely according to Corollary 9.17. Since $A \geqslant 0$, we have $\lambda_j \geqslant 0$ for $1 \leqslant j \leqslant k$. Now let

$$B := \sqrt{\lambda_1} P_1 + \cdots + \sqrt{\lambda_k} P_k.$$

B has eigenvalues $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k} \geqslant 0$, so $B \geqslant 0$. By Exercise 9.18, we get

$$B^2 = \left(\sqrt{\lambda_1}\right)^2 P_1 + \cdots + \left(\sqrt{\lambda_k}\right)^2 P_k = A.$$

To show uniqueness, suppose that $B, C \geqslant 0$ such that $B^2 = C^2 = A$. Using Corollary 9.17 again, decompose

$$\begin{aligned} B &= \mu_1 P_1 + \cdots + \mu_k P_k, \\ C &= \nu_1 Q_1 + \cdots + \nu_\ell Q_\ell. \end{aligned}$$

So,

$$B^2 = \mu_1^2 P_1 + \cdots + \mu_k^2 P_k = A = \nu_1^2 Q_1 + \cdots + \nu_\ell^2 Q_\ell = C^2.$$

Note that the μ_j are distinct and nonnegative (same with the ν_j), and therefore so are the μ_j^2 (same with the ν_j^2). Then since the decomposition of A from Corollary 9.17 is unique, we must

have $\{(P_1, \mu_1^2), \ldots, (P_k, \mu_k^2)\} = \{(Q_1, \nu_1^2), \ldots, (Q_\ell, \nu_\ell^2)\}$. Thus $k = \ell$ and $\{(P_1, \mu_1), \ldots, (P_k, \mu_k)\} = \{(Q_1, \nu_1), \ldots, (Q_k, \nu_k)\}$, because all the $\mu_j \geqslant 0$ and $\nu_j \geqslant 0$. So we must have $B = C$.

Notice that, since the same projectors are involved in the decompositions of A and B, it follows that the eigenvectors of A and B coincide, and the corresponding eigenvalues of B are the square roots of those of A.

The square root function applied to positive operators that we've just defined is an example of a scalar function applied to an operator that we discussed in the previous topic. The fact that any positive operator has a (positive) square root is useful in many places. For example, we get the following theorem:

Theorem 9.31 *Let A and B be any positive operators over a Hilbert space $\mathcal{H}$. Then $\langle A, B\rangle \geqslant 0$, with equality holding if and only if $AB = 0$. (Recall that $\langle A, B\rangle = \operatorname{tr}(A^*B)$.)*

Proof. A is Hermitean, so we have

$$\langle A, B\rangle = \operatorname{tr}(AB) = \operatorname{tr}\left(\sqrt{A}\sqrt{A}\sqrt{B}\sqrt{B}\right) = \operatorname{tr}\left(\sqrt{B}\sqrt{A}\sqrt{A}\sqrt{B}\right) = \operatorname{tr}(C^*C) = \langle C, C\rangle\,,$$

where $C := \sqrt{A}\sqrt{B}$. By positive definiteness, we have $\langle C, C\rangle \geqslant 0$ with equality holding if and only if $C = 0$. Now if $AB = 0$, then $\langle A, B\rangle = \operatorname{tr}(AB) = \operatorname{tr} 0 = 0$, which shows one direction of the "if and only if." Thus it only remains to show that if $C = 0$, then $AB = 0$. Suppose that $C = 0$. Then $AB = \sqrt{A}\sqrt{A}\sqrt{B}\sqrt{B} = \sqrt{A}C\sqrt{B} = 0$. □

Exercise 9.32 Show that if A and U are operators, $A \geqslant 0$, and U is unitary, then $UAU^* \geqslant 0$ and $\sqrt{UAU^*} = U\sqrt{A}U^*$. [Hint: By uniqueness, it suffices to show that $U\sqrt{A}U^* \geqslant 0$ and that $(U\sqrt{A}U^*)^2 = UAU^*$.]

Proposition 9.33 *Let A be an $n \times n$ matrix and B an $m \times n$ matrix, for positive integers m and n. If $A \geqslant 0$, then $BAB^* \geqslant 0$. (Note that BAB^* is an $m \times m$ matrix.)*

Proof. For any m-dimensional column vector v, we have

$$v^*BAB^*v = v^*B\sqrt{A}\sqrt{A}B^*v = \left\langle \sqrt{A}B^*v,\, \sqrt{A}B^*v\right\rangle = \left\| \sqrt{A}B^*v\right\|^2 \geqslant 0\,.$$

(The inner product is of n-dimensional vectors.) □

Exercise 9.34 Let $P_1, \ldots, P_k$ be nonzero projectors, and let $A = P_1 + \cdots + P_k$. Show that if A is a projector, then $P_iP_j = 0$ for all $i \neq j$. [Hint: Square both sides of the equation above, simplify, take the trace, then apply Theorem 9.31.]

Exercise 9.34 establishes that Condition (1) follows from Condition (2) in Definition 5.11.

If A is any operator, then A^*A is always positive: for any vector v, we have

$$\langle v, A^*Av\rangle = \langle Av, Av\rangle = \|Av\|^2 \geqslant 0.$$

We denote the positive operator $\sqrt{A^*A}$ by $|A|$. This is analogous to the absolute value of a scalar, but keep in mind that $|A|$ is an operator and not a scalar.

Exercise 9.35 Show that if A is any operator, then $A \geqslant 0$ if and only if $A = |A|$.

Exercise 9.36 Show that if A is any operator, then the eigenvalues of $|A|$ are the absolute values of the eigenvalues of A.

Here we summarize many equivalent ways of characterizing positive operators into a single proposition.

Proposition 9.37 *Let $\mathcal{H}$ be an n-dimensional Hilbert space for some $n > 0$, and let $A \in \mathcal{L}(\mathcal{H})$ be an operator (equivalently, a $n \times n$ matrix with respect to some standard basis $\{e_1, \ldots, e_n\}$ of $\mathcal{H}$). The following are equivalent:*

1. *$A \geqslant 0$ (i.e., $v^*Av \geqslant 0$ for all $v \in \mathcal{H}$).*
2. *A is normal with all eigenvalues $\geqslant 0$.*
3. *$A = |A|$.*
4. *There exists an $n \times n$ matrix B such that $A = B^*B$.*
5. *There exist a positive integer m and an $m \times n$ matrix B such that $A = B^*B$.*
6. *$\langle B, A\rangle \geqslant 0$ for every positive operator $B \in \mathcal{L}(\mathcal{H})$.*
7. *$\langle uu^*, A\rangle \geqslant 0$ for every unit vector $u \in \mathcal{H}$.*
8. *There exist vectors $v_1, \ldots, v_n \in \mathcal{H}$ such that $[A]_{ij} = \langle v_i, v_j\rangle$ for all $1 \leqslant i, j \leqslant n$.*

Proof. (1) $\Leftrightarrow$ (2) by Exercise 9.27. (1) $\Leftrightarrow$ (3) by Exercise 9.35. It then suffices show (3) $\Rightarrow$ (4) $\Rightarrow$ (5) $\Rightarrow$ (1) and (1) $\Rightarrow$ (6) $\Rightarrow$ (7) $\Rightarrow$ (1). For (3) $\Rightarrow$ (4), set $B := \sqrt{|A|}$. (4) $\Rightarrow$ (5) is obvious. For (5) $\Rightarrow$ (1), we have, for any $v \in \mathcal{H}$,

$$v^*Av = v^*B^*Bv = (Bv)^*(Bv) = \langle Bv, Bv\rangle = \|Bv\|^2 \geqslant 0\,,$$

hence $A \geqslant 0$. (1) $\Rightarrow$ (6) follows from Theorem 9.31. We have (6) $\Rightarrow$ (7) because $uu^* \geqslant 0$ (check it!). For (7) $\Rightarrow$ (1), for any nonzero $v \in \mathcal{H}$, let $u := v/\|v\|$. Then u is a unit vector, and we have

$$v^*Av = \operatorname{tr}(v^*Av) = \operatorname{tr}(vv^*A) = \operatorname{tr}((vv^*)^*A) = \langle vv^*, A\rangle = \|v\|^2\langle uu^*, A\rangle \geqslant 0\,.$$

(Obviously, $v^*Av = 0$ if $v = 0$.) We've now established that (1)–(7) are equivalent.

For (4) $\Rightarrow$ (8), let $A = B^*B$ for $B \in \mathcal{L}(\mathcal{H})$. Then for all $1 \leqslant i, j \leqslant n$,

$$[A]_{ij} = e_i^*Ae_j = e_i^*B^*Be_j = \langle Be_i, Be_j\rangle\,.$$

Set $v_k := Be_k$ for all $1 \leqslant k \leqslant n$. For (8) $\Rightarrow$ (1), let $v_1, \ldots, v_n \in \mathcal{H}$ be given satisfying (8). Then for any $v \in \mathcal{H}$, we can write $v = \sum_{i=1}^n x_ie_i$ for some $x_1, \ldots, x_n \in \mathbb{C}$. Then

$$v^*Av = \sum_{i,j=1}^n x_i^*x_j(e_i^*Ae_j) = \sum_{i,j} x_i^*x_j[A]_{ij} = \sum_{i,j} x_i^*x_j\langle v_i, v_j\rangle = \left\langle \sum_i x_iv_i,\ \sum_j x_jv_j \right\rangle = \langle u, u\rangle \geqslant 0$$

by the positive definiteness of $\langle \cdot, \cdot \rangle$, where $u := \sum_{k=1}^{n} x_k v_k$. □

Before leaving this topic, we define and give some basic properties of a binary $\leqslant$ relation on $\mathcal{L}(\mathcal{H})$, or equivalently, on square matrices. This relation arises naturally from the notion of operator positivity.

Definition 9.38 Let $\mathcal{H}$ be an n-dimensional Hilbert space and let $A, B \in \mathcal{L}(\mathcal{H})$ (equivalently, A and B are $n \times n$ matrices). We say that $A \leqslant B$ iff $B - A \geqslant 0$, *i.e.*, iff $B - A$ is a positive operator.

Most of Proposition 9.39, below, follows from the properties of positive operators we have established above. For technical convenience, we state things in terms of matrices rather than operators.

Proposition 9.39 *Let n and m be positive integers, and let A and B be any $n \times n$ matrices.*

1. *The binary relation $\leqslant$ on $n \times n$ matrices is reflexive and transitive.*
2. *If $A \leqslant B$ and $B \leqslant A$, then $A = B$.*
3. *For any $n \times n$ matrix C and scalars $0 \leqslant a \leqslant b$, if $A \leqslant B$ then $A + C \leqslant B + C$ and $aA \leqslant bB$.*
4. *For any $m \times n$ matrix D, if $A \leqslant B$, then $DAD^* \leqslant DBD^*$.*
5. *For any $n \times n$ matrix F, if A and B are Hermitean, $A \leqslant B$, and $F \geqslant 0$, then $\langle F, A \rangle \leqslant \langle F, B \rangle$ (and both quantities are real).*
6. *If A and B are Hermitean and $A \leqslant B$, then $\operatorname{tr} A \leqslant \operatorname{tr} B$ (and both quantities are real).*
7. *If A is a projector, then $0 \leqslant A \leqslant I$.*

Proof. Exercise. [Hint: (6) follows from (5) by setting $F := I$.] □

Items (1) and (2) are the axioms for $\leqslant$ being a partial order.

Corollary 9.40 *If A and B are operators, $A \leqslant I$, and $B \geqslant 0$, then $\operatorname{tr}(AB) \leqslant \operatorname{tr} B$ (and both quantities are real).*

Proof. Since $I - A \geqslant 0$ and hence is Hermitean, it follows that A is Hermitean. We then have

$$\operatorname{tr}(AB) = \operatorname{tr}(BA) = \langle B, A \rangle \leqslant \langle B, I \rangle = \operatorname{tr} B \, .$$

□

Commuting Operators. In this topic, we'll prove the fundamental result that commuting normal operators always share a common eigenbasis, and so they are simultaneously diagonalizable. This is a stronger version of the Spectral Theorem, which only deals with one normal operator.

Theorem 9.41 *Let $\mathcal{C}$ be an arbitrary family*[12] *of normal operators in $\mathcal{L}(\mathcal{H})$, any two of which commute, i.e., $AB = BA$ for all $A, B \in \mathcal{C}$. Then there is an orthonormal basis $\mathcal{B}$ of $\mathcal{H}$ that is an eigenbasis for all operators in $\mathcal{C}$ simultaneously.*

To prove Theorem 9.41, we will use Lemma 9.14 paired with the following fundamental property of commuting operators:

Lemma 9.42 *Let $A, B \in \mathcal{L}(\mathcal{H})$ be commuting operators, and let $E \subseteq \mathcal{H}$ be any eigenspace of A. Then B maps E into E.*

Proof. Let $E = \mathcal{E}_\lambda(A)$ be the eigenspace of A corresponding to some eigenvalue λ of A. Then for any $v \in E$, we have

$$ABv = BAv = B(\lambda v) = \lambda Bv\ .$$

Thus either $Bv = 0$ or Bv is an eigenvector of A with eigenvalue λ. In either case, $Bv \in E$, which proves the lemma. □

Proof of Theorem 9.41. This proof is somewhat similar to that of the Spectral Theorem. We proceed by induction on $n = \dim(\mathcal{H})$. If $n = 1$, then all operators in $\mathcal{C}$ are scalar multiples of the identity operator, making $\mathcal{H}$ itself a common eigenspace of all operators in $\mathcal{C}$, and any unit vector in $\mathcal{H}$ is then a common eigenbasis.

Now assume $n > 1$ and the theorem holds for any Hilbert space of dimension less than n. We prove it true for dimension n by first finding at least one common eigenvector for all the operators in $\mathcal{C}$. We will then continue as in the proof of the Spectral Theorem. To find a common eigenvector, we construct a finite, strictly descending chain of subspaces

$$\mathcal{H} = E_0 \supset E_1 \supset E_2 \supset \cdots \supset E_k\ ,$$

where E_i is a proper subspace of E_{i-1} for all $1 \leqslant i \leqslant k$, and $\dim(E_k) > 0$. (Any such chain must be finite, because the dimension decreases by at least 1 for each successive E_i.) We will do this in such a way that all nonzero vectors in E_k are common eigenvectors of all the operators in $\mathcal{C}$, that is, all operators in $\mathcal{C}$ are multiples of the identity when restricted to E_k. For convenience, for each $0 \leqslant i \leqslant k$, we also define $\mathcal{C}_i$ to be the set of all restrictions to E_i of operators in $\mathcal{C}$. We will maintain the invariant that every operator $A \in \mathcal{C}_i$ maps E_i into itself (that is, $A \in \mathcal{L}(E_i)$), from which it follows from Lemma 9.14 that A is a normal operator on E_i.

Now for the construction.[13] First, set $E_0 := \mathcal{H}$, whence $\mathcal{C}_0 := \mathcal{C}$. Note that the above invariant holds trivially for $i = 0$. Then for $i := 1, 2, 3, \ldots$ in increasing order (until we stop), do the following:

- If all operators in $\mathcal{C}_{i-1}$ are multiples of the identity on E_{i-1}, then set $k := i - 1$ and STOP.

[12] $\mathcal{C}$ need not be finite—or even countable.

[13] The construction makes a series of arbitrary choices, so it is not unique.

- Otherwise, choose an operator $A \in \mathcal{C}_{i-1}$ that is not a multiple of the identity, and let E_i be any eigenspace of A. Note that such an E_i exists (as one does for any operator), that E_i is a proper subspace of E_{i-1}, and that $\dim(E_i) > 0$. Also note that every operator in $\mathcal{C}_{i-1}$ commutes with A and thus maps E_i into itself by Lemma 9.42. From this one can see that the invariant is maintained for i.

By construction, every nonzero vector in E_k is an eigenvector of all the operators in $\mathcal{C}_k$ (and hence in $\mathcal{C}$). Let $\{b_1, \cdots, b_d\}$ be any orthonormal basis for E_k, where $d = \dim(E_k) > 0$. Since all operators in $\mathcal{C}$ map E_k to itself, by Lemma 9.14, all these operators also map $E_k^\perp$ into itself and act normally on $E_k^\perp$. Define $\mathcal{C}^\perp$ to be the set of restrictions to $E_k^\perp$ of operators in $\mathcal{C}$. We now apply the inductive hypothesis with space $E_k^\perp$ and set of operators $\mathcal{C}^\perp$ to obtain an orthonormal basis $\{b_{d+1}, \ldots, b_n\} \subseteq E_k^\perp$ of common eigenvectors of all operators in $\mathcal{C}^\perp$ (and hence all operators in $\mathcal{C}$). Combining, we now have an orthonormal basis $\mathcal{B} := \{b_1, \ldots, b_n\}$ of $\mathcal{H}$ consisting of common eigenvectors of all operators in $\mathcal{C}$. □

10 Week 5: Tensor products

Tensor Products and Combining Physical Systems. Suppose we have two physical systems S and T with state spaces $\mathcal{H}_S$ and $\mathcal{H}_T$, respectively, and we want to consider the two systems together as a single system ST. What is the state space of ST? Quantum mechanics says that the state space of ST is completely determined by $\mathcal{H}_S$ and $\mathcal{H}_T$ via a construction called the *tensor product*. We'll first describe the tensor product of matrices, then we'll discuss the tensor product in a basis-independent way.

Let A be an $m \times n$ matrix and let B be an $r \times s$ matrix (m, n, r, s are arbitrary positive integers). The *tensor product* of A and B (also called the *outer product* or the *direct product* or the *Kronecker product*) is the $mr \times ns$ matrix given in block form by

$$A \otimes B = \left[\begin{array}{c|c|c|c} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ \hline a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{array}\right].$$

We collect the standard, easily verifiable properties of the $\otimes$ operation here in one place.

Proposition 10.1 *For any matrices A, B, C, D and scalars $a, b \in \mathbb{C}$, the following equations hold provided the operations involved are well-defined:*

1. $a \otimes b = ab$, *where we identify scalars with 1×1 matrices as usual.*
2. *More generally, if A is $m \times 1$ (a column vector) and B is $1 \times n$ (a row vector), then $A \otimes B = AB$.*
3. $A \otimes (B + aC) = A \otimes B + a(A \otimes C)$ *and* $(A + aB) \otimes C = A \otimes C + a(B \otimes C)$, *that is, $\otimes$ is* bilinear *(linear in both arguments).*
4. $(A \otimes B) \otimes C = A \otimes (B \otimes C)$, *that is, $\otimes$ is associative.*

5. $(A \otimes B)(C \otimes D) = AC \otimes BD$. *(This is worth memorizing because we'll use it all the time.)*

6. $(A \otimes B)^* = A^* \otimes B^*$.

7. $\operatorname{tr}(A \otimes B) = (\operatorname{tr} A)(\operatorname{tr} B)$.

Exercise 10.2 Give the 4×4 matrices for $I \otimes X$, $X \otimes I$, $X \otimes Y$, and $Z \otimes Z$.

Exercise 10.3 Show that if A and B are Hermitean (respectively, unitary), then $A \otimes B$ is Hermitean (respectively, unitary).

A special case is when $u = (u_1, \ldots, u_m) \in \mathbb{C}^m$ and $v = (v_1, \ldots, v_n) \in \mathbb{C}^n$ are column vectors. Then

$$u \otimes v = \begin{bmatrix} \frac{u_1 v}{u_2 v} \\ \hline \vdots \\ \hline u_m v \end{bmatrix} = \begin{bmatrix} u_1 v_1 \\ \vdots \\ u_1 v_n \\ u_2 v_1 \\ \vdots \\ \vdots \\ u_m v_n \end{bmatrix} \in \mathbb{C}^{mn}.$$

If $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_n\}$ are the standard bases for $\mathbb{C}^m$ and $\mathbb{C}^n$ respectively as in Equation (5), then it is clear that $\{e_i \otimes f_j : 1 \leqslant i \leqslant m \;\&\; 1 \leqslant j \leqslant n\}$ is the standard basis for $\mathbb{C}^{mn}$. In fact, if we let $\{g_1, \ldots, g_{mn}\}$ be the standard basis of $\mathbb{C}^{mn}$, then

$$e_i \otimes f_j = g_{(i-1)n+j}$$

for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. If $w = (w_1, \ldots, w_m) \in \mathbb{C}^m$ and $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$, then for the standard inner product we have

$$\langle u \otimes v, w \otimes x \rangle = (u \otimes v)^*(w \otimes x) = (u^* \otimes v^*)(w \otimes x) = u^* w \otimes v^* x = \langle u, w \rangle \langle v, x \rangle.$$

From this it is easy to see that if $\{b_1, \ldots, b_m\}$ and $\{c_1, \ldots, c_n\}$ are *any* orthonormal bases for $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively, then $\{b_i \otimes c_j : 1 \leqslant i \leqslant m \;\&\; 1 \leqslant j \leqslant n\}$ is an orthonormal basis for $\mathbb{C}^{mn}$. Indeed, we have

$$\langle b_i \otimes c_j, b_k \otimes c_\ell \rangle = \langle b_i, b_k \rangle \langle c_j, c_\ell \rangle = \delta_{ik} \delta_{j\ell},$$

which is 1 if $i = k$ and $j = \ell$ and is 0 otherwise.

This last bit suggests that we can define the tensor product in a basis-independent way, applied to (abstract) vectors and operators. If $\mathcal{H}$ and $\mathcal{J}$ are Hilbert spaces, then we can define a Hilbert space $\mathcal{H} \otimes \mathcal{J}$ (the *tensor product* of $\mathcal{H}$ and $\mathcal{J}$) together with a bilinear map $\otimes : \mathcal{H} \times \mathcal{J} \to \mathcal{H} \otimes \mathcal{J}$, mapping any pair of vectors $u \in \mathcal{H}$ and $v \in \mathcal{J}$ to a vector $u \otimes v \in \mathcal{H} \otimes \mathcal{J}$, such that if $\{b_1, \ldots, b_m\}$ and $\{c_1, \ldots, c_n\}$ are orthonormal bases for $\mathcal{H}$ and $\mathcal{J}$, respectively, then $\{b_i \otimes c_j : 1 \leqslant i \leqslant m \;\&\; 1 \leqslant j \leqslant n\}$ is an orthonormal basis for $\mathcal{H} \otimes \mathcal{J}$. We'll call such a basis a *product basis*. We won't do it here, but it can be shown that these two rules—bilinearity and the basis rule—define in essence the Hilbert space $\mathcal{H} \otimes \mathcal{J}$ uniquely. Notice that the basis rule implies that the dimension of $\mathcal{H} \otimes \mathcal{J}$ is the product of the dimensions of $\mathcal{H}$ and $\mathcal{J}$.

It's worth pointing out that not all vectors in $\mathcal{H} \otimes \mathcal{J}$ are of the form $u \otimes v$ for $u \in \mathcal{H}$ and $v \in \mathcal{J}$. For example, the column vector $(1, 0, 0, 1) = e_1 + e_4$ cannot be written as the single tensor product of two 2-dimensional column vectors. It can, however, be written as the sum of two tensor products:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

In general a vector in $\mathcal{H} \otimes \mathcal{J}$ may not be a tensor product, but it is always a linear combination of them (which is clear by our discussion about bases, above), *i.e.*, the tensor products span the space $\mathcal{H} \otimes \mathcal{J}$.

We're not done overloading the $\otimes$ symbol. Given the definition of $\mathcal{H} \otimes \mathcal{J}$ just described, we can extend $\otimes$ to apply to operators as well as vectors. For example, we can extend it to a map $\otimes : \mathcal{L}(\mathcal{H}) \times \mathcal{L}(\mathcal{J}) \to \mathcal{L}(\mathcal{H} \otimes \mathcal{J})$ by *defining* the action of an operator $A \otimes B$ on a vector $u \otimes v \in \mathcal{H} \otimes \mathcal{J}$:

$$(A \otimes B)(u \otimes v) = Au \otimes Bv.$$

One can show that this definition is consistent, and since $\mathcal{H} \otimes \mathcal{J}$ is spanned by vectors of the form $u \otimes v$, this defines the operator $A \otimes B$ uniquely by linearity. We could define $\otimes$ on dual vectors and other kinds of linear maps, e.g., mapping from one space to another space.

Picking orthonormal bases for $\mathcal{H}$ and $\mathcal{J}$ allows us to represent objects such as vectors, dual vectors, operators, or what have you, in both spaces as matrices. When we do this, the abstract and matrix-based notions of $\otimes$ completely coincide, as is the case with the other linear algebraic constructs that we've seen, *e.g.*, adjoint, trace, et cetera. This idea (that the two notions should coincide) guides us in any further extensions of the $\otimes$ operation that we may wish to use.

Back to Combining Physical Systems. If S and T are physical systems with state spaces $\mathcal{H}_S$ and $\mathcal{H}_T$ as before, then the state space of the combined system is $\mathcal{H}_{ST} = \mathcal{H}_S \otimes \mathcal{H}_T$. If $|\varphi\rangle_S \in \mathcal{H}_S$ is a state of S and $|\psi\rangle_T \in \mathcal{H}_T$ is a state of T (we occasionally add subscripts to make clear which state goes with which system), then $|\varphi\rangle_S \otimes |\psi\rangle_T$ is a state of ST, which we interpret as saying, "The system S is in state $|\varphi\rangle_S$, and the system T is in state $|\psi\rangle_T$." (We'll often drop the $\otimes$ and write $|\varphi\rangle_S \otimes |\psi\rangle_T$ simply as $|\varphi\rangle_S|\psi\rangle_T$, or even just $|\varphi, \psi\rangle$ if the meaning is clear. The same holds for bras as well as kets.) As we've seen, however, there can be states of ST that can't be written as a single tensor product, for example, the two-qubit state $(|0\rangle|0\rangle + |1\rangle|1\rangle)/\sqrt{2}$. These states are called *entangled states*, whereas states of the form $|\varphi\rangle_S|\psi\rangle_T$ are called *separable states* or *tensor product states*. More on this later.

How does this look in the density operator formalism? Easy answer: exactly the same, at least for separable states. Let $\rho_S = |\varphi\rangle\langle\varphi|$ be the density operator corresponding to $|\varphi\rangle$ of system S, and let $\rho_T = |\psi\rangle\langle\psi|$ be the density operator corresponding to $|\psi\rangle$ of system T (subscripts dropped). Then the density operator for the combined system should be

$$\rho_{ST} = (|\varphi\rangle|\psi\rangle)(|\varphi\rangle|\psi\rangle)^* = (|\varphi\rangle|\psi\rangle)(|\varphi\rangle^*|\psi\rangle^*) = (|\varphi\rangle|\psi\rangle)(\langle\varphi|\langle\psi|) = |\varphi\rangle\langle\varphi| \otimes |\psi\rangle\langle\psi| = \rho_S \otimes \rho_T.$$

So we take the tensor product of the density operators just as we would do with the vectors in the original formulation. For the two-qubit entangled state example $(|0\rangle|0\rangle + |1\rangle|1\rangle)/\sqrt{2}$ above, which

we abbreviate as $(|00\rangle + |11\rangle)/\sqrt{2}$, the corresponding density operator is

$$\rho = \left(\frac{|00\rangle + |11\rangle}{\sqrt{2}}\right)\left(\frac{\langle 00| + \langle 11|}{\sqrt{2}}\right) = \frac{1}{2}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

If S and T are isolated from each other (and the outside world), then each evolves in time according to a unitary operator, say U for system S and V for system T. U and V are called *local operations*. In this case, $U \otimes V$ is the unitary giving the time evolution of the combined system: for tensor product state $|\varphi\rangle_S \otimes |\psi\rangle_T$, we have $(U \otimes V)(|\varphi\rangle_S \otimes |\psi\rangle_T) = U|\varphi\rangle_S \otimes V|\psi\rangle_T$, which is again a tensor product state. If S and T are brought together so that they *interact*, then the unitary giving the evolution of the combined system ST might not be able to be written as a single tensor product of unitaries for S and T respectively.

Exercise 10.4 Let $\mathcal{H}_S$ and $\mathcal{H}_T$ be Hilbert spaces, and let $P_1, \ldots, P_k \in \mathcal{L}(\mathcal{H}_S)$ be a complete set of orthogonal projectors for $\mathcal{H}_S$. Show that $P_1 \otimes I, \ldots, P_k \otimes I$ is a complete set of orthogonal projectors for $\mathcal{H}_S \otimes \mathcal{H}_T$, where I is the identity operator on $\mathcal{H}_T$. (The latter set represents a projective measurement on the system S when viewed from the combined system ST.)

Continuing the idea of Exercise 10.4, let $P_1, \ldots, P_k \in \mathcal{L}(\mathcal{H}_S)$ and $Q_1, \cdots, Q_\ell \in \mathcal{L}(\mathcal{H}_T)$ be complete sets of orthogonal projectors for systems S and T, respectively. Suppose that the combined system ST is in some arbitrary state $|\psi\rangle$ and that Alice measures system S using the first set of projectors. Then the exercise illustrates how she is actually measuring system ST with projectors $P_1 \otimes I, \ldots, P_k \otimes I$, where I is the identity operator on $\mathcal{H}_T$. She'll see some outcome i with some probability, and the state of ST will collapse to some $|\psi_i\rangle$ according to the usual rules. If Bob now measures system T using the second set of projectors when ST is in state $|\psi_i\rangle$ (which is tantamount to measuring ST with projectors $I \otimes Q_1, \ldots, I \otimes Q_\ell$, where I is the identity on $\mathcal{H}_S$), he will see some outcome j with some probability, and the system ST will then be in some state $|\psi_{ij}\rangle$, which depends on both Alice's outcome i and Bob's outcome j. Alternatively, Bob may do his measurement on T first and Alice does hers on S second. We won't bother to prove it here, but it can be easily shown mathematically that the joint probability $\Pr[i, j]$ of Alice seeing i and Bob seeing j is the same regardless of who does their measurement first, and the same goes for the post-measurement state $|\psi_{ij}\rangle$. Thus we can consider Alice and Bob doing their measurements simultaneously and independently of each other. Furthermore, we can consider the two measurements combined into a single projective measurement of ST, with projectors $\{P_i \otimes Q_j : 1 \leqslant i \leqslant k \;\&\; 1 \leqslant j \leqslant \ell\}$, where each projector $P_i \otimes Q_j$ corresponds to the outcome (i, j). Caveat: even though Alice's and Bob's measurements can be done independently of each other, the probabilities $\Pr[i]$ of Alice seeing i and $\Pr[j]$ of Bob seeing j may be correlated (*i.e.*, dependent) if $|\psi\rangle$ is an entangled state. We'll see a specific example of this later.

The No-Cloning Theorem. Quantum states cannot be duplicated in general. The following theorem makes this precise.

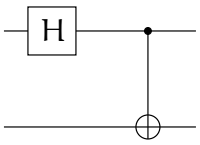


Figure 3: Sample quantum circuit with two qubits. Time moves from left to right in the figure. The gate H is applied first to the first qubit, then CNOT is applied to both qubits.

Theorem 10.5 (No-Cloning Theorem) *Let $\mathcal{H}$ be a Hilbert space of dimension at least two, and let $|0\rangle \in \mathcal{H}$ be a fixed unit vector. There is no unitary operator $U \in \mathcal{L}(\mathcal{H} \otimes \mathcal{H})$ such that $U|\psi\rangle|0\rangle \propto |\psi\rangle|\psi\rangle$ for any unit vector $|\psi\rangle \in \mathcal{H}$.*

Proof. Suppose U exists as above, and let $|\varphi\rangle, |\psi\rangle \in \mathcal{H}$ be any two unit vectors. Since U is unitary, we have

$$\begin{aligned}
\langle\varphi|\psi\rangle &= \langle\varphi|\psi\rangle\langle 0|0\rangle \\
&= (\langle\varphi|\langle 0|)(|\psi\rangle|0\rangle) \\
&= (\langle\varphi|\langle 0|)U^*U(|\psi\rangle|0\rangle) \\
&= (U|\varphi\rangle|0\rangle)^*U(|\psi\rangle|0\rangle) \\
&\propto (\langle\varphi|\langle\varphi|)(|\psi\rangle|\psi\rangle) \\
&= \langle\varphi|\psi\rangle^2,
\end{aligned}$$

and thus $|\langle\varphi|\psi\rangle| = |\langle\varphi|\psi\rangle|^2$, which implies $|\langle\varphi|\psi\rangle|$ is either 0 or 1, i.e., $|\varphi\rangle$ and $|\psi\rangle$ are either orthogonal or colinear. But clearly we can choose $|\varphi\rangle$ and $|\psi\rangle$ such that this is not the case. □

Quantum Circuits. The *quantum circuit* has become the *de facto* standard theoretical model of quantum computation. It is equivalent to the other standard model—the *quantum Turing machine*, or QTM—but it is easier to work with and represent visually. Quantum circuits are closely analogous to classical Boolean circuits, and we'll compare them occasionally.

A quantum circuit consists of some number of qubits, called a *quantum register*, represented by horizontal wires. The qubits start in some designated state, representing the input to the circuit. From time to time, we may act on one or more qubits in the circuit by applying a *quantum gate*, which is just a unitary operator applied to the corresponding qubits. A typical circuit with a two-qubit register is shown in Figure 3. To keep track, we number the qubits in the register from top to bottom, so that the topmost qubit is the first, etc. At any given time, the register is in some quantum state $|\psi\rangle \in \mathcal{H} \otimes \cdots \otimes \mathcal{H} = \mathcal{H}^{\otimes n}$, where $\mathcal{H}$ is here the state space of a single qubit, and n is the number of qubits in the register. We choose an orthonormal basis for $\mathcal{H}^{\otimes n}$ by taking tensor products of the individual one-qubit basis vectors $|0\rangle$ and $|1\rangle$. We call this basis the *computational basis* for the register. For example, a typical computational basis vector in $\mathcal{H}^{\otimes 5}$ is

$$|0\rangle \otimes |0\rangle \otimes |1\rangle \otimes |0\rangle \otimes |1\rangle = |0\rangle|0\rangle|1\rangle|0\rangle|1\rangle = |00101\rangle.$$

In this state, the first, second, and fourth qubits are 0, and the third and fifth qubits are 1. The state space of an n-qubit register has dimension 2^n, with computational basis vectors representing all

the 2^n possible values of n bits, listed in the usual binary order: $|00\cdots00\rangle$, $|00\cdots01\rangle$, $|00\cdots10\rangle$, etc., through $|11\cdots11\rangle$.

In the circuit diagram, the state of the register evolves in time from left to right. In Figure 3, for example, the first gate that is applied is the leftmost gate, i.e., the H gate applied to the first qubit. Here, we are not using H as a variable to describe any one-qubit gate, but rather we use H to denote a useful one-qubit gate, known as the *Hadamard gate*, given by

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

Note that

$$\begin{aligned} H|0\rangle &= (|0\rangle + |1\rangle)/\sqrt{2}, \\ H|1\rangle &= (|0\rangle - |1\rangle)/\sqrt{2}, \end{aligned}$$

or more succinctly,

$$H|b\rangle = \frac{|0\rangle + (-1)^b|1\rangle}{\sqrt{2}},$$

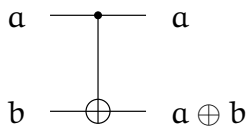
for any $b \in \{0,1\}$. Clearly, $H = (X+Z)/\sqrt{2}$ and $H^2 = I$. We also have $H \propto R_{(1,0,1)/\sqrt{2}}(\pi)$, and so H rotates the Bloch sphere 180° around the line through $(1,0,1)$, swapping the $+z$-axis with the $+x$-axis.

Note that although it looks as if we are only applying H to the first qubit, were are really transforming the state $|\psi\rangle \in \mathcal{H} \otimes \mathcal{H}$ of the entire two-qubit register via the unitary $H \otimes I$, where I is the one-qubit identity operator representing the fact that we are not acting on the second qubit. Suppose that the initial state of the register is $|00\rangle$. After the H gate is applied, the state becomes

$$|\psi_1\rangle = (H \otimes I)|00\rangle = \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)|0\rangle = \frac{|00\rangle + |10\rangle}{\sqrt{2}}.$$

11 Week 6: Quantum gates

The next gate in Figure 3 is another very useful, two-qubit gate called a *controlled NOT* or *C-NOT* gate, acting on both qubits. In a C-NOT gate, the small black dot connects to the *control* qubit (here, the first qubit) and the $\oplus$ end connects to the *target* qubit. If the control is $|0\rangle$, then the target does not change; if the control is $|1\rangle$, then the target's Boolean value is flipped $|0\rangle \leftrightarrow |1\rangle$ (logical NOT). The control qubit is unchanged regardless. Here it is schematically for any $a, b \in \{0, 1\}$ (here, $\oplus$ represents bitwise exclusive OR, *i.e.*, bitwise addition modulo 2):



The matrix for the C-NOT gate above, with the first qubit being the control and the second being the target, is

$$P_0 \otimes I + P_1 \otimes X = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Here X is the usual Pauli X operator, which swaps 0 with 1, and hence represents logical NOT. If the control and target qubits were reversed, then the gate would be

$$I \otimes P_0 + X \otimes P_1 = \left[\begin{array}{c|c} P_0 & P_1 \\ \hline P_1 & P_0 \end{array}\right] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

After the C-NOT gate is applied to the state $|\psi_1\rangle$ in Figure 3, the new and final state of the circuit is

$$|\psi_2\rangle = \text{C-NOT}|\psi_1\rangle = \text{C-NOT}\left(\frac{|00\rangle + |10\rangle}{\sqrt{2}}\right) = \frac{|00\rangle + |11\rangle}{\sqrt{2}}.$$

Keep in mind that the C-NOT gate (as with any quantum gate) acts *linearly* on the superposition $(|00\rangle + |10\rangle)/\sqrt{2}$, that is, it acts on each basis vector component of the superposition individually, and the overall result is the superposition of the individual results.

Every quantum circuit built this way represents a single unitary operator acting on the state space of all its qubits. Note that the individual gates are applied from left to right, which is opposite of how operators are applied in mathematical expressions.

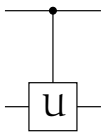
C-NOT is a *classical gate*. A classical gate is one that maps computational basis vectors to computational basis vectors. It can be described in non-quantum terms as a Boolean gate. Each column of its matrix has a single 1 with the other entries 0. In order to be a legitimate quantum gate, the matrix must be unitary, which means that the 1's must appear in all different rows. Such a matrix, with exactly one 1 in every row and every column and the other entries 0, is called a *permutation matrix* because it permutes the standard basis column vectors.

Exercise 11.1 Verify that every permutation matrix is unitary.

The C-NOT gate is one example of a controlled gate. More generally, if U is a unitary gate on k qubits, we can define the $(k+1)$-qubit *controlled* U *gate* to be

$$\text{C-}U = P_0 \otimes I + P_1 \otimes U = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & U \end{array}\right],$$

where in this case the control qubit is the first qubit. The matrix would be different if the control were not the first qubit, but the rule is the same in any case: If the control qubit is 0, then nothing happens with the other (target) qubits. If the control is 1, then U is applied to the target qubits. The control qubit is unchanged regardless. Here's how we draw it in the case where U acts on a single qubit:



In this context, the C-NOT gate is just a controlled X gate C-X.

We've seen two classical gates so far: X and C-NOT. We'll see some others in a bit. The other Pauli gates are not classical. The Pauli Z gate, for example, leaves the Boolean value (0 or 1) of the qubit unchanged, but introduces a phase factor (-1) if the value is 1. Z rotates the Bloch sphere 180 degrees about the $+z$-axis. Here are some other commonly used (nonclassical) gates:

$$S = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$$

is known as the *phase gate*. Note that $S \propto R_z(\pi/2)$ and that $S^2 = Z$. S rotates the Bloch sphere counterclockwise about the $+z$ axis 90 degrees.

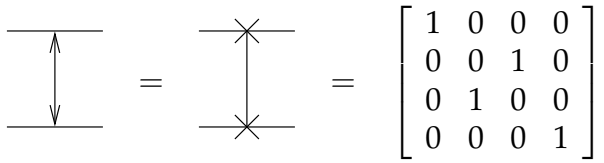
$$T = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1+i}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}.$$

For some obscure reason, this gate is known as the $\pi/8$ *gate*, maybe because

$$T \propto R_z(\pi/4) = \begin{bmatrix} e^{-i\pi/8} & 0 \\ 0 & e^{i\pi/8} \end{bmatrix}.$$

We have $T^2 = S$, and T rotates the Bloch sphere counterclockwise 45° about the $+z$-axis. Notice that T is the *only* one-qubit gate we've seen so far that does not map all axes to axes (*i.e.*, x-, y-, and z-axes) in the Bloch sphere. I'd call the three gates Z, S, and T *conditional phase-shift gates*, that leave the Boolean value of the qubit unchanged while introducing various phase factors conditioned on the qubit having Boolean value 1.

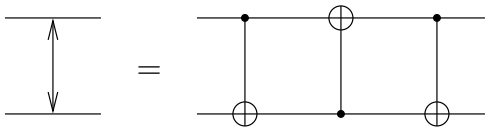
Here's another two-qubit classical gate, the *SWAP gate*:



$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The first depiction is mine and other people's; the second is the one the textbook uses. The SWAP gate just exchanges the Boolean values of the two qubits it acts on, fixing $|00\rangle$ and $|11\rangle$ but mapping $|01\rangle$ to $|10\rangle$ and vice versa.

Exercise 11.2 This is an entirely classical exercise. Show that



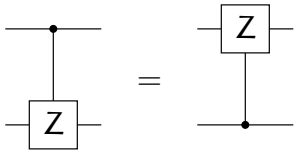
[Hint: Rather than multiplying matrices, which can be time-consuming, just compare what the two circuits do to the four possible basis states.]

Exercise 11.3 Do Exercise 4.16 on pages 178–179 of the text.

Exercise 11.4 This is a nonclassical exercise in several parts. It will help you to simplify circuits by inspection, based on some circuit identities. It mirrors Exercises 4.13 and 4.17–4.20 on pages 177–180 of the text. An item may use previous items.

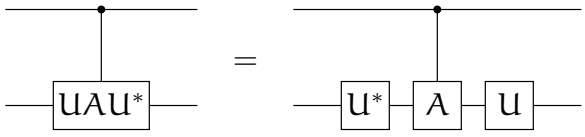
1. Verify directly that $HXH = Z$ and that $HZH = X$ (oh yes, and that $HYH = -Y$).

2. Verify that



   What is the matrix of this gate? The same is true for the C-S and C-T gates.

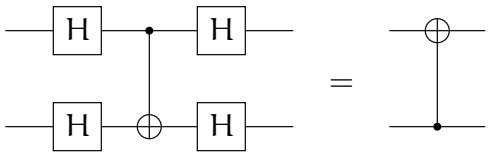
3. Show that, for any unitary gates U and A,



   (Remember that in the expression UAU^*, operators are applied from right to left, but in the circuit, gates are applied from left to right.) [Hint: Consider separately the case when the control qubit is $|0\rangle$ and when it is $|1\rangle$. To show equality of two linear operators generally, you only need to show that they both act the same on the vectors of some basis.]

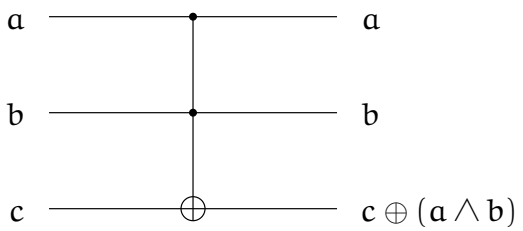
4. Construct a C-Z gate using a single C-NOT gate and two H gates. Similarly, construct a C-NOT gate using a single C-Z gate and two H gates.

5. Using the previous items, show that



   Note that gates acting on separate qubits commute, and so it doesn't matter which of the gates is applied first, and the order can be freely switched, provided that there are no gates in between that connect the qubits together. You can think of the gates as being applied simultaneously if you like.

Finally, we introduce a three-qubit classical gate known as the *Toffoli gate*, which is really a controlled controlled NOT gate:



There are two control qubits and one target qubit. The control qubits are unchanged, and the target is flipped if and only if both of the controls are 1.

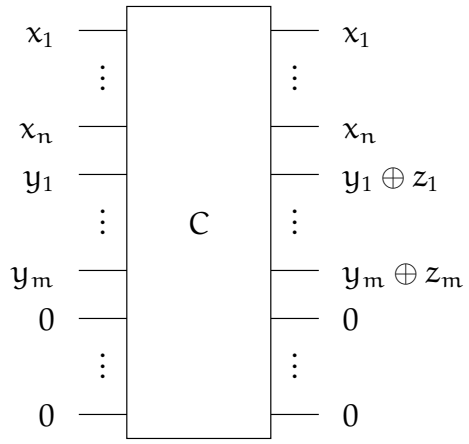
Quantum Circuits Versus Boolean Circuits. Are quantum circuits with unitary gates as powerful as classical Boolean circuits? You may have already noticed some similarities and differences between the two circuit models:

- Both types of circuits carry bit values on wires which are acted on by gates.
- Quantum gates can create superpositions from basis states, but Boolean gates are classical, mapping Boolean input values to definite Boolean output values.
- A Boolean gate may take some number of inputs (usually one or two), and has one output, which can be freely copied into any number of wires, and thus the number of wires from layer to layer may change. In quantum circuits, quantum gates are operators mapping the state space into itself, and so it always has the same number of outputs as inputs. Thus the number of qubits never changes, and each qubit retains its identity throughout the circuit.
- Boolean gates may lose information from inputs to output, *i.e.*, the input values are not uniquely recoverable from the output value (e.g., and AND gate or an OR gate). Any quantum unitary gate U can always be undone (at least theoretically) by applying U^* immediately

before or afterwards. Thus quantum unitary gates are *reversible*, *i.e.*, the input state is always uniquely recoverable from the output state.

A quantum circuit can use classical gates, provided that they are reversible. Does this pose a significant restriction on the power of quantum circuits to simulate classical computation? Not really. Every classical Boolean circuit can be simulated reversibly. More precisely, we have the following result:

Theorem 11.5 *For every Boolean function* $f : \{0,1\}^n \to \{0,1\}^m$ *with* n *inputs and* m *outputs, there is a reversible circuit* C *(equivalently, a quantum circuit using only classical gates) such that, for all* $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ *and all* $y = (y_1, \ldots, y_m) \in \{0,1\}^m$*, we have,*

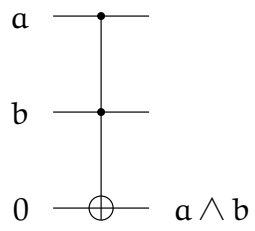


where $(z_1, \ldots, z_m) = f(x)$*. Furthermore,* C *uses only* X *and Toffoli gates, and if* C_f *is a (classical) Boolean circuit computing* f *using binary AND, OR, and unary NOT gates, then a description for* C *can be computed from a description of* C_f *in polynomial time.*

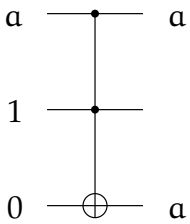
The circuit C acts on three quantum registers: the input qubits, whose initial values are $x_1, \ldots, x_n$; the output qubits (or target qubits), whose initial values are $y_1, \ldots, y_m$, and a set of "work" qubits, called an *ancilla*, whose initial and final value is always $00\cdots0$. When all the ancilla values are restored to 0 at the end of the circuit, we call this a *clean* circuit. The ancilla is used for temporary storage of intermediate results. If the $y_1, \ldots, y_m$ are all 0 initially, then $f(x)$ will appear as the final configuration of the output register. In quantum terms, if the initial state is the basis state $|x\rangle \otimes |y\rangle \otimes |0\cdots0\rangle = |x, y, 0\cdots0\rangle$, then the final state is the basis state $|x\rangle \otimes |y \oplus f(x)\rangle \otimes |0\cdots0\rangle = |x, y \oplus f(x), 0\cdots0\rangle$, where the three labels in the $|\cdot\rangle$ represent the contents of the three quantum registers. We often suppress the ancilla register and say that C takes $|x, y\rangle$ to $|x, y \oplus f(x)\rangle$.

Note that C is clearly reversible. In fact, C is its own inverse. If we feed the output values on the right as input values on the left, then C computes the original inputs as outputs.

We'll only sketch a proof of Theorem 11.5. If C_f is a Boolean circuit computing f, we build C by replacing each gate of C_f with one or more Toffoli gates. We replace NOT gates with Pauli X gates and AND gates with

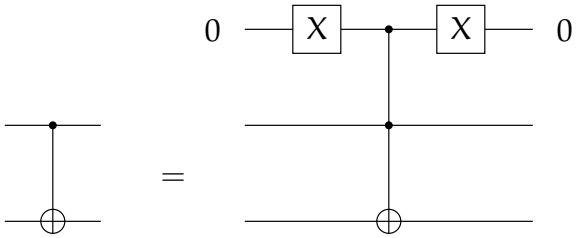


Here we use a fresh ancilla qubit for the second control wire.If we need to copy the Boolean value of a qubit, we can use



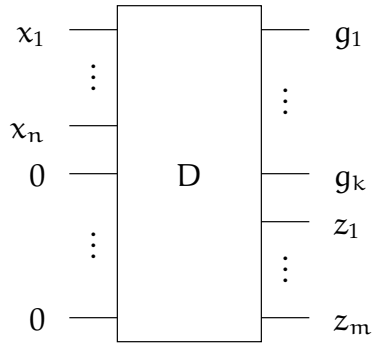
Here, we use a fresh ancilla qubit for the second control wire, and flip it from 0 to 1 with an X gate. To replace an OR gate, we can first express it with AND and NOT gates according to De Morgan's laws, then replace the AND and NOT gates as above.

Notice that the following one-gate circuit cleanly implements the C-NOT gate (the ancilla stays 0):



Thus we can use C-NOT gates in our simulation "for free."

After making all these replacements, we get a circuit that may behave something like this:



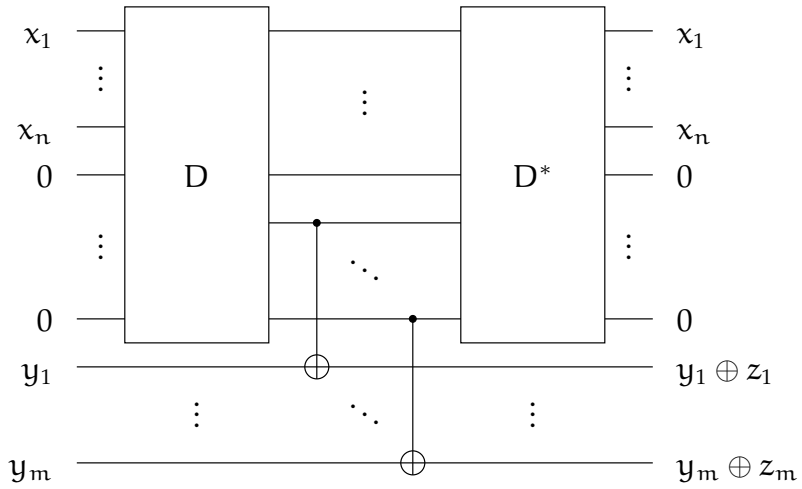
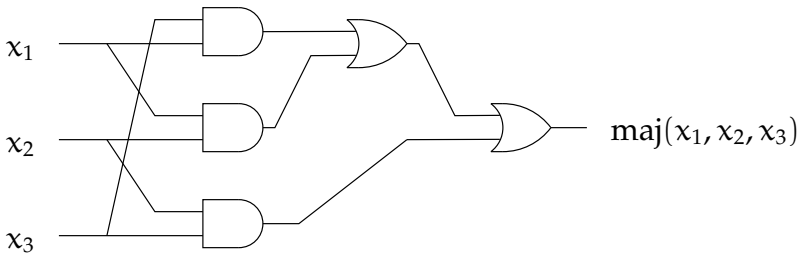


Figure 4: A full implementation of the circuit C. Inputs and ancilla values are restored by undoing the computation after copying the outputs to fresh qubits. The locations of the output register and the ancilla are swapped for ease of display. The circuit implementing D^*, the inverse of D is an exact mirror image of the circuit for D. The values on the qubits intermediate between the D and D^* subcircuits, from top down, are $g_1, \ldots, g_k, z_1, \ldots, z_m$. A C-NOT gate connects each z_i with the qubit carrying y_i. Some additional ancillæ (not shown) are used to implement the C-NOT gates via Toffoli gates.

The intended outputs $z_1, \ldots, z_m$ are somewhere on the right-hand side, and we show them below the other qubits, which contain unused garbage values $g_1, \ldots, g_k$. This circuit, which implements some unitary operator D, is reversible but may not be clean. We have to clean it up. First, we copy the intended outputs onto fresh wires using C-NOT gates, then we *undo* the D computation by applying the exact same gates as in D but in reverse order, taking note that both the Toffoli and X gates are their own inverses. The final circuit is shown in Figure 4.

Exercise 11.6 (Challenging because it's long) The circuit below outputs 1 if and only if at least two of x_1, x_2, x_3 are 1. The three gates in the left column are AND gates; the other two are OR gates.



Convert this circuit into a reversible circuit as in Theorem 11.5, above. Can you make any improvements to the construction?

Why Clean? We'd like to occasionally include one circuit as a subcircuit of another circuit. When we do this, we want to ignore any additional ancilla qubits used by the subcircuit, considering them "local" to the subcircuit, as we did in Figure 4 with the C-NOT gates. If we don't restore the ancilla qubits to their original values, then we can't ignore them as we'd like. Some of the computation will bleed into the unrestored ancilla qubits. This will be especially true with nonclassical quantum circuits.

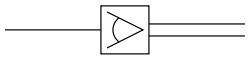
Let C be a circuit with unitary gates that acts on n input and output qubits, using m ancilla qubits. Let $\mathcal{H}$ be the 2^n-dimensional Hilbert space of the input/output qubits, and let $\mathcal{A}$ be the 2^m-dimensional space of the ancilla. Then C is a unitary operator in $\mathcal{L}(\mathcal{H} \otimes \mathcal{A})$. If C is clean, then it restores the ancilla to $|0\cdots0\rangle$, provided the ancilla started that way. Therefore, for every state $|\psi_{in}\rangle \in \mathcal{H}$ there is a unique state $|\psi_{out}\rangle \in \mathcal{H}$ such that $C(|\psi_{in}\rangle \otimes |0\cdots0\rangle) = |\psi_{out}\rangle \otimes |0\cdots0\rangle$. Let $C' : \mathcal{H} \to \mathcal{H}$ be the mapping that takes any $|\psi_{in}\rangle$ to the corresponding $|\psi_{out}\rangle$. C' is clearly a linear operator in $\mathcal{L}(\mathcal{H})$, and further, for any states $|\psi_1\rangle$ and $|\psi_2\rangle$ in $\mathcal{H}$, we have

$$\begin{aligned}
\langle\psi_1|\psi_2\rangle &= \langle\psi_1|\psi_2\rangle\langle 0\cdots0|0\cdots0\rangle \\
&= (\langle\psi_1|\langle 0\cdots0|)(|\psi_1\rangle|0\cdots0\rangle) \\
&= (\langle\psi_1|\langle 0\cdots0|C^*)(C|\psi_1\rangle|0\cdots0\rangle) && \text{(since } C \text{ is unitary)} \\
&= ((\langle\psi_1|C'^*)\langle 0\cdots0|)((C'|\psi_1\rangle)|0\cdots0\rangle) && \text{(by the definition of } C'\text{)} \\
&= \langle\psi_1|C'^*C'|\psi_2\rangle\langle 0\cdots0|0\cdots0\rangle \\
&= \langle\psi_1|C'^*C'|\psi_2\rangle.
\end{aligned}$$

Thus C' preserves the inner product on $\mathcal{H}$ and so must be unitary. This justifies our suppressing the ancilla when we use C as a new unitary "gate" in another circuit. We are really using C', which C implements with its "private" ancilla. We can't do this for a general unitary $C \in \mathcal{L}(\mathcal{H} \otimes \mathcal{A})$.

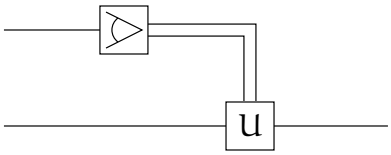
12 Week 6: Measurement gates

Measurement gates. So far, we've only seen unitary gates, reflecting unitary evolution of the qubit or qubits. To get any useful, classical information from a circuit, we must be able to make measurements. At the very least, it is only reasonable that we should be able to measure the (Boolean) value of a qubit, that is, we should be able to make a projective measurement $\{P_0, P_1\}$ of any qubit with respect to the computational basis. We represent such a measurement by the one-qubit gate



(For those of you failing to appreciate the artistry of my iconography, the gate depicts an eye in profile. In the book, this gate is depicted as the display of a gauge with a needle.) The incoming qubit is measured projectively in the computational basis, and the classical result (a single bit) is carried on the double wire to the right. If there are other qubits present in the system, then the projective measurement is really $\{P_0 \otimes I, P_1 \otimes I\}$, where I is the identity operator applying to the qubits not being measured (recall Exercise 10.4).

There are two uses for a qubit measurement. The first, obvious use is to read the answer from the final state of a computation. The second is to control future operations in the circuit by using the result of an intermediate measurement. For example, the result of a measurement may be used to control another gate:



The U gate is applied to the second qubit if and only if the result of measuring the first qubit is 1. Unlike a qubit, a classical bit can be duplicated freely and used to control many gates later in the circuit.

Exercise 12.1 A general three-qubit state can be written as

$$|\psi\rangle = \sum_{x \in \{0,1\}^3} \alpha_x |x\rangle,$$

where $\sum_x |\alpha_x|^2 = 1$. For each $i = 1, 2, 3$, give an expression for the probability of seeing 1 when the ith qubit is measured, and give the post-measurement state in each case.

Based on the discussion after Exercise 10.4, we may measure several different qubits at once, since the actual chronological order of the measurements does not matter. Here's a completely typical example: we decide to measure qubits 2, 3, and 5 of a n-qubit system (where $n \geqslant 5$, obviously). The state $|\psi\rangle$ of an n-qubit system can always be expressed as a linear combination of basis states:

$$|\psi\rangle = \sum_{x \in \{0,1\}^n} \alpha_x |x\rangle, \tag{47}$$

where each α_x is a scalar in $\mathbb{C}$, and

$$\sum_{x \in \{0,1\}^n} |\alpha_x|^2 = \langle\psi|\psi\rangle = 1. \tag{48}$$

If we measure qubits 2,3, and 5 when the system is in state $|\psi\rangle$, what is the probability that we will see, say, 101, *i.e.*, 1 for qubit 2, 0 for qubit 3, and 1 for qubit 5? The corresponding projector is $P = I \otimes P_1 \otimes P_0 \otimes I \otimes P_1 \otimes I \otimes I$, where I is the single-qubit identity operator. The probability is then

$$\Pr[101] = \langle\psi|P|\psi\rangle = \sum_{x \,:\, x_2 x_3 x_5 = 101} |\alpha_x|^2,$$

where we are letting x_j denote the jth bit of x. That is, we only retain those terms in the sum in (48) in which the corresponding bits of x match the outcome. Upon seeing 101, the post-measurement state will be

$$|\psi_{\text{post}}\rangle = \frac{P|\psi\rangle}{\Pr[101]} = \frac{1}{\Pr[101]} \sum_{x \,:\, x_2 x_3 x_5 = 101} \alpha_x |x\rangle.$$

We will often measure several qubits at once, so this example will come in handy.

Bell States and Quantum Teleportation. Recall the circuit of Figure 3. Let B be the two-qubit unitary operator realized by this circuit. The four states obtained by applying B to the four computational basis states are known as the *Bell states* and form the *Bell basis*:

$$\left|\Phi^+\right\rangle := B|00\rangle = (|00\rangle + |11\rangle)/\sqrt{2}, \tag{49}$$

$$\left|\Psi^+\right\rangle := B|01\rangle = (|01\rangle + |10\rangle)/\sqrt{2}, \tag{50}$$

$$\left|\Phi^-\right\rangle := B|10\rangle = (|00\rangle - |11\rangle)/\sqrt{2}, \tag{51}$$

$$\left|\Psi^-\right\rangle := B|11\rangle = (|01\rangle - |10\rangle)/\sqrt{2}. \tag{52}$$

These states are also called *EPR states* or *EPR pairs*. In a sense we will quantify later, these states represent maximally entangle pairs of qubits. EPR is an acronym for Einstein, Podolsky, and Rosen, who coauthored a paper describing apparent paradoxes in the rules of quantum mechanics involving pairs of qubits in states such as these. Suppose a pair of electrons is prepared whose spins are in one of the Bell states, say $|\Phi^+\rangle$. (There are actual physical processes that can do this.) The electrons can then (theoretically) be separated by a great distance—the first taken by Alice to a lab at UC Berkeley in California and the second taken by Bob to a lab at MIT in Massachusetts. If Alice measures her spin first, she'll see 0 or 1 with equal probability. Same with Bob if he measures his spin first. But if Alice measures her spin first and sees, say, 0, then according to the standard Copenhagen interpretation of quantum mechanics (which we are using), the state of the two spins collapses to $|00\rangle$, so if Bob measures his spin afterwards, he will see 0 with certainty. So Alice's measurement seems to affect Bob's somehow. Einstein called this phenomenon "spooky action at a distance." We'll talk about this more later, time permitting.

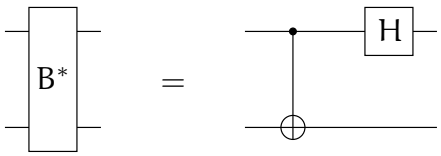
Philosophical problems aside, entangled pairs of qubits can be used in interesting and subtle ways. One of the earliest discovered uses of EPR pairs is to teleport an unknown quantum state across a distance using only *classical* communication, in a process called *quantum teleportation*. Suppose Alice and Bob share two qubits in the state $|\Phi^+\rangle$ as above, which may have been distributed to them long ago. Suppose also that Alice has another qubit in some arbitrary, unknown state

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle.$$

She wants Bob to have this state. She could mail her electron to Bob, but this won't work because the state $|\psi\rangle$ of the electron is very delicate and will be destroyed if the package is bumped, screened with X-rays, etc. Instead, she can transfer this state to Bob with only a phone call. No quantum states need to be physically transported between Alice and Bob. Here's how it works: The state of the three qubits initially is

$$|\psi\rangle\left|\Phi^+\right\rangle = (\alpha|0\rangle + \beta|1\rangle)(|00\rangle + |11\rangle)/\sqrt{2} = (\alpha|000\rangle + \alpha|011\rangle + \beta|100\rangle + \beta|111\rangle)/\sqrt{2}. \tag{53}$$

Alice possesses the first two qubits; Bob possesses the third. Alice applies the inverse B^* of the circuit of Figure 3:



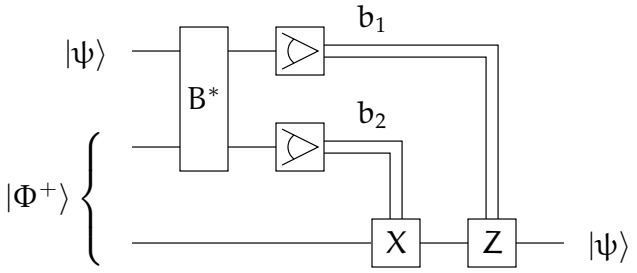


Figure 5: Quantum teleportation of a single qubit. Alice possesses the first qubit in some arbitrary, unknown state $|\psi\rangle$. The second and third qubits are an EPR pair prepared in the state $|\Phi^+\rangle$ sometime in the past, with the second qubit given to Alice and the third to Bob. Alice applies B^* to her two qubits, then measures both qubits, then communicates the results $b_1, b_2 \in \{0, 1\}$ of the measurements to Bob. Bob uses this information to decide whether to apply Pauli X and Z gates to his qubit.

to her two qubits. She then measures each qubit in the computational basis, getting Boolean values b_1 and b_2 for the first and second qubits, respectively. She then calls Bob on the phone and tells him the values she observed, *i.e.*, b_1 and b_2. Bob then does the following with his qubit (the third qubit): (i) if $b_2 = 1$, then Bob applies an X gate, otherwise he does nothing; then (ii) if $b_1 = 1$, then he applies a Z gate, otherwise he does nothing. At this point, Bob's qubit will be in state $|\psi\rangle$. We can illustrate the process by the circuit in Figure 5. Let's check that Bob actually does wind up with $|\psi\rangle$. It will make our work easier to first express the initial state of (53) using the Bell basis. It's easy to check that

$$\begin{aligned}
|00\rangle &= \left(\left|\Phi^+\right\rangle + \left|\Phi^-\right\rangle\right)/\sqrt{2}, \\
|01\rangle &= \left(\left|\Psi^+\right\rangle + \left|\Psi^-\right\rangle\right)/\sqrt{2}, \\
|10\rangle &= \left(\left|\Psi^+\right\rangle - \left|\Psi^-\right\rangle\right)/\sqrt{2}, \\
|11\rangle &= \left(\left|\Phi^+\right\rangle - \left|\Phi^-\right\rangle\right)/\sqrt{2},
\end{aligned}$$

so the initial state of (53) is

$$\begin{aligned}
&\frac{1}{2}\left[\alpha\left(\left|\Phi^+\right\rangle + \left|\Phi^-\right\rangle\right)|0\rangle + \alpha\left(\left|\Psi^+\right\rangle + \left|\Psi^-\right\rangle\right)|1\rangle + \beta\left(\left|\Psi^+\right\rangle - \left|\Psi^-\right\rangle\right)|0\rangle + \beta\left(\left|\Phi^+\right\rangle - \left|\Phi^-\right\rangle\right)|1\rangle\right] \\
&\quad = \frac{1}{2}\left[\left|\Phi^+\right\rangle(\alpha|0\rangle + \beta|1\rangle) + \left|\Psi^+\right\rangle(\alpha|1\rangle + \beta|0\rangle) + \left|\Phi^-\right\rangle(\alpha|0\rangle - \beta|1\rangle) + \left|\Psi^-\right\rangle(\alpha|1\rangle - \beta|0\rangle)\right].
\end{aligned}$$

Going back to Equations (49–52) and applying B^* to both sides, we see that B^* maps $|\Phi^+\rangle$ to $|00\rangle$ and so on. So after Alice applies B^* to her two qubits, the state becomes

$$\frac{1}{2}\left[|00\rangle(\alpha|0\rangle + \beta|1\rangle) + |01\rangle(\alpha|1\rangle + \beta|0\rangle) + |10\rangle(\alpha|0\rangle - \beta|1\rangle) + |11\rangle(\alpha|1\rangle - \beta|0\rangle)\right]. \tag{54}$$

Now Alice measures her two qubits. She'll get one of four possible values: 00, 01, 10, 11, all with probability 1/4. For $b_1, b_2 \in \{0, 1\}$, let $|\psi_{b_1 b_2}\rangle$ be the state of the three qubits after the measurement,

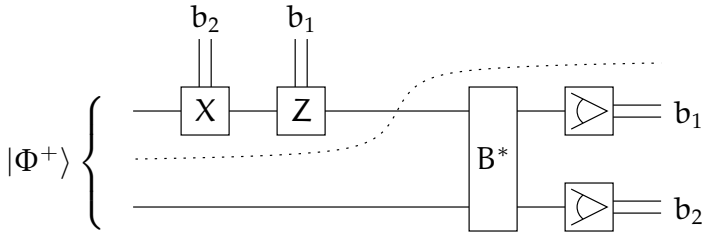


Figure 6: Dense coding. The EPR pair is initially distributed between Alice and Bob, with Alice getting the first qubit. The stuff above the dotted line belongs to Alice, and the rest belongs to Bob. The qubit crosses the dotted line when Alice sends it to Bob.

assuming the result is b_1, b_2. By applying the corresponding projectors to the state in (54) and normalizing, we get

$$\begin{aligned}
|\psi_{00}\rangle &= |00\rangle(\alpha|0\rangle + \beta|1\rangle) = |00\rangle|\psi\rangle, \\
|\psi_{01}\rangle &= |01\rangle(\alpha|1\rangle + \beta|0\rangle) = |01\rangle(X|\psi\rangle), \\
|\psi_{10}\rangle &= |10\rangle(\alpha|0\rangle - \beta|1\rangle) = |10\rangle(Z|\psi\rangle), \\
|\psi_{11}\rangle &= |11\rangle(\alpha|1\rangle - \beta|0\rangle) = |11\rangle(XZ|\psi\rangle).
\end{aligned}$$

We see that Bob's qubit is now in one of four possible states: $|\psi\rangle$, $X|\psi\rangle$, $Z|\psi\rangle$, or $XZ|\psi\rangle$, depending on whether the values measured by Alice are 00, 01, 10, or 11, respectively. Now Bob simply uses the information about b_1 and b_2 to undo the Pauli operators on his qubit, yielding $|\psi\rangle$ in every case.

This scenario can be used to teleport an n-qubit state from Alice to Bob by teleporting each qubit separately, just as above.

Note that Alice must tell Bob the values b_1 and b_2 so that Bob can recover $|\psi\rangle$ reliably. This means that quantum states cannot be teleported faster than the speed of light. Also note that after the protocol is finished, Alice no longer possesses $|\psi\rangle$. She can't, because that would violate the No-Cloning Theorem. Finally, note that the EPR state that Alice and Bob shared before the protocol no longer exists. It is used up, and can't be used to teleport additional states. Thus, teleporting an n-qubit state needs n separate EPR pairs.

Dense Coding. In quantum teleportation, with the help of an EPR pair, Alice can substitute transmitting a qubit to Bob with transmitting two classical bits. There is a converse to this: with the help of an EPR pair, Alice can substitute transmitting two classical bits to Bob with transmitting a single qubit. This inverse trade-off is known as *dense coding*.

Figure 6 illustrates how dense coding works. Alice has two classical bits b_1 and b_2 that she wants to communicate to Bob. She also shares an EPR pair with Bob in state $|\Phi^+\rangle$ as before. If $b_2 = 1$, Alice applies X to her half of the EPR pair, otherwise she does nothing. Then, if $b_1 = 1$, she applies Z to her qubit, otherwise she does nothing. She then sends her qubit to Bob. Bob now has both qubits. He applies B^* to them then measures each of his qubits, seeing b_1 and b_2 as outcomes with certainty.

Here are the four possible states of the two qubits when Alice sends her qubit to Bob, corresponding to the four possible values of $b_1 b_2$ (here, I is the one-qubit identity operator):

$$\begin{aligned}
|\psi_{00}\rangle &= (I \otimes I)\left|\Phi^+\right\rangle = \left|\Phi^+\right\rangle, \\
|\psi_{01}\rangle &= (X \otimes I)\left|\Phi^+\right\rangle = (|10\rangle + |01\rangle)/\sqrt{2} = \left|\Psi^+\right\rangle, \\
|\psi_{10}\rangle &= (Z \otimes I)\left|\Phi^+\right\rangle = (|00\rangle - |11\rangle)/\sqrt{2} = \left|\Phi^-\right\rangle, \\
|\psi_{11}\rangle &= (ZX \otimes I)\left|\Phi^+\right\rangle = (|01\rangle - |10\rangle)/\sqrt{2} = \left|\Psi^-\right\rangle.
\end{aligned}$$

So Alice is just preparing one of the four Bell states. So when Bob applies B^* to $|\psi_{b_1 b_2}\rangle$, he gets $|b_1 b_2\rangle$, yielding $b_1 b_2$ upon measurement.

Note that, as before, the EPR pair is consumed in the process.

Exercise 12.2 Recall the two-qubit swap operator SWAP satisfying $\text{SWAP}|a\rangle|b\rangle = |b\rangle|a\rangle$ for all $a, b \in \{0, 1\}$. Show that the four Bell states are eigenvectors of SWAP. What are the corresponding eigenvalues? For this and other reasons, the states $|\Phi^+\rangle$, $|\Phi^-\rangle$, and $|\Psi^+\rangle$ are often called *symmetric states*, *triplet states*, or *spin-1 states*, while the state $|\Psi^-\rangle$ is often called the *antisymmetric state*, the *singlet state*, or the *spin-0 state*.

13 Week 7: Basic quantum algorithms

Black-Box Problems. Many quantum algorithms solve what are called "black-box" problems. Typically, we are given some Boolean function $f : \{0,1\}^n \to \{0,1\}^m$ and we want to answer some question about the function as a whole, for example, "Is f constant?", "Is f the zero function?", "Is f one-to-one?", etc. We are allowed to feed an input $x \in \{0,1\}^n$ to f and get back the output $f(x)$. The input x is called a *query* to f and $f(x)$ is the *query answer*. Other than making queries to f, we are not allowed to inspect f in any way, hence the black-box nature of the function. (A black-box function f is sometimes called an *oracle*.) Generally, we would like to answer our question by making as few queries to f as we can, since queries may be expensive.

In the context of quantum computing, the function f is most naturally given to us as a classical, unitary gate U_f that acts on two quantum registers—the first with n qubits and the second with m qubits—and behaves as follows for all $x \in \{0,1\}^n$ and $y \in \{0,1\}^m$:

$$U_f|x, y\rangle = |x, y \oplus f(x)\rangle.$$

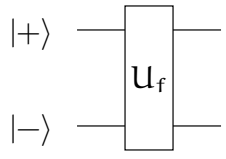
This is reasonable, given the restriction that unitary quantum gates must be reversible. U_f is called an f-*gate*. To solve a black-box problem involving f, we are allowed to build a quantum circuit using f-gates—as well as the other usual unitary gates. Each occurrence of an f-gate in the circuit counts as a query to f, so the number of queries is the number of f-gates in the circuit. The difference between classical queries to f and quantum queries to f is that we can feed a *superposition* of several classical inputs (basis states) into the f-gate, obtaining a corresponding superposition of the results. We'll see in a minute that we can use this idea, known as *quantum parallelism* to get more information out of f in fewer queries than any classical computation.

Deutsch's Problem and the Deutsch-Jozsa Problem. The first indication that quantum computation may be strictly more powerful than classical computation came with a black-box problem posed by David Deutsch: Given a one-bit Boolean function $f : \{0,1\} \to \{0,1\}$, is f constant, that is, is $f(0) = f(1)$? There are four possible functions $\{0,1\} \to \{0,1\}$: the constant zero function, the constant one function, the identity function, and the negation function. Deutsch's task is to determine whether f falls among the first two or the last two. Classically, it is clear that determining which is the case requires two queries to f, since we need to know both $f(0)$ and $f(1)$. Quantally, however, we can get by with only one query to f. Define

$$|+\rangle := H|0\rangle = (|0\rangle + |1\rangle)/\sqrt{2}, \tag{55}$$

$$|-\rangle := H|1\rangle = (|0\rangle - |1\rangle)/\sqrt{2}, \tag{56}$$

where H is the Hadamard gate. The states $|+\rangle$ and $|-\rangle$ correspond to the states $|+x\rangle$ and $|-x\rangle$ we defined earlier when we were discussing the Bloch sphere. If we feed these states into U_f like so:



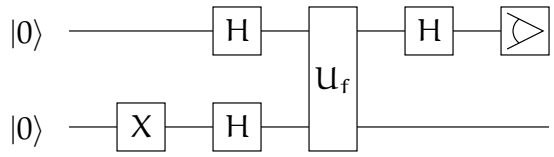


Figure 7: The full circuit for Deutsch's problem. The second qubit is not used after it emerges from the f-gate.

then the progression of states through the circuit from left to right is

$$\begin{aligned}
|+\rangle|-\rangle &= (|0\rangle + |1\rangle)(|0\rangle - |1\rangle)/2 \\
&= (|00\rangle - |01\rangle + |10\rangle - |11\rangle)/2 \\
&\overset{U_f}{\mapsto} (|0, f(0)\rangle - |0, 1 \oplus f(0)\rangle + |1, f(1)\rangle - |1, 1 \oplus f(1)\rangle)/2 \\
&=: |\psi_{out}\rangle.
\end{aligned}$$

If f is constant, *i.e.*, if $f(0) = f(1) = y$ for some $y \in \{0, 1\}$, then

$$\begin{aligned}
|\psi_{out}\rangle &= (|0, y\rangle - |0, 1 \oplus y\rangle + |1, y\rangle - |1, 1 \oplus y\rangle)/2 \\
&= (|0\rangle + |1\rangle)(|y\rangle - |1 \oplus y\rangle)/2 \\
&= (-1)^y(|0\rangle + |1\rangle)(|0\rangle - |1\rangle)/2 \\
&= (-1)^y|+\rangle|-\rangle.
\end{aligned}$$

If f is not constant, *i.e.*, if $f(0) = y = 1 \oplus f(1)$ for some $y \in \{0, 1\}$, then

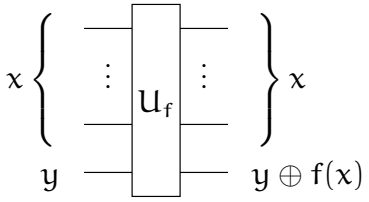
$$\begin{aligned}
|\psi_{out}\rangle &= (|0, y\rangle - |0, 1 \oplus y\rangle + |1, 1 \oplus y\rangle - |1, y\rangle)/2 \\
&= (|0\rangle - |1\rangle)(|y\rangle - |1 \oplus y\rangle)/2 \\
&= (-1)^y(|0\rangle - |1\rangle)(|0\rangle - |1\rangle)/2 \\
&= (-1)^y|-\rangle|-\rangle.
\end{aligned}$$

Now suppose we apply the Hadamard gate H to the first qubit of $|\psi_{out}\rangle$. We obtain

$$|\phi\rangle := (H \otimes I)|\psi_{out}\rangle = \begin{cases} \pm|0\rangle|-\rangle & \text{if f is contant,} \\ \pm|1\rangle|-\rangle & \text{if f is not constant.} \end{cases}$$

So now we measure the first qubit of $|\phi\rangle$. We get 0 with certainty if f is constant, and we get 1 with certainty otherwise. We can prepare the initial state $|+\rangle|-\rangle$ by applying two Hadamards and a Pauli X gate. The full circuit is in Figure 7. We only use the f-gate once, but in superposition. That is the key point.

Deutsch and Jozsa generalized this idea to a function $f : \{0, 1\}^n \to \{0, 1\}$ with n inputs and one output. The corresponding $(n + 1)$-qubit U_f gate looks like

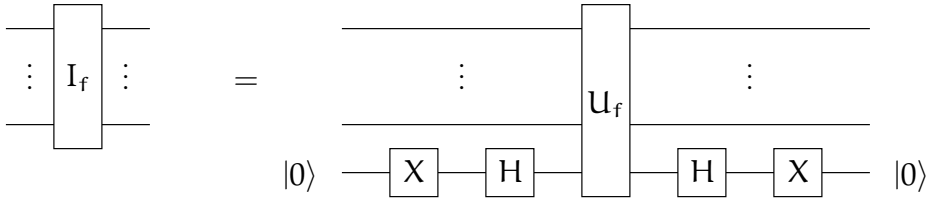


We say that f is *balanced* if the number of inputs x such that $f(x) = 0$ is equal to the number of inputs x such that $f(x) = 1$, namely, 2^{n-1}. The Deutsch-Jozsa problem is as follows: We are given f as above as a black-box gate, and we know (we are promised) that f is either constant or balanced, and we want to determine which is the case. Answering this question classically requires $2^{n-1}+1$ queries to f in the worst case, since it is possible that f is balanced but the first 2^{n-1} queries may all yield the same answer. Quantally, we can do *much* better; one query to f suffices.

The set-up is similar to what we just did, but instead of using an $(n+1)$-qubit f-gate directly, it is easier to work with an n-qubit *inversion f-gate* I_f defined as follows for every $x \in \{0,1\}^n$:

$$I_f|x\rangle = (-1)^{f(x)}|x\rangle.$$

That is, I_f leaves the values of the qubits alone but flips the sign iff $f(x) = 1$. We've defined I_f on computational basis vectors. Since I_f is linear, this defines I_f on all vectors in the state space of n qubits. I_f can be implemented cleanly (and easily) using U_f thus:

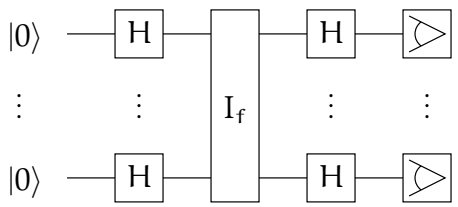


For any input state $|x\rangle$ where $x \in \{0,1\}^n$, the progression of states through the circuit from left to right is

$$\begin{aligned}
|x,0\rangle &\overset{X}{\mapsto} |x,1\rangle \\
&\overset{H}{\mapsto} |x\rangle|-\rangle \\
&\overset{U_f}{\mapsto} |x\rangle(|f(x)\rangle - |1 \oplus f(x)\rangle)\sqrt{2} \\
&= (-1)^{f(x)}|x\rangle(|0\rangle - |1\rangle)/\sqrt{2} \\
&= (-1)^{f(x)}|x\rangle|-\rangle \\
&\overset{H}{\mapsto} (-1)^{f(x)}|x\rangle|1\rangle \\
&\overset{X}{\mapsto} (-1)^{f(x)}|x,0\rangle
\end{aligned}$$

as advertized. Since only one f-gate is used to implement I_f, each occurrence of I_f in a circuit amounts to one occurrence of U_f in the circuit.

To determine whether f is constant or balanced, we use the following n-qubit circuit:



The dots indicate that all n qubits start in state $|0\rangle$, a Hadamard gate is applied to each qubit before and after I_f, and all qubits are measured at the end. Before we view the progression of states, let's see what happens when we apply a column of n Hadamard gates all at once to n qubits in the state $|x\rangle$, for any $x = x_1 x_2 \cdots x_n \in \{0,1\}^n$. (We denote the n-fold Hadamard operator as $H^{\otimes n}$.) Noting that, for all $b \in \{0,1\}$,

$$H|b\rangle = \frac{1}{\sqrt{2}}(|0\rangle + (-1)^b|1\rangle) = \frac{1}{\sqrt{2}} \sum_{c \in \{0,1\}} (-1)^{bc}|c\rangle,$$

we get

$$\begin{aligned}
|x\rangle \quad &\overset{H^{\otimes n}}{\mapsto} \quad \bigotimes_{i=1}^{n} H|x_i\rangle \\
&= \quad \frac{1}{2^{n/2}} \bigotimes_{i=1}^{n} \sum_{y_i \in \{0,1\}} (-1)^{x_i y_i}|y_i\rangle \\
&= \quad \frac{1}{2^{n/2}} \sum_{y_1 \in \{0,1\}} \cdots \sum_{y_n \in \{0,1\}} (-1)^{x_1 y_1 + \cdots + x_n y_n}|y_1\rangle \otimes \cdots \otimes |y_n\rangle \\
&= \quad \frac{1}{2^{n/2}} \sum_{y \in \{0,1\}^n} (-1)^{x \cdot y}|y\rangle,
\end{aligned}$$

where $x \cdot y = x_1 y_1 + \cdots + x_n y_n$ denotes the standard dot product of two n-bit vectors $x = x_1 \cdots x_n$ and $y = y_1 \cdots y_n$.

Now let's view the progression of states of the circuit above.

$$|00\cdots0\rangle \quad \overset{H^{\otimes n}}{\mapsto} \quad \frac{1}{2^{n/2}} \sum_{x \in \{0,1\}^n} |x\rangle \qquad \text{(because } (00\cdots0) \cdot x = 0) \tag{57}$$

$$\overset{I_f}{\mapsto} \quad \frac{1}{2^{n/2}} \sum_{x \in \{0,1\}^n} (-1)^{f(x)}|x\rangle \tag{58}$$

$$\overset{H^{\otimes n}}{\mapsto} \quad \frac{1}{2^n} \sum_{x \in \{0,1\}^n} (-1)^{f(x)} \sum_{y \in \{0,1\}^n} (-1)^{x \cdot y}|y\rangle \tag{59}$$

$$= \quad \frac{1}{2^n} \sum_{x,y \in \{0,1\}^n} (-1)^{f(x)+x \cdot y}|y\rangle \tag{60}$$

$$= \quad \frac{1}{2^n} \sum_{y \in \{0,1\}^n} \left(\sum_{x \in \{0,1\}^n} (-1)^{f(x)+x \cdot y} \right) |y\rangle \tag{61}$$

Suppose first that f is constant, and we let $|\psi_{\text{const}}\rangle$ denote this last state. Then $(-1)^{f(x)} = \pm 1$ independent of x, and so we can bring it out side the sum:

$$|\psi_{\text{const}}\rangle \quad = \quad \pm\frac{1}{2^n} \sum_{y \in \{0,1\}^n} \left(\sum_{x \in \{0,1\}^n} (-1)^{x \cdot y} \right) |y\rangle$$

$$= \pm\frac{1}{2^n}\left(\sum_x (-1)^0\right)|0^n\rangle \pm \frac{1}{2^n}\sum_{y\neq 0^n}\left(\sum_x (-1)^{x\cdot y}\right)|y\rangle$$
$$= \pm|0^n\rangle \pm \frac{1}{2^n}\sum_{y\neq 0^n}\left(\sum_x (-1)^{x\cdot y}\right)|y\rangle.$$

Since

$$1 = \langle\psi_{\text{const}}|\psi_{\text{const}}\rangle = 1 + \frac{1}{2^{2n}}\sum_{y\neq 0^n}\left|\sum_x (-1)^{x\cdot y}\right|^2,$$

we must have $\sum_x (-1)^{x\cdot y} = 0$ for all $y \neq 0^n$,[14] and thus

$$|\psi_{\text{const}}\rangle = \pm|0^n\rangle.$$

When we measure the qubits in state $|\psi_{\text{const}}\rangle$, we will see 0^n with certainty.

Now suppose that f is balanced, and we let $|\psi_{\text{bal}}\rangle$ denote the state of (61). Again separating the $|0^n\rangle$-term from the rest, we get

$$|\psi_{\text{bal}}\rangle = \frac{1}{2^n}\left(\sum_{x\in\{0,1\}^n}(-1)^{f(x)}\right)|0^n\rangle + \frac{1}{2^n}\sum_{y\neq 0^n}\left(\sum_{x\in\{0,1\}^n}(-1)^{f(x)+x\cdot y}\right)|y\rangle.$$

But f is balanced, and so $\sum_x (-1)^{f(x)} = 0$ because each term contributes $+1$ for $f(x) = 0$ and -1 for $f(x) = 1$. Thus,

$$|\psi_{\text{bal}}\rangle = \frac{1}{2^n}\sum_{y\neq 0^n}\left(\sum_{x\in\{0,1\}^n}(-1)^{f(x)+x\cdot y}\right)|y\rangle.$$

When we measure the qubits in state $|\psi_{\text{bal}}\rangle$, we see 0^n with probability *zero*. So we never see 0^n, but instead we'll see some random $y \neq 0^n$.

To summarize: when we measure the qubits, if we see 0^n, then we know that f is constant; if we see anything else, then we know that f is balanced.

Exercise 13.1 (Challenging) Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. We've implemented an I_f gate using U_f and a few standard gates. Show how to implement U_f given a single I_f gate and some standard gates. Thus U_f and I_f are computationally equivalent. Your circuit is allowed to depend on the value of $f(0^n)$, that is, you can have one circuit that works assuming $f(0^n) = 0$ and another (slightly different) circuit that works assuming $f(0^n) = 1$. [Hint: Build a quantum circuit with three registers: n input qubits; one output qubit; n ancilla qubits. Assume $x \in \{0,1\}^n$ is the classical input. Using one Hadamard and n Toffoli gates, convert the input state $|x, y, 0^n\rangle$ into the superposition

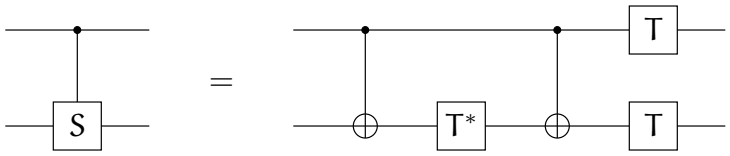
$$\frac{|x, 0, 0^n\rangle + (-1)^y|x, 1, x\rangle}{\sqrt{2}}.$$

Then feed the ancilla register into I_f, then undo what you did before applying I_f. What state do you wind up with? What else do you need to do, if anything?]

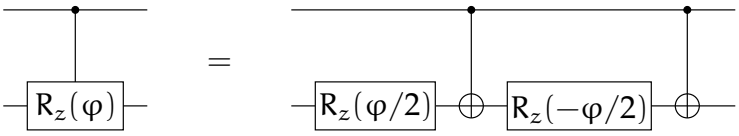
[14] Here's another way to see that $\sum_x (-1)^{x\cdot y} = 0$ for all $y \neq 0^n$: If $y \neq 0^n$, then one of y's bits is 1. For convenience, let's assume that the first bit of y is 1, and we let y' be the rest of y. Then $\sum_x (-1)^{x\cdot y} = \sum_{x_1\in\{0,1\}}\sum_{x'\in\{0,1\}^{n-1}}(-1)^{x_1 x'\cdot 1 y'} = \sum_{x_1}(-1)^{x_1}\sum_{x'}(-1)^{x'\cdot y'} = \sum_{x'}(-1)^{x'\cdot y'} - \sum_{x'}(-1)^{x'\cdot y'} = 0$.

Exercise 13.2 Here are some more circuit equalities for you to verify. Remember that circuits represent linear operators, and thus to show that two circuits are equal, it suffices to show that they act the same on the vectors of some basis, *e.g.*, the computational basis.

1. Check that

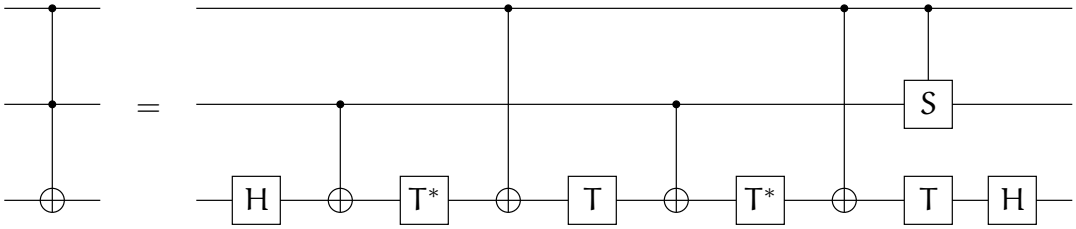


2. In a similar vein, for any $\varphi \in \mathbb{R}$ show that

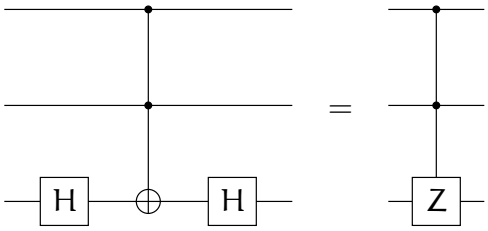


   [Hint: If the control qubit on the right-hand side is 1, then $XR_z(-\varphi/2)XR_z(\varphi/2)$ is applied to the target qubit. Note that $XR_z(-\varphi/2)X = Xe^{i\varphi Z/4}X = e^{i\varphi XZX/4} = e^{-i\varphi Z/4} = R_z(\varphi/2)$. The second equation follows from Exercise 9.3(7).]

3. (Challenging but recommended) Show that



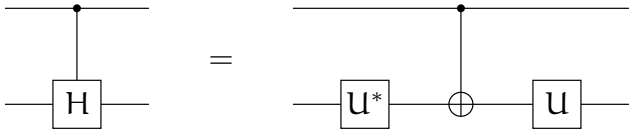
   Combining this with item 1 above gives a circuit implementing the Toffoli gate using only C-NOT, H, T, and T^* gates (and we could do without T^* explicitly by using T^7 instead, because $T^8 = I$). The Nielsen & Chuang textbook has a closely similar implementation of the Toffoli gate on page 182, but it's not optimal; it has one more gate than is necessary. [Hint: It will help first to transform this equation into an equivalent one by applying H gates on the third qubit to both sides of both circuits, *i.e.*, unitarily conjugating both sides of the equation by $I \otimes I \otimes H$. This has the effect of canceling out both the H gates on the right-hand circuit, and the left-hand side becomes



which flips the overall sign of the state (*i.e.*, gives an $e^{i\pi} = -1$ phase change) iff all three qubits are 1. The advantage of doing this is that now nothing in the right-hand circuit creates any superpositions; each gate maps a computational basis state to a computational basis state, up to a phase factor. Now proceed by cases, considering the possible 0, 1-combinations of the values of the three qubits, adding up the overall phase angles generated. You can simplify the task further by noticing a few general facts:

- A 0 on the control qubit of a C-NOT gate eliminates the gate.
- Adjacent T and T^* gates on the same qubit cancel.
- Adjacent C-NOT gates with the same control and target qubits cancel.]

Exercise 13.3 (Challenging) This exercise is a puzzler that is best solved by finding the right series of rotations of the Bloch sphere. Find a single-qubit unitary U such that



Furthermore, you are restricted to expressing U as the product of a sequence of operators, all of which are either H or T. [Hint: You are trying to find a U such that $UXU^* = H$. X gives a π-rotation of the Bloch sphere about the x-axis, and H gives a π-rotation about the line ℓ through the point in the x, z-plane halfway between the $+x$- and $+z$-axes, with spherical coordinates $(\pi/4, 0)$ (Cartesian coordinates $(1/\sqrt{2}, 0, 1/\sqrt{2})$). So U must necessarily give a rotation that moves the x-axis to ℓ, so that U^* (applied first) moves ℓ to the x-axis, then X (applied second) rotates π around the x-axis, then U (applied last) moves the x-axis back to ℓ, the net effect of all three being a π-rotation about ℓ. One possibility for U is a $(-\pi/4)$-rotation about the y-axis, but you must implement this using just H and T, the latter giving a $\pi/4$-rotation about the z-axis.]

14 Week 7: Simon's problem

Simon's Problem. The Deutsch-Jozsa problem is hard to decide classically, requiring exponentially many (in n) queries to f. But there is a sense in which this problem is easy classically: if we pick inputs to f *at random* and query f on those inputs, we quickly learn the right answer with high probability. If we ever see f output different values, then we know for certain that f is balanced, since it is nonconstant. Conversely, if f is balanced and we make 100 random queries to f, then the chances that f gives the same answer to all our queries is exceedingly small—2^{-99}. So we have an efficient randomized algorithm for finding the answer: Make m uniformly and independently random queries to f, where m is, say, 100. If the answers are all the same, output "constant"; otherwise, output "balanced." We will never output "balanced" incorrectly. We might output "constant" incorrectly, but only with probability 2^{1-m}, *i.e.*, exponentially small in m. This algorithm runs in time polynomial in n and m.

As with classical computation, quantum circuits can simulate classical randomized computation. We won't pursue that line further here, though. Instead, we'll now see a black-box problem—Simon's problem—that

- can be solved efficently with high probability on a quantum computer, but
- cannot be solved efficiently by a classical computer, even by a randomized algorithm that is allowed a probability of error slightly below 1/2.

In Simon's problem, we are given a black-box Boolean function $f : \{0,1\}^n \to \{0,1\}^m$, for some $n \leqslant m$. We are also given the promise that there is an $s \in \{0,1\}^n$ such that for all distinct $x, y \in \{0,1\}^n$,

$$f(x) = f(y) \iff x \oplus y = s.$$

This condition determines s uniquely: either $s = 0^n$ and f is one-to-one, or $s \neq 0^n$ in which case f is two-to-one with $f(x) = f(x \oplus s)$ for all x, and s is the unique nonzero input such that $f(s) = f(0)$. Our task is to find s.

The function f is given to us via the gate U_f as before, such that $U_f|x,y\rangle = |x, y \oplus f(x)\rangle$ for all $x \in \{0,1\}^n$ and $y \in \{0,1\}^m$. Consider the following quantum algorithm with two quantum registers—an n-qubit input register and an m-qubit output register.

1. We start with the two registers in the all-zero state $|0^n, 0^m\rangle$.
2. We then apply $H^{\otimes n}$ to the input register, obtaining the state $2^{-n/2} \sum_{x \in \{0,1\}^n} |x, 0^m\rangle$.
3. We then apply U_f to get the new state $2^{-n/2} \sum_{x \in \{0,1\}^n} |x, f(x)\rangle$.
4. We apply $H^{\otimes n}$ to the first register again to get the state
$$|\psi_{\text{out}}\rangle = 2^{-n} \sum_{x,y \in \{0,1\}^n} (-1)^{x \cdot y} |y, f(x)\rangle.$$
5. We now measure the first register (all n qubits), obtaining some value $y \in \{0,1\}^n$.

Exercise 14.1 Draw the quantum circuit implementing the algorithm above.

What z do we get in the last step? Note that $f(x) = f(x \oplus s)$ for all x, and that as x ranges through all of $\{0,1\}^n$, so does $x \oplus s$. Thus we can rewrite $|\psi_{\text{out}}\rangle$ as a split sum and combine terms in pairs:

$$\begin{aligned}
|\psi_{\text{out}}\rangle &= \frac{1}{2}\left(|\psi_{\text{out}}\rangle + |\psi_{\text{out}}\rangle\right) \\
&= 2^{-n-1}\left(\sum_{x,y}(-1)^{x \cdot y}|y, f(x)\rangle + \sum_{x,y}(-1)^{(x \oplus s) \cdot y}|y, f(x \oplus s)\rangle\right) \\
&= 2^{-n-1}\left(\sum_{x,y}(-1)^{x \cdot y}|y, f(x)\rangle + \sum_{x,y}(-1)^{(x \oplus s) \cdot y}|y, f(x)\rangle\right) \\
&= 2^{-n-1}\sum_{x,y}\left[(-1)^{x \cdot y} + (-1)^{x \cdot y + s \cdot y}\right]|y, f(x)\rangle \\
&= 2^{-n-1}\sum_{x,y}(-1)^{x \cdot y}\left[1 + (-1)^{s \cdot y}\right]|y, f(x)\rangle \\
&= 2^{-n}\sum_{x,y\,:\,s \cdot y \text{ is even}}(-1)^{x \cdot y}|y, f(x)\rangle.
\end{aligned}$$

The basis states $|y, f(x)\rangle$ for which $s \cdot y$ is odd cancel out, and we are left with a superposition of only states where $s \cdot y$ is even, with probability amplitudes differing only by a phase factor. So in Step 5 we will see an arbitrary such $y \in \{0,1\}^n$, uniformly at random. If $s = 0^n$, then $s \cdot y$ is even for all y, so each $y \in \{0,1\}^n$ will be seen with probability 2^{-n}. If $s \neq 0$, then $s \cdot y$ is even for exactly half of the $y \in \{0,1\}^n$, each of which will be seen with probability 2^{1-n}.

How does this help us find s? If $s \neq 0^n$ and we get some y in Step 5, then we know that $s \cdot y$ is even, which eliminates half the possibilities for s. Repeating the algorithm will give us some y' independent of y such that $s \cdot y'$ is even. This added constraint will most likely cut our search space in half again. After repeated executions of the algorithm, we will get a series of random constraints like this. After a modest number of repetitions, the constraints taken together will uniquely determine s with high probability. To show this, we need a brief linear algebraic digression, which will also help us when we discuss binary codes later.

Linear Algebra over $\mathbb{Z}_2$. Until now, we've been dealing with vectors and operators with scalars in $\mathbb{C}$ (and occasionally $\mathbb{R}$). These are not the only two possible scalar domains (known in algebra as *fields*) over which to do linear algebra. Another is the two-element field $\mathbb{Z}_2 := \{0,1\}$, with addition and multiplication defined thus:

$+$	0	1
0	0	1
1	1	0

$\times$	0	1
0	0	0
1	0	1

.

Addition and multiplication are the same as in $\mathbb{Z}$, except that $1+1=0$. Addition is also the same as the XOR operator $\oplus$. The additive identity is 0 and the multiplicative identity is 1. Since $x + x = 0$ in $\mathbb{Z}_2$ for any x, the negation $-x$ (additive inverse) of x is x itself. Thus subtraction is the same as addition. Finally, note that for all $x_1, \ldots, x_n \in \mathbb{Z}_2$, $x_1 + \cdots + x_n = 0$ (in $\mathbb{Z}_2$) if and only if $x_1 + \cdots + x_n$ (in $\mathbb{Z}$) is even.

Column vectors, row vectors, and matrices over $\mathbb{Z}_2$ are defined just as over $\mathbb{C}$, except that all the entries are in $\mathbb{Z}_2$ and all scalar arithmetic is done in $\mathbb{Z}_2$. We call these objects *bit vectors* and *bit matrices*. We can identify binary strings in $\{0,1\}^n$ with bit vectors in $\mathbb{Z}_2^n$.

Most of the basic concepts of linear algebra can be extended to $\mathbb{Z}_2$ (indeed, any field). Matrix addition and multiplication, trace and determinant of square matrices, and square matrix inversion are defined completely analogously to the case of $\mathbb{C}$. Same with vector spaces, subspaces, and linear operators. All the basic results of linear algebra carry over to $\mathbb{Z}_2$. For example,

- For any n, tr is a linear operator from the space of $n \times n$ matrices to $\mathbb{Z}_2$, and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any conformant bit matrices A and B such that AB is square.
- $\det(AB) = (\det A)(\det B)$ for any square A and B, and A is invertible iff $\det A \neq 0$.
- $\mathrm{char}_A(\lambda) = \det(A - \lambda I)$ as before. Its roots are the eigenvalues of A.
- Linear combination, linear (in)dependence, span, and the concept of a basis are the same as before. Every bit vector space has a basis, and any two bases of the same space have the same cardinality (the *dimension* of the space).

- The adjoint A^* is defined as the transpose conjugate as before, but in $\mathbb{Z}_2$ we define $0^* = 0$ and $1^* = 1$, and so the adjoint is the same as the transpose in this case.
- The scalar product of two (column) bit vectors x and y is $x^*y = x \cdot y$, but here the result is in $\mathbb{Z}_2$, where 0 represents an even number of 1s in the sum and 1 represents an odd number of 1s. In all of our uses of the dot product of bit vectors, we've only cared about whether the value was even or odd, so we're not losing any utility here.
- Orthogonality can be defined in terms of the dot product as before, as well as mutually orthogonal subspaces and the orthogonal complement $V^\perp$ of a subspace V of some bit vectors space $\mathcal{A}$. If $\mathcal{A}$ has dimension n and $V \subseteq \mathcal{A}$ is a subspace of dimension k, then $V^\perp$ has dimension $n - k$ as before, and $(V^\perp)^\perp = V$ as before.

Not everything works the same over $\mathbb{Z}_2$ as over $\mathbb{C}$. Here are some differences:

- An n-dimensional vector space over $\mathbb{Z}_2$ is finite, with exactly 2^n elements, one for each possible linear combination of the basis vectors
- There is no notion of "positive definite." We can have $x \cdot x = 0$ but $x \neq 0$ (*i.e.*, x has a positive but even number of 1s). The norm of a vector cannot be defined in the same way as with $\mathbb{C}$, however, a useful norm-like quantity associated with each bit vector x is the number of 1s in x, known as the *Hamming weight* of x and denoted $\mathrm{wt}(x)$.
- The concept of unit vector and orthonormal basis don't work over $\mathbb{Z}_2$ like they do over $\mathbb{C}$, and there is no Gram-Schmidt procedure.
- Mutually orthogonal subspaces may have nonzero vectors in their intersection. Indeed, it may be the case that $V \subseteq V^\perp$ for nontrivial V.
- $\mathbb{Z}_2$ is not algebraically closed. This means, for example, that a square matrix may not have any eigenvectors or eigenvalues.

Exercise 14.2 Let

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

be bit matrices. Compute AB, $\operatorname{tr} A$, $\operatorname{tr} B$, $\det A$, and $\det B$. All arithmetic is in $\mathbb{Z}_2$.

Exercise 14.3 Find the two 2×2 matrices over $\mathbb{Z}_2$ that have no eigenvalues or eigenvectors. (Challenging) Prove that there are only two.

Let A be an $m \times n$ matrix (over any field F). The *rank* of A, denoted $\operatorname{rank} A$, is the maximum number of linearly independent columns of A (or rows—it does not matter). Equivalently, it is the dimension of the span of the columns of A (or rows—it does not matter). An $m \times n$ matrix A has *full rank* if $\operatorname{rank} A = \min(m, n)$, and this is the highest rank and $m \times n$ matrix can have. A square matrix is invertible if and only if it has full rank. The *kernal* of A, denoted $\ker A$, is the set of column vectors $v \in F^n$ such that $Av = 0$. The kernal of A is a subspace of F^n. Its dimension is

known as the *nullity* of A. A standard theorem in linear algebra is that the sum of the rank and the nullity of A is equal to the number of columns of A, *i.e.*, n. The rank of any given bit matrix A is easy to compute; you can use Gaussian elimination, for example. If the nullity of A is positive, it is also easy to find a nonzero bit vector v such that $Av = 0$ (the right-hand side is the zero vector (a bit vector)).

Back to Simon's Problem. If we run the quantum algorithm above k times for some $k \geqslant n$, we get k independent, uniformly random vectors $y_1, \ldots, y_k \in \mathbb{Z}_2^n$ such that the following k linear equations hold:

$$\begin{aligned} y_1 \cdot s &= 0, \\ &\vdots \\ y_k \cdot s &= 0. \end{aligned}$$

Let A be the $k \times n$ bit matrix whose rows are the y_i. Then the above can be expressed as the single equation $As = 0$, where 0 denotes the zero vector in $\mathbb{Z}_2^k$. Thus, $s \in \ker A$.

The whole solution to Simon's problem is as follows: Run the algorithm above n times, obtaining $y_1, \ldots, y_n \in \mathbb{Z}_2^n$. Let A be the $n \times n$ bit matrix whose rows are the y_i.

1. If $\operatorname{rank} A < n - 1$, then give up (*i.e.*, output "I don't know").
2. If $\operatorname{rank} A = n$, *i.e.*, if A is invertible, then output 0.
3. Otherwise, $\operatorname{rank} A = n - 1$. Find the unique $s \neq 0$ such that $As = 0$. Using the U_f gate two more times, compute $f(0)$ and $f(s)$. If they are equal, then output s; otherwise, output 0.

Several things need explaining here. For one thing, the algorithm may fail to find s, outputting "I don't know." We'll see that this is reasonably unlikely to happen. For another thing, if we find that A is invertible in Step 2, then we know that $s = A^{-1}0 = 0$, so our output is correct. Finally, in Step 3 we know that an s exists and is unique: the nullity of A is $n - \operatorname{rank} A = n - (n-1) = 1$, so $\ker A$ is a one-dimensional space, which thus has $2^1 = 2$ elements, one of which is the zero vector. The final check is to determine which of these is the correct output. So if the algorithm does output an answer, that answer is always correct. Such a randomized algorithm (with low failure probability) is called a *Las Vegas algorithm*, as opposed to a *Monte Carlo algorithm* which is allowed to give a wrong answer with low probability.

What are the chances of the algorithm failing? If the algorithm fails, then $\operatorname{rank} A < n-1$, which certainly implies that the matrix formed from first $n-1$ rows of A has rank less than $n-1$. So if we bound the latter probability, we bound the probability of failure. For $1 \leqslant k \leqslant n$, let A_k be the bit matrix formed from the first k rows of A. Each row of A is a uniformly random bit vector in the space $S = \{0, s\}^\perp$, which has dimension $n-1$ (if $s \neq 0$) or n (if $s = 0$). Thus S has at least 2^{n-1} vectors. Consider the probability that $\operatorname{rank} A_{n-1} = n-1$, *i.e.*, that A_{n-1} has full rank. This is true iff all rows of A_{n-1} are linearly independent, or equivalently, iff the A_k have full rank for all $1 \leqslant k \leqslant n-1$. We can express this probability as a product of conditional probabilities:

$$\Pr[\operatorname{rank} A_{n-1} = n-1] = \Pr[\operatorname{rank} A_1 = 1] \prod_{k=2}^{n-1} \Pr[\operatorname{rank} A_k = k \mid \operatorname{rank} A_{k-1} = k-1].$$

Clearly, rank $A_1 = 1$ iff its row is a nonzero bit vector in S, and so

$$\Pr[\operatorname{rank} A_1 = 1] = \frac{|S| - 1}{|S|} \geqslant \frac{2^{n-1} - 1}{2^{n-1}} = 1 - 2^{1-n}.$$

Now what is $\Pr[\operatorname{rank} A_k = k \mid \operatorname{rank} A_{k-1} = k-1]$ for $k > 1$? If $\operatorname{rank} A_{k-1} = k-1$, then the rows of A_{k-1} are linearly independent, and thus span a $(k-1)$-dimensional subspace of $D \subseteq S$ that has 2^{k-1} elements. Assuming this, A_k will have full rank iff its last row is linearly independent of the other rows, *i.e.*, the last row is an element of $S - D$. Thus,

$$\Pr[\operatorname{rank} A_k = k \mid \operatorname{rank} A_{k-1} = k-1] = \frac{|S| - |D|}{|S|} \geqslant \frac{2^{n-1} - 2^{k-1}}{2^{n-1}} = 1 - 2^{k-n}.$$

Putting this together, we have

$$\Pr[\operatorname{rank} A_{n-1} = n-1] \geqslant \prod_{k=1}^{n-1} (1 - 2^{k-n}) = \prod_{k=1}^{n-1} (1 - 2^{-k}) = p_{n-1},$$

where we define

$$p_m := \prod_{k=1}^{m} (1 - 2^{-k}) \tag{62}$$

for all $m \geqslant 0$. Clearly, $1 = p_0 > p_1 > \cdots > p_n > \cdots > 0$, and it can be shown that if $p := \lim_{m\to\infty} p_n$, then $1/4 < p < 1/3$. Thus the chances are better than $1/4$ that A_{n-1} will have full rank, and so the algorithm will fail with probability less than $3/4$. This seems high, but if we repeat the whole process r times independently, then the chances that we will fail on *all* r trials is less than $(3/4)^r$, which goes to zero exponentially in r. The expected number of trials necessary to succeed at least once is thus at most $\sum_{r=1}^{\infty} (r/4)(3/4)^{r-1} = 4$.

Shor's Algorithm for Factoring. In the early 1990s, Peter Shor showed how to factor an integer N on a quantum computer in time polynomial in $\lg N$ (which is roughly the number of bits needed to represent N in binary). All known classical algorithms for factoring run exponentially slower than this (with a somewhat liberal definition of "exponentially slower"). Although it has not been shown that no fast classical factorization algorithm exists, it is widely believed that this is the case (and RSA security depends on this being the case). Shor's algorithm is the single most important quantum algorithm to date, because of its implications for public key cryptography. Using similar techniques, Shor also gave quantum algorithms for quickly solving the discrete logarithm problem, which also has cryptographic (actually cryptanalytical) implications. To do Shor's algorithm correctly, we need a couple more mathematical detours.

Modular Arithmetic. If a and m are integers and $m > 0$, then we can divide a by m and get two integer results—quotient and remainder. Put another way, there are unique integers q, r such that $0 \leqslant r < m$ and $a = qm + r$. We let $a \bmod m$ denote the number r. For any integer $m > 1$, we let $\mathbb{Z}_m = \{0, 1, \ldots, m-1\} = \{a \bmod m : a \in \mathbb{Z}\}$, and we define addition and multiplication in $\mathbb{Z}_m$ just as in $\mathbb{Z}$ except that we take the result mod m. Our previous discussion about $\mathbb{Z}_2$ is a special case of this. Arithmetic in $\mathbb{Z}_m$ resembles arithmetic in $\mathbb{Z}$ in several ways:

- Both operations are associative and commutative.
- Multiplication distributes over addition, *i.e.*, $x(y+z) = xy + xz$ in $\mathbb{Z}_m$.
- 0 is the additive identity, and 1 is the multiplicative identity of $\mathbb{Z}_m$.
- A unique additive inverse (negation) $-x \in \mathbb{Z}_m$ exists for each element $x \in \mathbb{Z}_m$, such that $x + (-x) = 0$. In fact, $-0 = 0$, and $-x = m - x$ if $x \neq 0$. Clearly, $-(-x) = x$, and $(-x)y = -xy$ in $\mathbb{Z}_m$. Subtraction is defined as addition of the negation as usual: $x - y = x + (-y)$.
- A multiplicative inverse (reciprocal) may or may not exist for any given element $x \in \mathbb{Z}_m$ (that is, a $b \in \mathbb{Z}_m$ such that $xb = 1$ in $\mathbb{Z}_m$). If it does, it is unique and written x^{-1} or $1/x$, and we say that x is *invertible* or a *unit*. If x is a unit, then so is x^{-1}, and $(x^{-1})^{-1} = x$. 0 is never a unit, but 1 and -1 are always units. Division can be defined as multiplication by the reciprocal as usual, provided the denominator is a unit: $x/y = x(1/y)$, provided $1/y$ exists.
- We define exponentiation as usual: x^n is the product of x with itself n times, where $x \in \mathbb{Z}_m$ and $n \in \mathbb{Z}$ with $n > 0$. We let $x^0 = 1$ by convention. If x is a unit, then we can define $x^{-n} = (1/x)^n$ as usual.

We let $\mathbb{Z}_m^*$ be the set of all units in $\mathbb{Z}_m$. $\mathbb{Z}$ has only two units—1 and -1—but $\mathbb{Z}_m$ may have many units other than ± 1. The units of $\mathbb{Z}_m$ are exactly those elements x that are relatively prime to m (*i.e.*, $\gcd(x, m) = 1$). If m is prime, then all nonzero elements of $\mathbb{Z}_m$ are units. In any case, $\mathbb{Z}_m^*$ contains 1 and is closed under multiplication and reciprocals, but not necessarily under addition.

Exercise 14.4 What is $\mathbb{Z}_{30}^*$? Pair the elements of $\mathbb{Z}_{30}^*$ with their multiplicative inverses.

For any $x \in \mathbb{Z}_m^*$ we define the *order* of x in $\mathbb{Z}_m^*$ to be the least $r > 0$ such that $x^r = 1$. Such an r must exist: The elements of the sequence $1, x, x^2, x^3, \ldots$ are all in $\mathbb{Z}_m$, which is finite, so by the Pigeon Hole Principle there must exist some $0 \leqslant s < t$ such that $x^s = x^t$. Multiplying both sides by x^{-s}, we get $1 = x^{-s}x^s = x^{-s}x^t = x^{t-s}$, and incidentally, $t - s > 0$.

Factoring Reduces to Order Finding. Shor's algorithm does not factor N directly. Instead it solves problem of finding the order of an element $x \in \mathbb{Z}_N^*$. This is enough, as we will now see.

Let N be a large composite integer, and let x be an element of $\mathbb{Z}_N^*$. Suppose that you had at your disposal a black box into which you could feed x and N, and the box would promptly output the order of x in $\mathbb{Z}_N$. Then you could use this box to find a nontrivial factor of N quickly and with high probability via the following (classical!) Las Vegas algorithm:

1. Input: a composite integer $N > 0$.
2. If N is even, then output 2 and quit.
3. If $N = a^b$ for some integers $a, b \geqslant 2$, then output a and quit. (To see that this can be done quickly, note that if $a, b \geqslant 2$ and $a^b = N$, then $2^b \leqslant a^b = N$ and so $2 \leqslant b \leqslant \lg N$. For each b, you can try finding an integer a such that $a^b = N$ by binary search.)

4. (At this point, N is odd and not a power. This means that N has at least two distinct odd prime factors, in particular, there are odd, coprime $p, q > 1$ such that $N = pq$.) Pick a random $x \in \mathbb{Z}_N$.

5. Compute $\gcd(x, N)$ with the Euclidean Algorithm. If $\gcd(x, N) > 1$, then output $\gcd(x, N)$ and quit.

6. (At this point, $x \in \mathbb{Z}_N^*$.) Use the order-finding black box to find the order r of x in $\mathbb{Z}_N^*$.

7. If r is odd, then give up (*i.e.*, output "I don't know" and quit).

8. (r is even.) Compute $y = x^{r/2}$ in $\mathbb{Z}_N$. If $y = -1$ (in $\mathbb{Z}_N$), then give up.

9. ($y \neq -1$.) Compute $\gcd(y - 1, N)$ and output the result.

Shor's quantum algorithm provides the order-finding black box for this reduction.

15 Week 8: Factoring and order finding (cont.)

This algorithm (really a randomized reduction of Factoring to Order Finding) is clearly efficient (polynomial time in $\lg N$), given black-box access to Order Finding. We need to check two things: (i) the algorithm, if it does not give up, outputs a nontrivial factor of N, and (ii) the probability of it giving up is not too big—at most $1 - \varepsilon$ for some constant ε, say.

Notation 15.1 For $a, b \in \mathbb{Z}$, we let $a \mid b$ mean that a divides b, or that b is a multiple of a, precisely, there is a $c \in \mathbb{Z}$ such that $ac = b$. Clearly, if $a > 0$, then $a \mid b$ iff $b = 0$ in $\mathbb{Z}_a$. We write $a \nmid b$ to mean that a does not divide b.

Anything the algorithm outputs in Steps 2, 3, or 5 is clearly correct. The only other output step is Step 9. We claim that $\gcd(y - 1, N)$ is a nontrivial factor of N: We have $y \neq -1$ in $\mathbb{Z}_N$ by assumption, or equivalently, $N \nmid y + 1$. Also, $y \neq 1$ in $\mathbb{Z}_N$, since otherwise $x^{r/2} = 1$ in $\mathbb{Z}_N$, which contradicts the fact that r is the least such exponent. Thus $N \nmid y - 1$. Yet we have $y^2 = x^r = 1$ in $\mathbb{Z}_N$, which means that $N \mid y^2 - 1 = (y + 1)(y - 1)$. So N divides $(y + 1)(y - 1)$ but neither of its two factors. The only way this can happen is when $y + 1$ includes some, but not all, of the prime factors of N, and likewise with $y - 1$. Thus $1 < \gcd(y - 1, N) < N$, and so we output a nontrivial factor of N in Step 9.

The algorithm could give up in Steps 7 or 8. Giving up in Step 7 means that r is odd. We show that at most half the elements of $\mathbb{Z}_N^*$ have odd order, and so the algorithm gives up in Step 7 with probability at most $1/2$. In fact, we show that if $x \in \mathbb{Z}_N^*$ has odd order r, then $-x$ in $\mathbb{Z}_N$ (which is also in $\mathbb{Z}_N^*$) has order $2r$. So at least one element of each pair $\pm x$ has even order, and so we're done since $\mathbb{Z}_N^*$ is made up of such disjoint pairs. First, we have

$$(-x)^{2r} = (-1)^{2r} x^{2r} = ((-1)^2)^r (x^r)^2 = 1^r 1^2 = 1,$$

where all arithmetic is in $\mathbb{Z}_N$. So $-x$ has order at most $2r$. Now suppose that $-x$ has order $s < 2r$. Then $1 = (-x)^s = (-1)^s x^s$. We must have $s \neq r$, for otherwise this would become $1 = (-1)^r x^r = (-1)^r = -1$, since r is odd (and since $N > 2$, we have $1 \neq -1$ in $\mathbb{Z}_N$). Now since $0 < s < 2r$ but $s \neq r$, we must have $x^s \neq 1$, and because $(-1)^s x^s = 1$, we cannot have $(-1)^s = 1$. Thus $(-1)^s = x^s = -1$. But now,

$$-1 = (-1)^r = (x^s)^r = x^{rs} = (x^r)^s = 1^s = 1,$$

contradiction. Therefore, $-x$ has order $2r$.

We claim that, if the algorithm makes it to Step 8, then it gives up in this step at most half the time. We won't prove the claim, since that would get us too much into number theoretic waters, but we'll give some reasonable evidence that it is true. Recall that by Step 8, we have $N = pq$, where p and q are odd and coprime. Define a map $d : \mathbb{Z}_N \to \mathbb{Z}_p \times \mathbb{Z}_q$ such that $d(x) = (x \bmod p, x \bmod q)$ for all $x \in \mathbb{Z}_N$. Here are some easy-to-prove facts about d. To avoid confusion, for any $n > 1$ we'll use $+_n$ and $\cdot_n$ to denote addition in $\mathbb{Z}_n$ and multiplication in $\mathbb{Z}_n$, respectively. Let $x, y \in \mathbb{Z}_N$ be arbitrary, and let $d(x) = (x_1, x_2)$ and $d(y) = (y_1, y_2)$.

- $d(x +_N y) = (x_1 +_p y_1, x_2 +_q y_2)$.

- $d(x \cdot_N y) = (x_1 \cdot_p y_1, x_2 \cdot_q y_2)$.
- $d(1) = (1, 1)$.
- $d(-1) = (-1, -1)$. More generally, $d(-x) = (-x_1, -x_2)$.
- $x \in \mathbb{Z}_N^*$ if and only if $x_1 \in \mathbb{Z}_p^*$ and $x_2 \in \mathbb{Z}_q^*$.

It turns out that d is a bijection from $\mathbb{Z}_N$ to $\mathbb{Z}_p \times \mathbb{Z}_q$. This is a consequence of the following classic theorem in number theory:

Theorem 15.2 (Chinese Remainder Theorem (dyadic version)) *Let $p, q > 0$ be coprime and let $N = pq$. Define $d : \mathbb{Z}_N \to \mathbb{Z}_p \times \mathbb{Z}_q$ by $d(x) = (x \bmod p, x \bmod q)$. Then d is a bijection, i.e., for every $x_1 \in \mathbb{Z}_p$ and $x_2 \in \mathbb{Z}_q$, there exists a unique $x \in \mathbb{Z}_N$ such that $d(x) = (x_1, x_2)$.*

I'll include the proof here for you to read on your own if you want, but I won't present it in class.

Proof. Set $\tilde{p} = p \bmod q$ and $\tilde{q} = q \bmod p$. Since $\gcd(p, q) = 1$, we also have $\gcd(\tilde{p}, q) = \gcd(p, \tilde{q}) = 1$, and hence $\tilde{p} \in \mathbb{Z}_q^*$ and $\tilde{q} \in \mathbb{Z}_p^*$. Let $\tilde{p}^{-1}$ and $\tilde{q}^{-1}$ be the reciprocals of $\tilde{p}$ in $\mathbb{Z}_q^*$ and of $\tilde{q}$ in $\mathbb{Z}_p$, respectively. Given any $x_1 \in \mathbb{Z}_p$ and $x_2 \in \mathbb{Z}_q$, let $x = (x_1\tilde{q}^{-1}q + x_2\tilde{p}^{-1}p) \bmod N$ (normal arithmetic in $\mathbb{Z}$). Clearly, $x \in \mathbb{Z}_N$. Then letting $d(x) = (y_1, y_2)$, we get

$$\begin{aligned} y_1 &= [(x_1\tilde{q}^{-1}q + x_2\tilde{p}^{-1}p) \bmod N] \bmod p \\ &= (x_1\tilde{q}^{-1}q + x_2\tilde{p}^{-1}p) \bmod p \\ &= x_1\tilde{q}^{-1}q \bmod p \\ &= x_1\tilde{q}^{-1}\tilde{q} \bmod p \\ &= x_1 \bmod p \\ &= x_1, \end{aligned}$$

and similarly,

$$\begin{aligned} y_2 &= [(x_1\tilde{q}^{-1}q + x_2\tilde{p}^{-1}p) \bmod N] \bmod q \\ &= (x_1\tilde{q}^{-1}q + x_2\tilde{p}^{-1}p) \bmod q \\ &= x_2\tilde{p}^{-1}p \bmod q \\ &= x_2\tilde{p}^{-1}\tilde{p} \bmod q \\ &= x_2 \bmod q \\ &= x_2. \end{aligned}$$

Thus $d(x) = (x_1, x_2)$, which proves that d is surjective. To see that d is injective, let $x, y \in \mathbb{Z}_N$ be such that $d(x) = d(y) = (x_1, x_2)$. Then $d(x -_N y) = (x_1 -_p x_1, x_2 -_q x_2) = (0, 0)$, and so we have $(x - y) \bmod p = (x - y) \bmod q = 0$, or equivalently, $p \mid x - y$ and $q \mid x - y$. But since p and q are coprime, we must have $N \mid x - y$, and so,

$$x = x \bmod N = y \bmod N = y,$$

which shows that d is an injection. □

We won't discuss it here, but given x_1, x_2, one can quickly (and classically) compute inverses in $\mathbb{Z}_n^*$, and thus find the unique x such that $d(x) = (x_1, x_2)$, using the Extended Euclidean Algorithm.

Exercise 15.3 In this exercise, you will prove some standard results about the cardinality of $\mathbb{Z}_n^*$ for any integer $n > 1$. For any such n, the *Euler totient function* is defined as

$$\varphi(n) := |\mathbb{Z}_n^*|,$$

which is the number of elements of $\mathbb{Z}_n$ that are relatively prime to n. By convention, $\varphi(1) := 1$.

1. Show that if $a, b > 0$ are coprime, then $\varphi(ab) = \varphi(a)\varphi(b)$. [Hint: Show that the bijection d defined in Theorem 15.2 above (with $p = a$ and $q = b$) matches elements of $\mathbb{Z}_{ab}^*$ with elements of $\mathbb{Z}_a^* \times \mathbb{Z}_b^*$ and vice versa.]

2. Show that if n is some power of a prime p, then $\varphi(n) = n(p-1)/p$. [Hint: An element $x \in \mathbb{Z}_n$ is relatively prime to n iff x is not a multiple of p.]

3. Conclude that if $n = q_1^{e_1} q_2^{e_2} \cdots q_k^{e_k}$ is the prime factorization of n, where $q_1 < q_2 < \ldots < q_k$ are all prime and $e_1, e_2, \ldots, e_k > 0$, then

$$\varphi(n) = \prod_{j=1}^{k} q_j^{e_j - 1}(q_j - 1).$$

   Dividing this by n, we get

$$\frac{\varphi(n)}{n} = \prod_{j=1}^{k} \left(1 - \frac{1}{q_j}\right). \tag{63}$$

4. Using Equation (63), prove that for all integers $n > 0$,

$$\frac{\varphi(n)}{n} \geqslant \frac{1}{1 + \lg n}. \tag{64}$$

   [Hint: Use the fact that if $n = q_1^{e_1} q_2^{e_2} \cdots q_k^{e_k}$ is the prime factorization of n as above, then $k \leqslant \lg n$ (why?) and the fact that $q_j \geqslant j + 1$ for all $1 \leqslant j \leqslant k$ (why?).]

Exercise 15.4 (Challenging) Show that $\varphi(n)/n \geqslant 1/\lg n$ for all integers $n > 1$ except 2 and 6. [Hint: For $\ell > 0$, let n_ℓ be the product of the first ℓ primes. Using the inequality (64) above, show that, for any $\ell > 0$, if $\varphi(n_\ell)/n_\ell \geqslant 1/\lg n_\ell$, then $\varphi(n)/n \geqslant 1/\lg n$ for all $n \geqslant n_\ell$. Then find an ℓ for which the hypothesis is true.]

Back to the issue at hand. When $y = x^{r/2}$ is computed in Step 8, we have $y^2 = x^r = 1$, and so y is one of the square roots of 1 in $\mathbb{Z}_N$. Both 1 and -1 are square roots of 1 in $\mathbb{Z}_N$ for any N, but in this case ($N = pq$ as above) there are at least two others. Whereas $d(1) = (1, 1)$ and $d(-1) = (-1, -1)$, by the Chinese Remainder Theorem, there is an $x \in \mathbb{Z}_N$ such that $d(x) = (1, -1)$. By the bijective nature of d, we have $x \neq \pm 1$, and so x and $-x$ are two additional square roots of 1 besides ± 1. There could be still others. We won't prove it, but if x is chosen uniformly at random among those elements of $\mathbb{Z}_N^*$ with even order, then $x^{r/2}$ is at least as likely to be one of the other square roots of 1 than ± 1, where r is the order of x. Thus Step 8 gives up with probability at most $1/2$.

So the whole reduction succeeds in outputting a nontrivial factor of N with probability at least $1/4$. As with Simon's algorithm, we can expect to run this reduction about four times to find such a factor. Running it additional times decreases the likelihood of failure exponentially.

Geometric series. This elementary fact will be useful in what is to come.

Proposition 15.5 *For any $r \in \mathbb{C}$ such that $r \neq 1$, and for any integer $n \geqslant 0$,*

$$\sum_{i=0}^{n-1} r^i = \frac{r^n - 1}{r - 1} .$$

You can prove this by induction on n. If $n = 0$, then both sides are 0. Now assume the equation holds for fixed $n \geqslant 0$. Then

$$\sum_{i=0}^{n} r^i = r^n + \sum_{i=0}^{n-1} r^i = r^n + \frac{r^n - 1}{r - 1} = \frac{r^n(r-1) + r^n - 1}{r - 1} = \frac{r^{n+1} - 1}{r - 1} .$$

The sum $\sum_{i=0}^{n-1} r^i$ is called a *finite geometric series* with *ratio* r.

The Quantum Fourier Transform. The Fourier transform is of fundamental importance in many areas of science, math, and engineering. For example, it is used in signal processing to pick out component frequencies in a periodic signal (and we will see how this applies to Shor's order-finding algorithm). The auditory canal inside your ear acts as a natural Fourier transformer, allowing your brain to register different frequencies (of musical notes, say) inherent in the sound waves entering the ear.

A quantum version of the Fourier transform, known as the *quantum Fourier transform* or QFT, is a crucial ingredient in Shor's algorithm.

Let $m > 1$ be an integer. We will define the m-dimensional *discrete Fourier transform*[15] DFT_m is a linear map $\mathbb{C}^m \to \mathbb{C}^m$ that takes a vector $x = (x_0, \ldots, x_{m-1}) \in \mathbb{C}^m$ and maps it to the vector $y = (y_0, \ldots, y_{m-1}) \in \mathbb{C}^m$ satisfying

$$y_j = \frac{1}{\sqrt{m}} \sum_{k=0}^{m-1} e^{2\pi i j k/m} x_k$$

for all $0 \leqslant j < m$.[16] Set $\omega_m := e^{2\pi i/m}$. Clearly, m is the least positive integer such that $\omega_m^m = 1$. We call ω_m the *principal m-th root of unity*. Note that $\omega_m^a = \omega_m^{a \bmod m}$ for any $a \in \mathbb{Z}$, so we can consider the exponent of ω_m to be an element of $\mathbb{Z}_m$.

The matrix corresponding to DFT_m is the $m \times m$ matrix whose (j,k)th entry is $[\mathrm{DFT}_m]_{jk} = \omega_m^{jk}/\sqrt{m}$, for all $0 \leqslant j, k < m$, *i.e.*, for all $j, k \in \mathbb{Z}_m$. (It will be more convenient for the time being to start the indexing at zero rather than one.) In fact, DFT_m is unitary, and it is worth seeing why this is so. We check that $(\mathrm{DFT}_m)^* \mathrm{DFT}_m$ has diagonal entries 1 and off-diagonal entries 0. For general j, k, we have

$$[(\mathrm{DFT}_m)^* \mathrm{DFT}_m]_{jk} = \frac{1}{m} \sum_{\ell \in \mathbb{Z}_m} \omega_m^{-\ell j} \omega_m^{\ell k} = \frac{1}{m} \sum_{\ell \in \mathbb{Z}_m} \omega_m^{\ell(k-j)}. \tag{65}$$

[15] There are continuous versions of the Fourier transform.

[16] There is some variation in the definition of DFT_m in different sources; for example, there may be a minus sign in the exponent of e, or there may be no factor $1/\sqrt{m}$ in front. The current definition is the most useful for us.

If $j = k$, then the right-hand side is $(1/m)\sum_{\ell \in \mathbb{Z}_m} 1 = 1$. Now suppose $j \neq k$. Then $0 < |k - j| < m$, and so $\omega_m^d \neq 1$, where $d = k - j$. To see that the sum on the right-hand side of (65) is 0, notice that it is a finite geometric series with ratio $\omega_m^d \neq 1$, and so we have

$$\sum_{\ell \in \mathbb{Z}_m} (\omega_m^d)^\ell = \frac{(\omega_m^d)^m - 1}{\omega_m^d - 1} = \frac{(\omega_m^m)^d - 1}{\omega_m^d - 1} = 0\,,$$

because $(\omega_m^m)^d = (\omega_m^0)^d = 1^d = 1$.

Naively applying DFT_m to a vector in $\mathbb{C}^m$ requires $\Theta(m^2)$ scalar arithmetic operations. A much faster method, known as the *Fast Fourier Transform* (FFT), can do this with $O(m \lg m)$ scalar arithmetic operations. The FFT was described by Cooley & Tukey in 1965, but the same idea can be traced back to Gauss. It uses divide-and-conquer, and is easiest to describe when m is a power of 2. The FFT is also easily parallelizable: it can be computed by an arithmetic circuit of width m and depth $\lg m$ called a *butterfly network*. Because of this, the FFT has been rated as the second most useful algorithm ever, second only to fast sorting. Besides its use in digital signal processing, it is also used to implement the asymptotically fastest known algorithms, due to Schönhage & Strassen, for multiplying integers and polynomials.

It was Shor who first showed that DFT_{2^n} could be implemented by a quantum circuit on n qubits with size polynomial in n, and his idea is based on the Fast Fourier Transform. From now on, the dimension will be a power of 2, so I'll define the n-qubit *quantum Fourier transform* QFT_n to be DFT_{2^n}. For notational convenience, I'll also define

$$e_n(x) := \omega_{2^n}^x = e^{2\pi i x/2^n}$$

for all $n, x \in \mathbb{Z}$ with $n \geqslant 0$. Note that

- $e_n(x + y) = e_n(x)e_n(y)$ for all $y \in \mathbb{Z}$, and
- $e_n(x) = e_n(x \bmod 2^n)$.

Thus, for any $x \in \mathbb{Z}_{2^n}$, we have

$$\mathrm{QFT}_n|x\rangle = \frac{1}{2^{n/2}} \sum_{y \in \mathbb{Z}_{2^n}} e_n(xy)|y\rangle\,.$$

Interestingly, this sum factors completely.

$$\begin{aligned} \mathrm{QFT}_n|x\rangle &= \frac{1}{2^{n/2}}(\,|0\rangle + e_1(x)|1\rangle\,) \otimes (\,|0\rangle + e_2(x)|1\rangle\,) \otimes \cdots \otimes (\,|0\rangle + e_n(x)|1\rangle\,) \\ &= \frac{1}{2^{n/2}} \bigotimes_{k=1}^{n} (\,|0\rangle + e_k(x)|1\rangle\,). \end{aligned}$$

Exercise 15.6 (Challenging) Verify this fact.

Before we describe a circuit for QFT_n, we will sketch out and analyze Shor's quantum algorithm for order-finding, which is a Monte Carlo algorithm. This description and the QFT_n circuit layout later on are adapted with modifications from a paper by Cleve & Watrous in 2000.

1. Input: $N > 1$ and $a \in \mathbb{Z}_N^*$ with $a > 1$. (The algorithm attempts to find the order of a in $\mathbb{Z}_N^*$.) Let $n = \lceil \lg N \rceil$.

2. Initialize a $2n$-qubit register and an n-qubit register in the state $|0\rangle|0\rangle$. Here we will label the basis states of a register with nonnegative integers via their usual binary representations.

3. Apply a Hadamard gate to each qubit of the first register, obtaining the state

$$(H^{\otimes 2n} \otimes I)|0\rangle|0\rangle = \frac{1}{2^n} \sum_{x \in \mathbb{Z}_{2^{2n}}} |x\rangle|0\rangle.$$

4. Apply a classical quantum circuit for modular exponentiation that sends $|x\rangle|0\rangle$ to $|x\rangle|a^x \bmod N\rangle$, obtaining the state

$$|\varphi\rangle = \frac{1}{2^n} \sum_{x \in \mathbb{Z}_{2^{2n}}} |x\rangle|a^x \bmod N\rangle \,. \tag{66}$$

   (We can imagine that N and a are hard-coded into the circuit, which means that the circuit must be built in a preprocessing step after the inputs N and a are known. Alternatively, we can keep N and a in separate quantum registers that don't change during the course of the computation, then feed them into this circuit when they're needed.)

5. (Optional) Measure the second register in the computational basis, obtaining some classical value $w \in \mathbb{Z}_N$, which is ignored.[17]

6. Apply QFT_{2^n} to the first register.

7. Measure the first register (in the computational basis), obtaining some value $y \in \mathbb{Z}_{2^{2n}}$. (This ends the quantum part of the algorithm.)

8. Find the smallest coprime integers k and $r > 0$ such that

$$\left| \frac{y}{2^{2n}} - \frac{k}{r} \right| \leqslant 2^{-2n-1}. \tag{67}$$

   (We are just finding a reasonably good rational approximation to $y/2^{2n}$ that has small denominator r. This can be done classically using continued fractions. See below.)

9. Classically compute $a^r \bmod N$. If the result is 1, then output r. Otherwise, give up.

16 Week 8: Shor's algorithm (cont.)

Let R be the order of a in $\mathbb{Z}_N^*$. The whole key to proving that Shor's algorithm works is to show that in Step 9 the algorithm outputs R with high probability. First, we'll show that a single run of the algorithm above outputs R with probability at least $4/(\pi^2 n) - O(n^{-2})$, and so if we run the

[17] Since we ignore the result of the measurement, this step is entirely superfluous; the algorithm would would do just as well without it. Including this step, however, collapses the state, which simplifies the analysis greatly and allows us to ignore the second register altogether.

algorithm about $\pi^2 n/4$ times, we will succeed with high probability. The actual single-run success probability is usually much higher than $4/(\pi^2 n)$, but $4/(\pi^2 n)$ is a good enough approximate lower bound, and it is easier to derive than a tighter lower bound. After the analysis, we'll discuss how the quantum Fourier transform and the (classical) continued fraction algorithm used in Step 8 are implemented.

Shor's algorithm, if it succeeds, will be guaranteed to output some $r > 0$ such that $a^r = 1$ in $\mathbb{Z}_N$. It is possible—although very unlikely—that r is a multiple of R, but not equal to it. If we run the algorithm until it succeeds k times and take the gcd of the k results, then the chances of not getting R are at most $(1 - 4/(\pi^2 n))^k$, which decrease exponentially with k. If we only want to find a nontrivial factor of N, then we use this algorithm to implement the black box in the Factoring-to-Order-Finding reduction. As the next exercise shows, we don't need to worry about the value returned by the black box if it succeeds.

Exercise 16.1 Suppose that on input N and $x \in \mathbb{Z}_N^*$, the black box used in the reduction from Factoring to Order Finding is only guaranteed to output *some* r with $0 < r < 2^{2n}$ such that $x^r = 1$ in $\mathbb{Z}_N$, where $n = \lceil \lg N \rceil$. Show how to modify the reduction slightly so that it succeeds with the same probability as it did before when the box always outputted the order of x in $\mathbb{Z}_N^*$. [Hint: Let R be the order of x in $\mathbb{Z}_N^*$. First, given any multiple r of R, show how to find an *odd* multiple of R (that is, a number of the form cR where c is odd) that is no bigger than r. Second, show that the probability of success of the reduction is the same if the black box returns some odd multiple of R.]

Analysis of Shor's Algorithm. Let R be the order of a in $\mathbb{Z}_N^*$. We can express x uniquely as $qR + s$ with $s \in \mathbb{Z}_R$ and note that, owing to the periodicity of $a^x \bmod N$,

$$a^x \bmod N = a^{qR+s} \bmod N = (a^R)^q a^s \bmod N = 1^q a^s \bmod N = a^s \bmod N .$$

Then letting $Q := 2^{2n}$ and

$$M := \left\lfloor \frac{2^{2n}}{R} \right\rfloor = \left\lfloor \frac{Q}{R} \right\rfloor ,$$

we rewrite the state $|\varphi\rangle$ of Equation (66) as

$$\begin{aligned} |\varphi\rangle &= \frac{1}{\sqrt{Q}} \left(\sum_{q \in \mathbb{Z}_M} \sum_{s \in \mathbb{Z}_R} |qR + s\rangle |a^s \bmod N\rangle + \sum_{s=0}^{(2^{2n} \bmod R)-1} |MR + s\rangle |a^s \bmod N\rangle \right) \\ &= \frac{1}{\sqrt{Q}} \sum_{q \in \mathbb{Z}_M} \sum_{s \in \mathbb{Z}_R} |qR + s\rangle |a^s \bmod N\rangle + O\left(2^{-n/2}\right) . \end{aligned}$$

Now when the second register is measured in the next step, we obtain some $w = a^s \bmod N$ corresponding to some unique $s \in \mathbb{Z}_R$. The state after this measurement then collapses to either

$$\frac{1}{\sqrt{M+1}} \left(\sum_{q \in \mathbb{Z}_{M+1}} |qR + s\rangle \right) |w\rangle ,$$

if $0 \leqslant s < 2^{2n} \bmod R$, or to

$$\frac{1}{\sqrt{M}} \left(\sum_{q \in \mathbb{Z}_M} |qR + s\rangle \right) |w\rangle \ ,$$

if $2^{2n} \bmod R \leqslant s < R$. It does not really matter which is the case, as the analysis is nearly identical and the conclusions (particularly Corollary 16.4, below) are the same either way, so for simplicity, we'll assume the latter case applies.[18] Also, the second register will no longer participate in the algorithm, so we can ingore it from now on. To summarize, the post-measurement state of the first register is then given as

$$|\eta\rangle = \frac{1}{\sqrt{M}} \sum_{q \in \mathbb{Z}_M} |qR + s\rangle \ . \tag{68}$$

The next step of the algorithm applies QFT_{2^n} to this state to obtain

$$\begin{aligned} |\psi\rangle &= QFT_{2^n}|\eta\rangle = \frac{1}{\sqrt{M}} \sum_{q \in \mathbb{Z}_M} QFT_{2^n}|qR + s\rangle & (69) \\ &= \frac{1}{\sqrt{QM}} \sum_{q \in \mathbb{Z}_M} \sum_{y \in \mathbb{Z}_{2^{2n}}} e_{2^n}((qR + s)y)|y\rangle & (70) \\ &= \frac{1}{\sqrt{QM}} \sum_{y} \left[\sum_{q} e_{2^n}((qR + s)y) \right] |y\rangle & (71) \\ &= \frac{1}{\sqrt{QM}} \sum_{y} e_{2^n}(sy) \left[\sum_{q} e_{2^n}(qRy) \right] |y\rangle \ . & (72) \end{aligned}$$

Finally, the first register is measured, obtaining $y \in \mathbb{Z}_{2^{2n}}$ with probability

$$\Pr[y] = \left| \frac{e_{2^n}(sy)}{\sqrt{QM}} \sum_{q} e_{2^n}(qRy) \right|^2 = \frac{|e_{2^n}(sy)|^2}{QM} \left| \sum_{q} e_{2^n}(qRy) \right|^2 = \frac{1}{QM} \left| \sum_{q \in \mathbb{Z}_M} e_{2^n}(qRy) \right|^2 .$$

We'll show that $\Pr[y]$ spikes when y/Q is close to a multiple of $1/R$, but first some intuition. Permit me an acoustical analogy. Think of the column vector $|\eta\rangle$ as a periodic signal with period R, *i.e.*, the entries at indices $x = qR + s$ (for integral q) have value $1/\sqrt{M}$, and all the other entries are 0. The "frequency" of this signal is then $1/R$, and since the Fourier transform is good at picking out frequencies, we'd expect to see a "spike" in the probability amplitude of the Fourier transformed state $|\psi\rangle$ of Equation (72) right around the frequencies $1/R$, $2/R$, $3/R$, ..., with $1/R$ being the fundamental component of the signal and the others being overtones (higher harmonics). This is exactly what happens, and it is the whole point of using the QFT. The larger the signal sample, the sharper and narrower the spikes will be. We choose a sample of length Q, which is at least N^2, giving us at least $N^2/R \geqslant N$ periods of the function. This turns out to give us sufficiently sharp spikes to approximate R with high probability.

[18] For the former case, just substitute $M + 1$ for M in the analysis to follow.

We now concentrate on the scalar quantity

$$\sum_{q \in \mathbb{Z}_M} e_{2n}(qRy) \tag{73}$$

in the expression for $\Pr[y]$, above. For every $y \in \mathbb{Z}_Q$ define

$$s_y := \begin{cases} Ry \bmod Q & \text{if } Ry \bmod Q \leqslant Q/2, \\ (Ry \bmod Q) - Q & \text{if } Ry \bmod Q > Q/2. \end{cases}$$

That is, s_y is the remainder of Ry divided by Q with least absolute value. We have $|s_y| \leqslant Q/2$, and in addition, $s_y \equiv Ry \pmod{Q}$, and thus

$$e_{2n}(qRy) = e_{2n}(qs_y) \tag{74}$$

for all q, and so (73) becomes

$$\sum_{q \in \mathbb{Z}_M} e_{2n}(qRy) = \sum_{q \in \mathbb{Z}_M} e_{2n}(qs_y) = \begin{cases} \dfrac{e_{2n}(Ms_y) - 1}{e_{2n}(s_y) - 1} & \text{if } s_y \neq 0, \\ \\ M & \text{if } s_y = 0, \end{cases} \tag{75}$$

noting that $\sum_{q \in \mathbb{Z}_M} e_{2n}(qs_y)$ is a finite geometric series with ratio $e_{2n}(s_y)$, provided $s_y \neq 0$.

If $|s_y|$ is small, then Ry/Q is close to an integer, and so y/Q is close to an integer multiple of $1/R$, which makes Step 8 of the algorithm more likely to find $r = R$. So we want to show that $|s_y|$ is small with reasonably high probability. The following claim shows that if $|s_y|$ is small enough, then (75) has large absolute value. This is true intuitively because the terms of the sum on the left are all pointing roughly in the same direction in the complex plane and so they add constructively. (Conversely, if $|s_y|$ is large, then the terms in the sum wrap around the unit circle many times, mostly canceling each other out and giving (75) a small absolute value.)

Definition 16.2 We say that $y \in \mathbb{Z}_Q$ is *okay* if $|s_y| \leqslant R/2$.

Claim 16.3 *If y is okay, then $\left|\sum_{q \in \mathbb{Z}_M} e_{2n}(qRy)\right| \geqslant 2M/\pi$.*

Proof. Fix y and suppose that $|s_y| \leqslant R/2$. If $s_y = 0$, then the claim clearly holds by (75), so assume $s_y \neq 0$. Starting from Equation (75) and using Exercise 2.5, we have

$$\begin{aligned} \left|\frac{e_{2n}(Ms_y) - 1}{e_{2n}(s_y) - 1}\right| &= \left|\frac{e_{2n}(Ms_y/2)[e_{2n}(Ms_y/2) - e_{2n}(-Ms_y/2)]}{e_{2n}(s_y/2)[e_{2n}(s_y/2) - e_{2n}(-s_y/2)]}\right| \\ &= \frac{|e_{2n}(Ms_y/2)|\,|e_{2n}(Ms_y/2) - e_{2n}(-Ms_y/2)|}{|e_{2n}(s_y/2)|\,|e_{2n}(s_y/2) - e_{2n}(-s_y/2)|} \\ &= \left|\frac{e_{2n}(Ms_y/2) - e_{2n}(-Ms_y/2)}{e_{2n}(s_y/2) - e_{2n}(-s_y/2)}\right| \\ &= \left|\frac{2i\sin(M\theta)}{2i\sin\theta}\right| = \left|\frac{\sin(M\theta)}{\sin\theta}\right|, \end{aligned}$$

where $\theta := \pi s_y/Q$. Since we have

$$|Ms_y| \leqslant \frac{Q|s_y|}{R} \leqslant \frac{Q}{2} ,$$

we know that $|\theta| \leqslant \pi/(2M)$ and $|M\theta| \leqslant \pi/2$. This gives

$$\left|\frac{\sin(M\theta)}{\sin\theta}\right| = \frac{\sin|M\theta|}{\sin|\theta|} \geqslant \frac{\sin|M\theta|}{|\theta|} .$$

It is readily checked that the function $\frac{\sin x}{x}$ is decreasing in the interval $(0, \pi/2]$, so

$$\frac{\sin|M\theta|}{|\theta|} \geqslant \frac{\sin(\pi/2)}{\pi/(2M)} = \frac{2M}{\pi}$$

as desired. □

Corollary 16.4 *If y is okay, then*

$$\Pr[y] \geqslant \frac{4M}{Q\pi^2} \geqslant \frac{4}{R\pi^2} - O(1/Q) = \frac{4}{R\pi^2} - O\left(2^{-2n}\right) .$$

So for each individual okay $y \in \mathbb{Z}_Q$, we get a relatively large (but still exponentially small) probability of seeing that particular y. We'll need the additional fact that there are many okay y. The following claim is obvious, so we'll give it without proof:

Claim 16.5 *For every $k \in \mathbb{Z}_R$, there exists $y \in \mathbb{Z}_Q$ such that Ry is in the closed interval $[Qk - R/2, Qk + R/2]$. Each such y is okay.*

These intervals are pairwise disjoint for different k. This means there are at least R many okay y. By Corollary 16.4, the chances of finding an okay y in Step 7 of the algorithm are then

$$\Pr[y \text{ is okay}] = \sum_{y \text{ is okay}} \Pr[y] \geqslant R\left[\frac{4}{R\pi^2} - O(2^{-2n})\right] \geqslant \frac{4}{\pi^2} - O(2^{-n}) .$$

Let y be the value measured in Step 7, and suppose that y is okay. Then there is some least $k_y \in \mathbb{Z}$ such that

$$Qk_y - R/2 \leqslant Ry \leqslant Qk_y + R/2. \tag{76}$$

(Actually, k_y is unique satisfying (76) because the intervals don't overlap.) Dividing by QR and rearranging, (76) becomes

$$\left|\frac{y}{Q} - \frac{k_y}{R}\right| \leqslant \frac{1}{2Q} = 2^{-2n-1},$$

and so k_y/R satisfies Equation (67). Now Step 8 produces the *least* k and r satisfying (67), so we have two possible issues to address:

1. The k and r found in Step 8 are such that $k/r \neq k_y/R$.

2. The k and r found in Step 8 satisfy $k/r = k_y/R$, but $r < R$ because the fraction k_y/R is not in lowest terms (k/r is always given in lowest terms).

It turns out that the first issue never arises. To see this, first notice that k/r and k_y/R must be close to each other, because they are both close to y/Q:

$$\left|\frac{k_y}{R} - \frac{k}{r}\right| = \left|\frac{k_y}{R} - \frac{y}{Q} + \frac{y}{Q} - \frac{k}{r}\right| \leqslant \left|\frac{y}{Q} - \frac{k_y}{R}\right| + \left|\frac{y}{Q} - \frac{k}{r}\right| \leqslant \frac{1}{Q} = 2^{-2n}\,, \tag{77}$$

since both k/r and k_y/R satisfy (67). Now suppose for the sake of contradiction that $k/r \neq k_y/R$. Recall that $R < N \leqslant 2^n = \sqrt{Q}$, and also note that $r \leqslant R$ by the minimality of r. Then we also have

$$0 \neq \left|\frac{k_y}{R} - \frac{k}{r}\right| = \left|\frac{k_y r - kR}{rR}\right| = \frac{|k_y r - kR|}{rR} \geqslant \frac{1}{rR} > \frac{1}{\sqrt{Q}\,\sqrt{Q}} = \frac{1}{Q} = 2^{-2n}\,,$$

which contradicts (77). (The first inequality comes from the fact that $k_y r - kR$ is a nonzero integer; the second comes from the fact that $r \leqslant R < \sqrt{Q}$.) Thus if y is okay, we must have $k/r = k_y/R$.

The second issue is more of a problem. It arises when k_y and R are not coprime, whence $k = k_y/g$ and $r = R/g$, where $g = \gcd(k_y, R) > 1$.

Definition 16.6 We say that $y \in \mathbb{Z}_Q$ is *good* if y is okay and k_y is relatively prime to R.

Claim 16.7 *If Step 7 produces a good y, then $r = R$ is found in Step 8.*

Proof. Let $y \in \mathbb{Z}_Q$ be good. We have $k/r = k_y/R$, since y is okay. But since both fractions are in lowest terms, we must have $k = k_y$ and $r = R$. □

Claim 16.8 *There are at least $\varphi(R)$ many good $y \in \mathbb{Z}_Q$.*

(Recall that $\varphi(n)$ is Euler's totient function, defined in Exercise 15.3.)

Proof. Claim 16.5 says that every $k \in \mathbb{Z}_R$ is equal to k_y for some okay $y \in \mathbb{Z}_Q$. There are $\varphi(R)$ many k coprime with R, so there are at least $\varphi(R)$ many good y. □

Now we can combine all our claims to get our main Theorem 16.9, below.

Theorem 16.9 *The probability that $r = R$ is found in Step 8 is at least $4/(\pi^2 n) - O(n^{-2})$.*

Proof. By Claim 16.7, it suffices to show that a good y is found in Step 7 with probability at least $4/(\pi^2 n) - O(n^{-2})$. By Claim 16.8 and Equation (64), there are at least

$$\varphi(R) \geqslant \frac{R}{1 + \lg R} \geqslant \frac{R}{1 + \lg N} \geqslant \frac{R}{n+1}$$

many good y. By Corollary 16.4, each good y occurs with probability at least about $4/(R\pi^2)$, and so

$$\Pr[y \text{ is good}] \geqslant \frac{4}{\pi^2 n} - O(n^{-2})\,.$$

□

There are some tricks to (modestly) boost the probability of success of Shor's algorithm while keeping the number of repetitions of the whole quantum computation to a minimum. For example, if an okay y is returned in Step 8 that is not good, it may be that $\gcd(k_y, R)$ is reasonably small, in which case R is a small multiple of r. If you can only afford to run the quantum portion of the computation once, then in Step 9, you could try computing $a^r, a^{2r}, a^{3r}, \ldots, a^{nr}$ (all mod N) and return the least exponent yielding 1, if there is one. If not, you could try relaxing the distance bound 2^{-2n-1} in (67) to something bigger, in the hope that the y you found, if not okay, is close to okay (if y is not okay, it is more likely than random of being close to one that is). If you can afford to run the quantum computation twice, obtaining r_1 and r_2 respectively in Step 8, then taking the least common multiple $\text{lcm}(r_1, r_2)$ yields R with much higher probability than you can get by running the quantum computation just once. Using ideas like these, it can be shown that one can boost the probability of finding R (using one run of the quantum part of the algorithm) to a positive constant, independent of n.

This concludes the analysis of Shor's algorithm. The only things left are (i) to show how the QFT is implemented efficiently with a quantum circuit, and (ii) describe how Step 8 is implemented by a classical algorithm. We'll take these in reverse order.

17 Week 9: Best rational approximations

The Continued Fraction Algorithm. The book illustrates continued fractions as part of the order-finding algorithm, with Theorem 5.1 on page 229, and Box 5.3 on the next page. We actually don't need to talk about continued fractions explicitly. All we need is to find an efficient classical algorithm to implement Step 7, which we'll do directly now.

For any real numbers $a < b$, there are infinitely many rational numbers in the interval $[a, b]$. We want to find one with smallest denominator and numerator.

Definition 17.1 Let $a, b \in \mathbb{R}$ with $0 < a < b$. Define d to be the least positive denominator of any fraction in $[a, b]$. Now define $n \in \mathbb{Z}$ to be least such that $n/d \in [a, b]$. We call the fraction n/d the *simplest rational interpolant*,[19] or *SRI*, of a and b, and we denote it $\mathrm{SRI}(a, b)$.

The fraction k/r found in Step 7 is just $\mathrm{SRI}((2y-1)/2^{2n+1}, (2y+1)/2^{2n+1})$.

Here is a simple, efficient, recursive algorithm to find $\mathrm{SRI}(a, b)$ for positive rational $a < b$. Each step will include a comment explaining why it is correct.

$\mathrm{SRI}(a, b)$:

Input: Rational numbers a, b with $0 < a < b$, each given in numerator/denominator form, where both numerator and denominator are in binary.

Base Case: If $a \leqslant 1 \leqslant b$, then return $1 = 1/1$. (Clearly, this is the simplest possible fraction!)

First Recursive Case: If $1 < a$, then

1. Let $q = \lceil a - 1 \rceil$ be the largest integer strictly less than a.
2. Recursively compute $r = \mathrm{SRI}(a - q, b - q)$.
3. Return $r + q$.

(Obviously, shifting the interval $[a, b]$ by an integral amount shifts the SRI the same amount. Also note that $q \geqslant a/2$—a fact that will be useful later.)

Second Recursive Case: Otherwise, $b < 1$.

1. Recursively compute $r = \mathrm{SRI}(1/b, 1/a)$.
2. Return $1/r$.

(We claim that if $d'/n' = \mathrm{SRI}(1/b, 1/a)$, then $n'/d' = \mathrm{SRI}(a, b)$. Let $n/d = \mathrm{SRI}(a, b)$. We show that $n'/d' = n/d$. Since $n/d \in [a, b]$, we clearly have $d/n \in [1/b, 1/a]$, and so $n' \leqslant n$ by minimality of n'. Similarly, since $d'/n' \in [1/b, 1/a]$, we have $n'/d' \in [a, b]$, and so $d \leqslant d'$ by minimality of d. Thus we have $n'/d' \leqslant n/d$. Suppose $n'/d' < n/d$. We have $n'/d' \leqslant n'/d \leqslant n/d$, so $n'/d \in [a, b]$ and $d/n' \in [1/b, 1/a]$. We also have either $n'/d' < n'/d$ or $n'/d < n/d$. We can't have the latter, owing to the minimality of n. But we can't have the former, either, for otherwise, $d'/n' > d/n'$, and this contradicts the minimality of d'. Thus we must have $n'/d' = n/d$, and so $\mathrm{SRI}(a, b) = 1/\mathrm{SRI}(1/b, 1/a)$.)

[19] I'm making this term up. I'm sure there must be an official name for it, but I haven't found what it is.

The comments suggest that the SRI algorithm is correct as long as it halts. It does halt, and quickly, too. Let the original inputs be $a = a_0 = n_0/d_0$ and $b = N_0/D_0$, given as fractions in lowest terms (n_0, d_0, N_0, and D_0 are all positive integers). Similarly, for $0 < k$, let $a_k = n_k/d_k$ and $b_k = N_k/D_k$ be respectively the first and second argument to the kth recursive call to SRI. We consider the product $P_k := n_k d_k N_k D_k$ and how it changes with k. If the kth recursive call occurs in the second case, then the numerators and denominators are simply swapped, so $P_k = P_{k-1}$. If the kth recursive call occurs in the first case, then $d_k = d_{k-1}$ and $D_k = D_{k-1}$, but (letting $q := \lceil a_{k-1} - 1 \rceil$)

$$n_k = n_{k-1} - q d_{k-1} \leqslant n_{k-1} - (a_{k-1}/2) d_{k-1} \leqslant n_{k-1}/2 \,,$$

and

$$N_k = N_{k-1} - q D_{k-1} < N_{k-1} \,,$$

because $q \geqslant a_{k-1}/2$. Thus in this case, $P_k < P_{k-1}/2$. The two recursive cases alternate, so P_k decreases by at least half with every other recursive call. Since $P_k > 0$, we must hit the base case after at most $2 \lg P_0 = 2(\lg n_0 + \lg d_0 + \lg N_0 + \lg D_0)$ recursive calls. For each $k \geqslant 0$, $\lg P_k$ approximates the size of the input (in bits) up to an additive constant, and this size never increases from call to call, so the whole algorithm is clearly polynomial time.

Exercise 17.2 What is SRI(7/25, 3/10)?

Exercise 17.3 (Challenging) Using your favorite programming language, implement the SRI algorithm above. You can decide to accept either exact rational or floating point inputs.

Implementing the QFT. Recall that for all $x \in \mathbb{Z}_{2^n}$,

$$\mathrm{QFT}_n |x\rangle = \frac{1}{2^{n/2}} \sum_{y \in \mathbb{Z}_{2^n}} e_n(xy) |y\rangle.$$

It was Peter Shor who first showed how to implement QFT_n efficiently with a quantum circuit, in the same paper as his factoring algorithm. The following recursive description is taken from Cleve & Watrous (2000). When $n = 1$, you can easily check that $\mathrm{QFT}_1 = H$, *i.e.*, the one-qubit Hadamard gate. Now suppose that $n > 1$ and let $1 \leqslant m < n$ be an integer. QFT_n can be decomposed into a circuit using QFT_{n-m}, QFT_m, and two other subcircuits, as shown in Figure 8. The $P_{n,m}$ gate acts on two numbers—an $(n-m)$-bit number $x \in \mathbb{Z}_{2^{n-m}}$ and an m-bit number $y \in \mathbb{Z}_{2^m}$—such that

$$P_{n,m} |x, y\rangle = e_n(xy) |x, y\rangle.$$

That is, $P_{n,m}$ adjusts the phase of $|x, y\rangle$ by $e^{2\pi i xy/2^n}$. (The P stands for "phase.") In the figure, x is fed to $P_{n,m}$ in the upper $n-m$ qubits, and y in the lower m qubits. We'll see shortly how $P_{n,m}$ can be implemented in terms of simple gates. The unnamed gate on the far right of the figure merely serves to move the qubits around, bringing the top $n-m$ qubits to the bottom and bringing the bottom m qubits to the top.[20] These qubit-permuting gates can be left out when recursively

[20]This, as well as any other rearrangement of qubits, can always be achieved by two layers of swap gates. Proving this makes a great exercise.

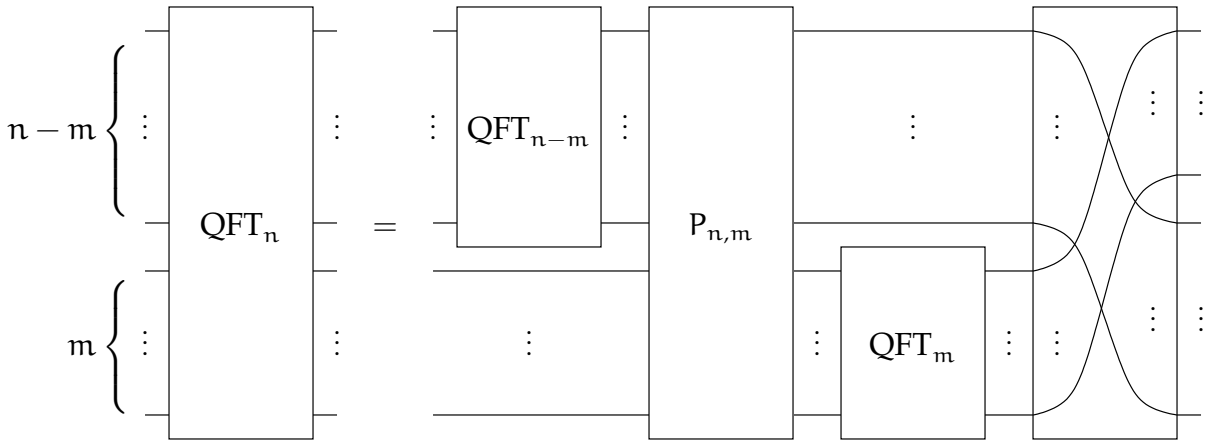


Figure 8: QFT_n in terms of QFT_{n-m} and QFT_m.

expanding QFT_{n-m} and QFT_m, as long as you keep track of which qubit is which and adjust the elementary gates accordingly.

Many recursive decompositions are possible, based on the choice of m at each stage. Shor's original circuit for QFT_n is obtained by recursively decomposing with $m = 1$ throughout. A smaller depth circuit is achieved by a divide-and-conquer approach, letting m be roughly $n/2$ each time.

Let's check that the decomposition of Figure 8 is correct. Given any n-bit number $x \in \mathbb{Z}_{2^n}$, we split its binary representation into its $n - m$ high-order bits $x_h \in \mathbb{Z}_{2^{n-m}}$ and its m low-order bits $x_\ell \in \mathbb{Z}_{2^m}$. So we have $x = x_h 2^m + x_\ell$, and we may write the state $|x\rangle$ as $|x_h, x_\ell\rangle$ or as $|x_h\rangle|x_\ell\rangle$. Applying QFT_n to $|x\rangle$ gives

$$\mathrm{QFT}_n|x\rangle = \frac{1}{2^{n/2}} \sum_{y \in \mathbb{Z}_{2^n}} e_n(xy)|y\rangle = \frac{1}{2^{n/2}} \sum_y e_n((x_h 2^m + x_\ell)y)|y\rangle. \tag{78}$$

Expressing each y as $y_h 2^{n-m} + y_\ell$ for unique $y_h \in \mathbb{Z}_{2^m}$ and $y_\ell \in \mathbb{Z}_{2^{n-m}}$, (78) becomes

$$\frac{1}{2^{n/2}} \sum_y e_n((x_h 2^m + x_\ell)(y_h 2^{n-m} + y_\ell))|y\rangle = \frac{1}{2^{n/2}} \sum_y e_{n-m}(x_h y_\ell) e_m(x_\ell y_h) e_n(x_\ell y_\ell)|y\rangle. \tag{79}$$

(Notice that there is no $x_h y_h$ exponent, since it is multiplied by 2^n.) Now let's see what happens when the right-hand circuit of Figure 8 acts on $|x\rangle$. We have

$$\begin{aligned}
|x\rangle \quad &= \quad |x_h\rangle|x_\ell\rangle \\
&\overset{\mathrm{QFT}_{n-m}}{\longmapsto} \quad \frac{1}{2^{(n-m)/2}} \sum_{y_\ell \in \mathbb{Z}_{2^{n-m}}} e_{n-m}(x_h y_\ell)|y_\ell\rangle|x_\ell\rangle \\
&\overset{P_{n,m}}{\longmapsto} \quad \frac{1}{2^{(n-m)/2}} \sum_{y_\ell} e_{n-m}(x_h y_\ell) e_n(y_\ell x_\ell)|y_\ell\rangle|x_\ell\rangle \\
&\overset{\mathrm{QFT}_m}{\longmapsto} \quad \frac{1}{2^{n/2}} \sum_{y_\ell} \sum_{y_h \in \mathbb{Z}_{2^m}} e_{n-m}(x_h y_\ell) e_n(y_\ell x_\ell) e_m(x_\ell y_h)|y_\ell\rangle|y_h\rangle
\end{aligned}$$

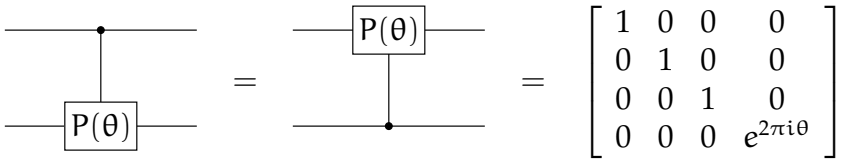
$$\begin{aligned} &\longmapsto \quad \frac{1}{2^{n/2}} \sum_{y_\ell} \sum_{y_h} e_{n-m}(x_h y_\ell) e_n(y_\ell x_\ell) e_m(x_\ell y_h) |y_h\rangle |y_\ell\rangle \\ &= \quad \frac{1}{2^{n/2}} \sum_{y \in \mathbb{Z}_2^n} e_{n-m}(x_h y_\ell) e_m(x_\ell y_h) e_n(x_\ell y_\ell) |y\rangle, \end{aligned}$$

where we set $y := y_h 2^{n-m} + y_\ell$ as before. The last arrow represents the action of the qubit-permuting gate. The final state is evidently the same as in (79), so the two circuits are equal.

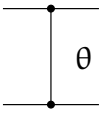
Finally, we get to implementing the $P_{n,m}$ gate. We'll implement $P_{n,m}$ entirely using controlled phase-shift gates. For any $\theta \in \mathbb{R}$, define the conditional phase-shift gate as

$$P(\theta) := e^{\pi i \theta} R_z(2\pi\theta) = \begin{bmatrix} 1 & 0 \\ 0 & e^{2\pi i \theta} \end{bmatrix}.$$

For example, $I = P(1)$, $Z = P(1/2)$, $S = P(1/4)$, and $T = P(1/8)$. For the controlled $P(\theta)$ gate—the C-$P(\theta)$ gate—we clearly have



Owing to the symmetry between the control and target qubits, we will display this gate as



where we place the value θ somewhere nearby. Our θ-values will always be of the form 2^{-k} for integers $k > 0$. Notice that for any $a, b \in \mathbb{Z}_2$,

$$\text{C-}P(2^{-k})|a\rangle|b\rangle = e_k(ab)|a\rangle|b\rangle. \tag{80}$$

It is easiest to think of $P_{n,m}$ as acting on two quantum registers—the first with $n - m$ qubits and the second with m qubits. What gates do we need to implement $P_{n,m}$? Let's consider $P_{n,m}$ applied to the state $|x\rangle|y\rangle = |x_1 x_2 \cdots x_{n-m}\rangle|y_1 y_2 \cdots y_m\rangle$, where $x_1, \ldots, x_{n-m}$ and $y_1, \ldots, y_m$ are all bits in $\mathbb{Z}_2$. We have

$$\frac{x}{2^{n-m}} = 0.x_1 x_2 \cdots x_{n-m} = \sum_{j=1}^{n-m} x_j 2^{-j} \qquad \text{and} \qquad \frac{y}{2^m} = 0.y_1 y_2 \cdots y_m = \sum_{k=1}^{m} y_k 2^{-k},$$

where the "decimal" expansions are actually base 2. Multiplying these two quantities gives

$$\frac{xy}{2^n} = \sum_{j=1}^{n-m} \sum_{k=1}^{m} x_j y_k 2^{-j-k},$$

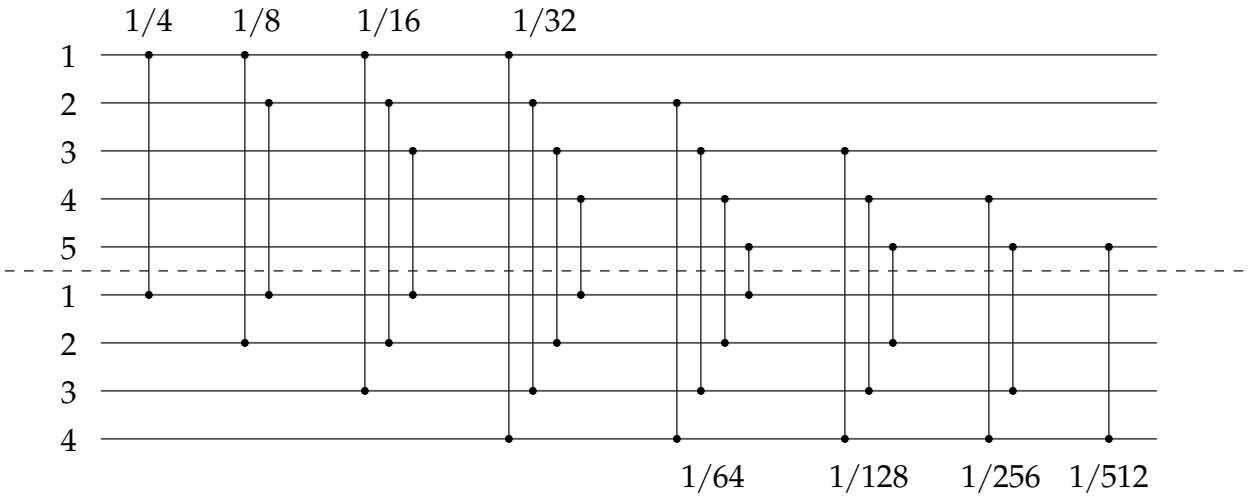


Figure 9: The circuit implementing $P_{9,4}$. C-$P(\theta)$ gates are grouped according to the values of θ. Within each group, gates act on disjoint pairs of qubits, so they can form a single layer of gates acting in parallel.

and so

$$e_n(xy) = \exp(2\pi i xy/2^n) = \prod_{j,k} \exp(2\pi i x_j y_k 2^{-j-k}) = \prod_{j,k} e_{j+k}(x_j y_k).$$

And thus we get

$$P_{n,m}|x\rangle|y\rangle = \left(\prod_{j,k} e_{j+k}(x_j y_k)\right)|x\rangle|y\rangle.$$

Recalling (80), notice that for each j and k, we can get the (j,k)th factor in the product above if we connect the jth qubit of the first register (carrying x_j) with the kth qubit of the second register (carrying y_k) with a C-$P(2^{-j-k})$ gate (which then acts on the state $|x_j y_k\rangle$ to get an overall phase contribution of $e_{j+k}(x_j y_k)$). So to implement $P_{n,m}$ we just need to do this for all $1 \leqslant j \leqslant n-m$ and all $1 \leqslant k \leqslant m$. That's it. All these gates will combine to give the correct overall phase shift of $e_n(xy)$. The order of the gates does not matter, because they all commute with each other (they are all diagonal matrices in the computational basis). For example, Figure 9 shows the $P_{9,4}$ circuit.

Exercise 17.4 Give two complete decompositions of QFT_4 as circuits, the first using $m = 1$ throughout, and the second using $m = 2$ for the initial decomposition. Both circuits should use only H and C-$P(2^{-k})$ gates for $k > 0$. Do not cross wires except at the end of the entire circuit, that is, shift any wire crossings in the recursive QFT circuits to the end of the overall circuit.

In Exercise 17.4, if you correctly shifted all the wire crossings to the end of the circuit, you may have noticed that in the end you simply reverse the order of the qubits. That is not a coincidence; an easy inductive argument shows that this must always be the case.

Exercise 17.5 (Challenging) Asymptotically, what is the size (number of elementary gates) of QFT_n when decomposed using $m = 1$ throughout (Shor's circuit)? What is the size using the divide-and-conquer method with $m = n/2$ throughout? The same questions for the depth (minimum possible number of layers of gates acting on disjoint sets of qubits). Use big-O notation. In all cases, you can ignore the qubit-permuting gates. [Hint: Find recurrence equations satisfied by the size and the depth in each case.]

Actually, there is another way to implement $P_{n,m}$: Classically compute xy as an n-bit binary integer, then for each $k \in \{1, 2, \ldots, n\}$, send the kth qubit of the result through the gate $P(2^{-k})$. There are fast parallel circuits for multiplication, with polynomial size and depth $O(\lg n)$. This log-depth implementation of $P_{n,m}$ together with the divide-and-conquer decomposition method for QFT_n give an $O(n)$-depth, polynomial-size circuit that exactly implements QFT_n.

18 Week 9: Approximate QFT

Exact versus Approximate. The QFT_n circuit we described above for Shor's algorithm blithely uses C-$P(2^{-k})$ gates where k ranges between 2 and n. If Shor's algorithm is to significantly outperform the best classical factoring algorithms, then n must be on the order of 10^3 and above, which means that we will be using gates that produce conditional phase shifts of $2\pi/2^{1000}$ or less. No one in their right mind imagines that we could ever tune our instruments so precisely as to produce so small a phase shift, which is required for any exact implementation of QFT_{1000}. The bottom line is that implementing QFT_n exactly for large n will just never be feasible.

Fortunately, an exact implementation is unnecessary for Shor's algorithm or for any other probabilistic quantum algorithm that uses the QFT. We can actually tolerate a lot of imprecision in the implementation of the C-$P(2^{-k})$ gates. In fact, if $k \gg \lg n$, then C-$P(2^{-k})$ is close enough to the identity operator that we can omit these gates entirely. The resulting circuit is *much* smaller and produces a good approximation to QFT_n that can be used in Shor's algorithm. Good enough so that the probability of finding R in Step 7 of the algorithm is at worst only slightly smaller than with the exact implementation, thus requiring only a few more repetitions of the algorithm to produce R with high probability.

In the next few topics, we'll make this all quantitative. The concepts and techniques we introduce are useful in other contexts. Before we do, we need a basic inequality known as the *Cauchy-Schwarz inequality*.

The Cauchy-Schwarz inequality. We mentioned this inequality early in the course as proving the triangle inequality for complex scalars, but this is the first time since then that we actually need it. We'll use it here to bound the effects of unitary errors in implementing a quantum circuit. We'll use it again in other contexts.

Theorem 18.1 (Cauchy-Schwarz Inequality) *Let $\mathcal{H}$ be a Hilbert space. For any vectors $u, v \in \mathcal{H}$,*

$$|\langle u, v\rangle| \leqslant \|u\| \cdot \|v\|,$$

with equality holding if and only if u and v are linearly dependent.

Proof. There are many ways to prove this theorem. The Nielsen & Chuang textbook has a proof in Box 2.1 on page 68, which we loosely paraphrase here. See Section B.1 of the background material in Appendix B for another proof. Equality clearly holds if u and v are linearly dependent, since then one vector is a scalar multiple of the other. So assume that u and v are linearly independent. By the Gram-Schmidt procedure, we can find orthonormal vectors b_1, b_2 such that $b_1 = u/\|u\|$ and $b_2 = (v - \langle b_1, v\rangle b_1)/\|v - \langle b_1, v\rangle b_1\|$. We thus have

$$\begin{aligned} u &= a b_1, \\ v &= c b_1 + d b_2, \end{aligned}$$

for some $a, c, d \in \mathbb{C}$ with $a > 0$ and $d > 0$. We now get

$$\|u\| \cdot \|v\| = a(|c|^2 + d^2)^{1/2} > a(|c|^2)^{1/2} = a|c| = |ac| = |\langle a b_1, c b_1 + d b_2\rangle| = |\langle u, v\rangle| \,.$$

□

Exercise 18.2 Show that $\|u + v\| \leqslant \|u\| + \|v\|$ for any two vectors $u, v \in \mathcal{H}$, with equality holding if and only if one is a nonnegative scalar times the other. This is another example of a triangle inequality. [Hint: Use Cauchy-Schwarz (Theorem 18.1) and the fact that $\Re[z] \leqslant |z|$ for any $z \in \mathbb{C}$.]

A Hilbert Space Is a Metric Space. For any two vectors $u, v \in \mathcal{H}$, the *Euclidean distance* between u and v is defined as

$$d(u, v) := \|u - v\|.$$

The function d satisfies the following axioms:

1. $d(u, v) \geqslant 0$,
2. $d(u, v) = 0$ iff $u = v$,
3. $d(u, v) = d(v, u)$, and
4. $d(u, v) \leqslant d(u, w) + d(w, v)$ for any $w \in \mathcal{H}$.

These are the axioms for a *metric* on the set $\mathcal{H}$. The last item is known as the *triangle inequality*, which can be seen as follows:

$$d(u, v) = \|u - v\| = \|u - w + w - v\| \leqslant \|u - w\| + \|w - v\| = d(u, w) + d(w, v),$$

where the inequality follows from Exercise 18.2. All the other axioms are straightforward.

Suppose that you could run an ideal quantum algorithm to produce a state $|\psi\rangle$ that you then subject to some projective measurement. You would get certain probabilities for the various possible outcomes. Suppose instead that you actually ran an imperfect implementation of the algorithm and produced a state $|\varphi\rangle$ that was close to $|\psi\rangle$ in Euclidean distance, and you subjected $|\varphi\rangle$ to the same projective measurement. The next proposition shows that the probabilities of the outcomes are close to those of the ideal situation.

Proposition 18.3 *Let $\{P_a : a \in \mathcal{I}\}$ be some complete set of orthogonal projectors on $\mathcal{H}$. Let $u, v \in \mathcal{H}$ be any two unit vectors, and let $\Pr_u[a]$ and $\Pr_v[a]$ be the probability of seeing outcome $a \in \mathcal{I}$ when measuring the state u and v respectively using this complete set. Then for every outcome $a \in \mathcal{I}$,*

$$|P_u[a] - \Pr_v[a]| \leqslant 2d(u, v).$$

Proof. We have

$$\begin{aligned}
|\Pr_u[a] - \Pr_v[a]| &= |u^* P_a u - v^* P_a v| \\
&= |u^* P_a u - u^* P_a v + u^* P_a v - v^* P_a v| \\
&= |u^* P_a (u - v) + (u - v)^* P_a v| \\
&\leqslant |u^* P_a (u - v)| + |(u - v)^* P_a v| \\
&= |\langle P_a u, u - v\rangle| + |\langle u - v, P_a v\rangle| \\
&\leqslant \|P_a u\| \cdot \|u - v\| + \|u - v\| \cdot \|P_a v\| \\
&\leqslant 2\|u - v\|.
\end{aligned}$$

The second inequality is an application of Cauchy-Schwarz (Theorem 18.1); the third follows from the fact that $\|Pw\| \leqslant \|w\| = 1$ for any projector P and unit vector w (see Exercise 5.12). □

The next definition extends the notion of distance to operators. Here we give one of many possible ways to do this.

Definition 18.4 Let $A \in \mathcal{L}(\mathcal{H})$ be an operator. The *operator norm* of A is defined as

$$\|A\| := \sup_{v \in \mathcal{H} : \|v\| = 1} \|Av\| = \sup_{v \neq 0} \frac{\|Av\|}{\|v\|}.$$

This norm is also sometimes called the ℓ_∞*-norm* on operators.

$\|A\|$ is thus the maximum of $\|Av\|$ taken over all unit vectors v. Don't confuse $\|A\|$, which is a scalar, with $|A| = \sqrt{A^* A}$, which is an operator. It can be shown that the maximum is actually achieved by some vector, *i.e.*, there is always a unit vector v such that $\|A\| = \|Av\|$. Here are some basic properties of the operator norm that follow quickly from the definition:

1. $\|A\| \geqslant 0$, with $\|A\| = 0$ iff $A = 0$.
2. $\|zA\| = |z| \cdot \|A\|$ for any scalar $z \in \mathbb{C}$.
3. $\|A + B\| \leqslant \|A\| + \|B\|$, for any $B \in \mathcal{L}(\mathcal{H})$.
4. $\|I\| = 1$, where I is the identity operator.
5. $\|UA\| = \|AU\| = \|A\|$ for any unitary $U \in \mathcal{L}(\mathcal{H})$.
6. $\|Av\| \leqslant \|A\| \cdot \|v\|$ for any $v \in \mathcal{H}$.
7. $\|AB\| \leqslant \|A\| \cdot \|B\|$ for any $B \in \mathcal{L}(\mathcal{H})$.

8. $\|A\| = \| \, |A| \, \|$.

Exercise 18.5 Verify each of these items, based on the definition of $\|\cdot\|$.

We can use the operator norm to define a metric d on $\mathcal{L}(\mathcal{H})$ just as we did with $\mathcal{H}$.

Definition 18.6 For $A, B \in \mathcal{L}(\mathcal{H})$ define

$$d(A, B) := \|A - B\|,$$

the *operator distance* between A and B.

Picking up on the last item, above, we see that A has the same norm as $|A|$. Since $|A| \geqslant 0$, there is an eigenbasis $\{b_1, \ldots, b_n\}$ of $|A|$ with respect to which $|A| = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_n \geqslant 0$ are the eigenvalues of $|A|$. We claim that $\|A\| = \lambda_1$, *i.e.*, $\|A\|$ is the largest eigenvalue of $|A|$. To see why, let $v = (v_1, \ldots, v_n)$ be any unit column vector with respect to this basis $\{b_j\}_{1 \leqslant j \leqslant n}$. Then we have

$$\|Av\|^2 = \langle Av, Av \rangle = \sum_{j=1}^{n} \lambda_j^2 |v_j|^2 = \sum_j \lambda_j^2 a_j,$$

where we set $a_j := |v_j|^2$. We have $a_j \geqslant 0$ for all $1 \leqslant j \leqslant n$, and since v is a unit vector, we have $\sum_j a_j = 1$. So,

$$\begin{aligned} \|Av\|^2 &= \sum_{j=1}^{n} \lambda_j^2 a_j \\ &= \lambda_1^2 a_1 + \sum_{j=2}^{n} \lambda_j^2 a_j \\ &= \lambda_1^2 \left(1 - \sum_{j=2}^{n} a_j \right) + \sum_{j=2}^{n} \lambda_j^2 a_j \\ &= \lambda_1^2 + \sum_{j=2}^{n} (\lambda_j^2 - \lambda_1^2) a_j. \end{aligned}$$

Since $\lambda_j^2 - \lambda_1^2 \leqslant 0$ for all $2 \leqslant j \leqslant n$, the right-hand side is clearly maximized by setting $a_2 = \cdots = a_n = 0$ (and so $a_1 = 1$). So we must have $\|A\| = \| \, |A| \, \| = \||A|b_1\| = \lambda_1$ as claimed.

The next property follows from the claim, above.

9. If A and B are operators (not necessarily over the same space), then $\|A \otimes B\| = \|A\| \cdot \|B\|$. In particular, $\|A \otimes I\| = \|A\|$ and $\|I \otimes B\| = \|B\|$.

This property is useful when we take the norm of a single gate in a circuit. The unitary operator corresponding to the action of the gate is generally of the form $U \otimes I$, where U corresponds to the

gate acting on the space of its own qubits, and the identity I acts on the qubits not involved with the gate. Property 9 says that we can ignore the I when taking the norm of this operator.

To prove Property 9, we first prove that $|A \otimes B| = |A| \otimes |B|$. To show this, we only need to verify two things: (i) $(|A| \otimes |B|)^2 = (A \otimes B)^*(A \otimes B)$ and (ii) $|A| \otimes |B| \geqslant 0$. We leave (i) as an exercise. For (ii), we first pick eigenbases for $|A|$ and $|B|$, respectively. Then if $|A| = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with respect to the first basis and $|B| = \mathrm{diag}(\mu_1, \ldots, \mu_m)$ with respect to the second, then with respect to the product of the two bases (itself an orthonormal basis), $|A| \otimes |B|$ is a diagonal matrix whose diagonal entries are $\lambda_j \mu_k$ for all $1 \leqslant j \leqslant n$ and $1 \leqslant k \leqslant m$. Since all the λ_j and μ_k are nonnegative, the diagonal entries of $|A| \otimes |B|$ are all nonnegative. Hence, $|A| \otimes |B| \geqslant 0$, which proves (ii), and thus $|A \otimes B| = |A| \otimes |B|$. Now the largest eigenvalue of $|A| \otimes |B|$ is clearly $\lambda\mu$, where $\lambda = \max(\lambda_1, \ldots, \lambda_n) = \|A\|$ and $\mu = \max(\mu_1, \ldots, \mu_m) = \|B\|$ by the claim. Since $|A| \otimes |B| = |A \otimes B|$, the product $\lambda\mu$ is also the largest eigenvalue of $|A \otimes B|$, and so using the claim again, we get Property 9.

Exercise 18.7 Verify by direct calculation that $(|A| \otimes |B|)^2 = (A \otimes B)^*(A \otimes B)$.

While we're on the subject, one more property of the operator norm will find use later on. If you want, you can skip down to after the proof of Claim 18.8, below, and refer back to it later when you need to.

10. $\|A^*\| = \|A\|$ for any operator A.

This property follows immediately from the following claim:

Claim 18.8 *For any operator A, the operators $|A|$ and $|A^*|$ are unitarily conjugate, i.e., there is a unitary operator U such that $|A^*| = U|A|U^*$.*

Since unitarily conjugate operators have the same spectrum, Claim 18.8 implies that $|A|$ and $|A^*|$ have the same largest eigenvalue, *i.e.*, $\|A\| = \|A^*\|$. Claim 18.8 itself follows from a fundamental decomposition theorem known as the *polar decomposition*. For a proof of this decomposition, see Section B.3 in Appendix B. The polar decomposition is closely related (in fact, equivalent) to the *singular value decomposition*, which is also proved in Section B.3.

Theorem 18.9 (Polar Decomposition, Theorem B.8 in Section B.3) *For every operator A there is a unitary U such that $A = U|A|$. In fact, $|A|$ is the unique positive operator H such that $A = UH$ for some unitary U.*

If $z \in \mathbb{C}$ is a scalar, then obviously $z = u|z|$ for some $u \in \mathbb{C}$ with unit norm (*i.e.*, a phase factor). Furthermore, $|z|$ is the unique nonnegative real factor in any such decomposition, and if $z \neq 0$ then u is unique as well. Theorem 18.9 generalizes this fact to operators in an analogous way. (If A is nonsingular (invertible), then U is unique as well: it can be easily shown that if A is nonsingular then $|A|$ is nonsingular, whence $U = A|A|^{-1}$.)

Proof of Claim 18.8. Let A be an operator. By the polar decomposition (Theorem 18.9), there is a unitary U such that $A = U|A|$. We have, using Exercise 9.32,

$$|A^*| = \sqrt{AA^*} = \sqrt{U|A|^2U^*} = U\sqrt{|A|^2}U^* = U|A|U^*.$$

□

Now we consider an arbitrary idealized quantum circuit C with m many unitary gates, which basically consists of a succession of unitary operators $U_1, \ldots, U_m$ applied to some initial state $|\text{init}\rangle$, producing the state $|\psi\rangle = U_m \cdots U_1|\text{init}\rangle$, which is then projectively measured somehow. When implementing C we might implement each gate U_j imperfectly, getting some unitary V_j instead, where hopefully, V_j is close to U_j. I will call this a *unitary error*. The actual circuit produces the state $|\psi'\rangle = V_m \cdots V_1|\text{init}\rangle$. Assuming $d(U_j, V_j) \leqslant \varepsilon$ for all $1 \leqslant j \leqslant m$, what can we say about $d(|\psi\rangle, |\psi'\rangle)$?

Classical calculations are often numerically unstable, and errors may compound multiplicatively. Fortunately for us, unitary errors only compound additively rather than multiplicatively, so we can tolerate a fair amount of imperfection in our gates—only $O(\lg n)$ bits of precision per gate for a circuit with a polynomially bounded (in n) number of gates.

Back to the question above. Using the basic properties of the operator norm listed above, we get

$$\begin{aligned} d\left(|\psi\rangle, |\psi'\rangle\right) &= \|(U_m \cdots U_1 - V_m \cdots V_1)|\text{init}\rangle\| \\ &\leqslant \|U_m \cdots U_1 - V_m \cdots V_1\| \cdot \||\text{init}\rangle\| \\ &= \|U_m \cdots U_1 - V_m \cdots V_1\|. \end{aligned}$$

The operator inside the $\|\cdot\|$ on the right can be expressed as a telescoping sum:

$$U_m \cdots U_1 - V_m \cdots V_1 = \sum_{k=1}^{m} U_m \cdots U_{k+1}(U_k - V_k)V_{k-1} \cdots V_1. \tag{81}$$

Therefore,

$$\begin{aligned} \|U_m \cdots U_1 - V_m \cdots V_1\| &= \left\| \sum_{k=1}^{m} U_m \cdots U_{k+1}(U_k - V_k)V_{k-1} \cdots V_1 \right\| \\ &\leqslant \sum_k \|U_m \cdots U_{k+1}(U_k - V_k)V_{k-1} \cdots V_1\| \\ &= \sum_k \|U_k - V_k\| \\ &\leqslant \sum_k \varepsilon \\ &= m\varepsilon, \end{aligned}$$

and so $d(|\psi\rangle, |\psi'\rangle) \leqslant m\varepsilon$.

Suppose we want the probability of some outcome to differ from the ideal probability by no more than some $\delta > 0$. Then by Proposition 18.3, it suffices that $2m\varepsilon \leqslant \delta$, or that

$$\varepsilon \leqslant \frac{\delta}{2m}.$$

For example, the entire quantum circuit for Shor's algorithm has size polynomial in n—let's say at most cn^k gates for some constants c and k. (I'm not sure, but I believe that $k \leqslant 3$. The dominant contribution is not the QFT but rather the classical modular exponentiation circuit.) The algorithm produces a good y (one that will lead to finding R) with probability at least $4/(\pi^2 n)$, ignoring a quadratically small correction term. We could settle instead for a success probability of at least $2/(\pi^2 n)$, say, which would require up to twice as many trials on average for success. But then, choosing $\delta := 4/(\pi^2 n) - 2/(\pi^2 n) = 2/(\pi^2 n)$, we could implement each gate to within an error (operator distance) of

$$\varepsilon_{\text{Shor}} := \frac{2/(\pi^2 n)}{2cn^k} = \frac{1}{\pi^2 cn^{k+1}} = \Theta(n^{-k-1})$$

away from the ideal. This has major implications for the QFT part of the circuit. The QFT has size $\Theta(n^2)$, uses n Hadamard gates, and the rest of the gates are C-P(2^{-j}) gates, where $2 \leqslant j \leqslant n$. (We can do without the swap gates by keeping track of which qubit is which, and rearranging the bits of the y value that we measure.) Note that for any $\theta \in \mathbb{R}$,

$$\text{C-P}(\theta) - I = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{2\pi i\theta} - 1 \end{bmatrix} = 2ie^{i\pi\theta} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sin(\pi\theta) \end{bmatrix}.$$

It follows that

$$d(\text{C-P}(\theta), I) = \|\text{C-P}(\theta) - I\| = 2|\sin(\pi\theta)| \leqslant 2\pi\theta.$$

This means that if $2\pi 2^{-j} \leqslant \varepsilon_{\text{Shor}}$, or equivalently,

$$j \geqslant \lg(2\pi/\varepsilon_{\text{Shor}}) = (k+1)\lg n + O(1),$$

then any C-P(2^{-j}) in the QFT circuit is close enough to I that we can just omit it. It's easy to see that *most* of the QFT gates are like this and can be omitted, shrinking the QFT portion of the circuit from quadratic size to linear size in n. This fact was first observed by Coppersmith.

For $n = 10^3$ and assuming $k = 3$ we can get by with implementing each gate with error $O(n^{-4})$, which is on the order of one part per trillion. This is still a very tall order, but unlike 2^{-1000} it is at least close to the realm of sanity. Optimizing other aspects of Shor's algorithm and its analysis increases the error tolerance considerably.

19 Midterm Exam

Do all problems. Hand in your anwsers in class on Wednesday, March 28, just as you would a homework problem set. The only difference between this exam and the homeworks is that you may not discuss exam questions or answers with anyone inside or outside of class except me. It goes without saying that if you do, you have cheated and I'll have to summarily fail you, which is my usual policy about cheating. I know you won't, though, so I'll sleep well at night.

All questions with Roman numerals carry equal weight, but may not be of equal difficulty.

Recall that for two vectors or operators a, b, we say that $a \propto b$ if there is a phase factor $e^{i\theta}$ where $\theta \in \mathbb{R}$ such that $a = e^{i\theta} b$.

I) (Rotating the Bloch sphere) Find a unit vector $\hat{n} = (x, y, z) \in \mathbb{R}^3$ on the Bloch sphere and an angle $\varphi \in [0, 2\pi)$ such that

$$\begin{aligned} R_{\hat{n}}(\varphi)|+x\rangle &\propto |+y\rangle, \\ R_{\hat{n}}(\varphi)|+y\rangle &\propto |+z\rangle, \end{aligned}$$

where $R_{\hat{n}}(\varphi)$ is defined in Exercise 9.4, and $|+x\rangle$, $|+y\rangle$, and $|+z\rangle$ are given by Equations (18–20). Give the 2×2 matrix corresponding to your solution in the standard computational basis, simplified as much as possible. What can you say about $R_{\hat{n}}(\varphi)|+z\rangle$? There are exactly two possible solutions to this problem.

II) (Phase factors and density operators) Let U and V be unitary operators over $\mathcal{H}$. It is easy to see that if $U \propto V$, then $U\rho U^* = V\rho V^*$ for every state ρ. (Here, by "state" we mean a state in the density operator formalism, i.e., a one-dimensional projection operator of the form $|\psi\rangle\langle\psi|$ for some unit vector $|\psi\rangle$.) Show the converse: If U and V are unitary and $U\rho U^* = V\rho V^*$ for all states ρ, then $U \propto V$. [Hint: Consider U and V in matrix form and show that every entry of U is equal to the corresponding entry of V multiplied by the same phase factor. Use the equation above for specific values of ρ. This technique is similar to that used in Exercise 9.26.]

III) (Tensor products of matrices) Let A be an arbitrary $n \times n$ matrix and let B be an arbitrary $m \times m$ matrix.

(a) If A and B are both upper triangular, explain why $A \otimes B$ is also upper triangular.

(b) Suppose that A has eigenvalues $\lambda_1, \ldots, \lambda_n$ (with multiplicities), and that B has eigenvalues $\mu_1, \ldots, \mu_m$ (with multiplicities). Describe the eigenvalues of $A \otimes B$. Note that here, A and B are not necessarily upper triangular. [Hint: Use the previous item and things we know about the eigenvalues of upper triangular matrices.]

IV) (Teleportation gone wrong) Alice and Bob think they are sharing a pair of qubits in the state $|\Phi^+\rangle$, but instead the pair of qubits that they share is in one of the other three Bell states. Suppose that they now attempt to do the standard one-qubit teleportation protocol to teleport the state $|\psi\rangle$ from Alice to Bob using this pair.

(a) Show that the state that Bob possesses at the end is, up to a phase factor, some Pauli operator (X, Y, or Z) applied to $|\psi\rangle$. [Hint: You can save yourself a lot of calculation by observing that the four Bell states are of the form $(I \otimes \sigma)|\Phi^+\rangle$ for $\sigma \in \{I, X, Z, XZ\}$.]

(b) Supposing Alice and Bob know that they share a pair of qubits in the state $|\Psi^-\rangle$, show how they can alter their protocol to faithfully teleport $|\psi\rangle$. [Hint: Use the previous item.]

V) (A black-box problem) Suppose $f : (\mathbb{Z}_2)^n \to \mathbb{Z}_2$ is such that there is some $s = s_1 \cdots s_n \in (\mathbb{Z}_2)^n$ such that for all $x = x_1 \cdots x_n \in (\mathbb{Z}_2)^n$,

$$f(x) = s \cdot x,$$

Where $s \cdot x = \left(\sum_{j=1}^{n} s_j x_j\right) \bmod 2$ is the standard dot product of s and x over $\mathbb{Z}_2$. Recall the inversion gate I_f such that

$$I_f|x\rangle = (-1)^{f(x)}|x\rangle$$

for all $x \in (\mathbb{Z}_2)^n$. The following describes a circuit that uses I_f once to find s:

(a) Initialize an n-qubit register in the state $|0^n\rangle$.

(b) Apply a Hadamard gate H to each of the n qubits. (This is a single layer.)

(c) Apply I_f to the n qubits.

(d) Apply a Hadamard gate H to each of the n qubits. (This is a single layer.)

(e) Measure the n qubits in the computational basis, obtaining some $y \in (\mathbb{Z}_2)^n$.

Do the following:

(a) Draw the circuit described above.

(b) Give the state of the n qubits after each unitary gate—or layer of gates—is applied.

(c) Show that $y = s$ with certainty.

(d) Show how to find s classically by evaluating f on exactly n elements of $(\mathbb{Z}_2)^n$.

20 Week 10: Grover's algorithm

Quantum Search. You are given an array $A[1 \ldots N]$ of N values, one of which is a recognizable target value t. You want to find the position w of t in the list. The values are not necessarily sorted or arranged in any particular way. Classically, the best you can do in the worst case is to probe all $A[j]$ for $1 \leqslant j \leqslant N$, and find the target on the last probe. On average, you will need about $N/2$ probes before finding the target with high probability.

With a quantum algorithm, you can find the target with (extremely) high probability using only $O(\sqrt{N})$ many probes, giving a quadratic speed-up. This result is due to Lov Grover, and is known as *Grover's quantum search algorithm*. It has many variants, but we only give the simplest one here to give an idea of how it works.

We assume that $N = 2^n$ for some n and that we have a black-box Boolean function $f : \{0,1\}^n \to \{0,1\}$ available such that there is a unique $w \in \{0,1\}^n$ such that $f(w) = 1$ and $f(z) = 0$ for all $z \neq w$. Think of f as the target detector. Our task is to find w.

We assume that we can use n-qubit I_f gates, where we recall that

$$I_f|z\rangle = (-1)^{f(z)}|z\rangle.$$

In the present case, we have $I_f|w\rangle = -|w\rangle$ and $I_f|z\rangle = |z\rangle$ if $z \neq w$. Note that given the promise about f, we have $I_f = \mathrm{diag}(1, \ldots, 1, -1, 1, \ldots, 1)$, where the -1 occurs at position w. Thus,

$$I_f = I - 2|w\rangle\langle w|.$$

Each use of an I_f gate will count as a probe. We will also use the gate

$$I_0 = I - 2|0^n\rangle\langle 0^n|,$$

which flips the sign of $|0^n\rangle$ but leaves all other basis states alone. I_0 can be implemented by an $O(n)$-size $O(\lg n)$-depth circuit using H, X, and CNOT gates. Finally we assume that we have some n-qubit unitary U available such that $\langle w|U|0^n\rangle \neq 0$. Setting $x := \langle w|U|0^n\rangle$ and by adjusting U by a phase factor if necessary, we can assume that $x > 0$. The larger x is, the better. If we let $U = H^{\otimes n}$ be a layer of n Hadamard gates, then we can get

$$x = \langle w|U|0^n\rangle = 2^{-n/2} \sum_{x \in \{0,1\}^n} \langle w|x\rangle = 2^{-n/2} = \frac{1}{\sqrt{N}}.$$

It turns out that we can't do better than this in the worst case. Grover's algorithm now works as follows:

1. Initialize an n-qubit register in the state $|0^n\rangle$.
2. Apply U to get the state $|s\rangle = U|0^n\rangle$. We call $|s\rangle$ the *start state*. Note that $x = \langle w|s\rangle = \langle s|w\rangle > 0$. We'll assume that $x < 1$, or equivalently, that $|s\rangle$ and $|w\rangle$ are linearly independent; otherwise, $|s\rangle \propto |w\rangle$ and we can skip the next step entirely. For U implemented with Hadamards as above, this assumption clearly holds.

3. Apply G to $|s\rangle$ $\lfloor \pi/(4\sin^{-1} x)\rfloor$ many times, where

$$G := -UI_0U^*I_f$$

is known as the *Grover iterate*.

4. Measure the n qubits in the computational basis, obtaining a value $y \in \{0,1\}^n$.

We'll show that $y = w$ with high probability. Note that if $x = 1/\sqrt{N}$, then $\lfloor \pi/(4\sin^{-1} x)\rfloor \doteq \pi/(4x) = \Theta(\sqrt{N})$, and so there are $\Theta(\sqrt{N})$ many probes, since G consists of one probe.

We expand G:

$$\begin{aligned} G &= -UI_0U^*I_f \\ &= -U(I - 2|0^n\rangle\langle 0^n|)U^*(I - 2|w\rangle\langle w|) \\ &= (I - 2U|0^n\rangle\langle 0^n|U^*)(I - 2|w\rangle\langle w|) \\ &= (I - 2|s\rangle\langle s|)(I - 2|w\rangle\langle w|) \\ &= -I + 2|s\rangle\langle s| + 2|w\rangle\langle w| - 4x|s\rangle\langle w|. \end{aligned}$$

Applying the right-hand side to $|s\rangle$ and $|w\rangle$ immediately gives us

$$\begin{aligned} G|s\rangle &= (1 - 4x^2)|s\rangle + 2x|w\rangle, \\ G|w\rangle &= -2x|s\rangle + |w\rangle. \end{aligned}$$

So we see that $G|s\rangle$ and $G|w\rangle$ are both (real) linear combinations of $|s\rangle$ and $|w\rangle$. Thus G maps the plane spanned by $|s\rangle$ and $|w\rangle$ into itself, and all intermediate states of the algorithm lie in this plane. Thus we can now restrict our attention to this two-dimensional subspace S.

Using Gram-Schmidt, we pick an orthonormal basis for S, with $|w\rangle$ being one vector and $|r\rangle := |r'\rangle/\||r'\rangle\|$ being the other, where $|r'\rangle := |s\rangle - x|w\rangle$. We have

$$\||r'\rangle\|^2 = \langle r'|r'\rangle = (\langle s| - x\langle w|)(|s\rangle - x|w\rangle) = 1 - x^2 - x^2 + x^2 = 1 - x^2,$$

and so

$$|r\rangle = \frac{|s\rangle - x|w\rangle}{\sqrt{1 - x^2}}.$$

It is easily checked that $\langle r|w\rangle = 0$. Let $0 < \theta < \pi/2$ be such that $x = \sin\theta$. Expressing $|s\rangle$ in the $\{|r\rangle, |w\rangle\}$ basis, we get

$$|s\rangle = \sqrt{1 - x^2}|r\rangle + x|w\rangle = \cos\theta|r\rangle + \sin\theta|w\rangle = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}.$$

Let's express G with respect to the same $\{|r\rangle, |w\rangle\}$ basis. Note that restricted to the subspace S, the identity I has the same effect as the orthogonal projector $P_S = |r\rangle\langle r| + |w\rangle\langle w|$ projecting onto S: they both fix all vectors in S. It follows that, restricted to S,

$$\begin{aligned} G &= -P_S + 2|s\rangle\langle s| + 2|w\rangle\langle w| - 4x|s\rangle\langle w| \\ &= -|r\rangle\langle r| - |w\rangle\langle w| + 2(\cos\theta|r\rangle + \sin\theta|w\rangle)(\cos\theta\langle r| + \sin\theta\langle w|) \\ &\qquad + 2|w\rangle\langle w| - 4\sin\theta(\cos\theta|r\rangle + \sin\theta|w\rangle)\langle w| \\ &= (2\cos^2\theta - 1)|r\rangle\langle r| - 2\cos\theta\sin\theta|r\rangle\langle w| + 2\sin\theta\cos\theta|w\rangle\langle r| + (1 - 2\sin^2\theta)|w\rangle\langle w| \\ &= \begin{bmatrix} \cos(2\theta) & -\sin(2\theta) \\ \sin(2\theta) & \cos(2\theta) \end{bmatrix}. \end{aligned}$$

Geometrically, if we identify $|r\rangle$ with the point $(1,0) \in \mathbb{R}^2$ and $|w\rangle$ with the point $(0,1) \in \mathbb{R}^2$, then $|s\rangle$ is the point in the first quadrant of the unit circle, forming angle θ with $|r\rangle$. Also, G is seen to give a counterclockwise rotation of the circle through angle 2θ. We want the state to wind up as close to $|w\rangle$ as possible, which makes an angle $\pi/2$ with $|r\rangle$. Applying G m times puts the state at an angle $(2m+1)\theta$ from $|r\rangle$, so we solve

$$(2m+1)\theta = \frac{\pi}{2} \iff m = \frac{\pi}{4\theta} - \frac{1}{2} = \frac{\pi}{4\sin^{-1} x} - \frac{1}{2}.$$

Rounding to the nearest integer gives $m = \lfloor \pi/(4\sin^{-1} x) \rfloor$, which is the number of times we apply G to $|s\rangle$. The final state is within an angle θ of $|w\rangle$, so the probability of getting w as the result of the measurement is at least $\cos^2\theta = 1 - x^2 = 1 - 2^{-n} = 1 - 1/N$ (if $x = 2^{-n/2}$), which is exponentially close to 1.

Interestingly, if we apply G too many times, then we start drifting away from $|w\rangle$ and the probability of getting w in the measurement will start going down again to about zero at $2m$ applications, then it will oscillate back to one at about $3m$, then close to zero again at $4m$, et cetera.

Some Variants of Quantum Search. An obvious variant is to assume that $f(z) = 1$ for *at most* one z, rather than for exactly one z. For this variant, one can run Grover's algorithm just as before, but check that the final result y is such that $f(y) = 1$, using one more probe of f. If not, then you can conclude that f is the constant zero function, and you'd be wrong with exponentially small probability.

Another variant is when there are exactly k many z such that $f(z) = 1$, where k is known, and your job is to find the location of any one of them. This is the subject of the next exercise.

Exercise 20.1 (Somewhat challenging) Show that if there are exactly k many z such that $f(z) = 1$, where $0 < k < 2^n$ is known, then one of the targets can be found with high probability using $O(\sqrt{N/k})$ probes to f. [Hint: Let $U = H^{\otimes n}$, let $|s\rangle = U|0^n\rangle = 2^{-n/2}\sum_{z\in\{0,1\}^n} |z\rangle$ be the start state, and let $G = -UI_0U^*I_f = -(I - 2|s\rangle\langle s|)I_f$ be the Grover iterate, all as before. Run Grover's algorithm as before, applying G some number of times to $|s\rangle$. To see how many times to apply G:

1. Define the state $|w\rangle$ to be an equal superposition of all target locations:

$$|w\rangle := \frac{1}{\sqrt{k}} \sum_{z:f(z)=1} |z\rangle.$$

2. Likewise, define the state $|r\rangle$ to be the superposition of all nontarget locations:

$$|r\rangle := \frac{1}{\sqrt{2^n - k}} \sum_{z:f(z)=0} |z\rangle.$$

   Notice that $|r\rangle$ and $|w\rangle$ are orthogonal unit vectors.

3. Let $x := \langle s|w\rangle$ as before. Show that now, $x = \sqrt{k}/2^{n/2}$.

4. Define $0 < \theta < \pi/2$ such that $x = \sin\theta$, just as before, and show that $|s\rangle = \cos\theta|r\rangle + \sin\theta|w\rangle$, just as before.

5. (The crucial step) Show directly that

$$\begin{aligned} G|r\rangle &= \cos(2\theta)|r\rangle + \sin(2\theta)|w\rangle, \\ G|w\rangle &= -\sin(2\theta)|r\rangle + \cos(2\theta)|w\rangle, \end{aligned}$$

just as before. Note that $G = -(I - 2|s\rangle\langle s|)I_f \neq -(I - 2|s\rangle\langle s|)(I - 2|w\rangle\langle w|)$, so the calculation must be a bit different from before. You might observe that I_f has the same effect as $I - 2|w\rangle\langle w|$ within the space spanned by $|r\rangle$ and $|w\rangle$, but you can't use this fact until you establish that G maps this space into itself. Better to just do the calculations above directly.

6. Conclude that G maps the space spanned by the orthonormal set $\{|r\rangle, |s\rangle\}$ into itself, and its matrix looks the same as before.

7. Conclude that $\lfloor \pi/(4\theta) \rfloor$ is the right number of applications of G, since measuring the qubits in a state close to $|w\rangle$ returns some target location with high probability. Show that $\lfloor \pi/(4\theta) \rfloor = \Theta(\sqrt{N/k})$.]

21 Week 10: Quantum search lower bound

A Lower Bound on Quantum Search. The number of probes to the function f in Grover's search algorithm is asymptotically tight. That is, no quantum algorithm can find a unique target in a search space of size N with high probability using $o(\sqrt{N})$ probes. This bound is due to Bennett, Bernstein, Brassard, and Vazirani, and predates Grover's algorithm. It is one of the earliest results in the area of *quantum query complexity*.

Suppose we are given an arbitrary r-qubit quantum circuit C of unitary gates followed by an n-qubit measurement in the computational basis. We assume that the initial state of the r qubits is some fixed $|0\rangle$, and that C may contain some number of n-qubit I_f gates, which allow it to make queries to a Boolean function $f : \{0,1\}^n \to \{0,1\}$. To prove a lower bound, our goal is to find some f corresponding to a unique target $w \in \{0,1\}^n$ (*i.e.*, $f(w) = 1$ and $f(z) = 0$ for all $z \neq w$) such that w is unlikely to be the final measurement result. The particular w that we choose will depend on the circuit C.

Here's the basic intuition. Suppose C contains some number of I_f gates. Just before one of these gates is applied, the state of its input qubits is generally some superposition of states $|z\rangle$ with $z \in \{0,1\}^n$. There are 2^n many such z, and since the state is a unit vector, most of the corresponding probability amplitudes must be close to zero. If the probability amplitude of some $|w\rangle$ is small, then changing $f(w)$ from 0 to 1 just flips the sign of this term in the superposition, which in turn makes little difference to the overall state and is likely to go unnoticed. We want to choose w so that this is true for all the I_f gates in C, as well as the final state of the measured qubits.

Now the details. This development is loosely adapted from pages 269–271 of the textbook, except that, unlike the textbook, we do not implicitly assume that our circuit C has only n qubits. Suppose that the circuit C has m many I_f gates, for some $m \geqslant 0$. For any f, the circuit C corresponds to the unitary transformation $U_m I_f^{(m)} U_{m-1} I_f^{(m-1)} \cdots U_1 I_f^{(1)} U_0$, where

- each $I_f^{(j)}$ is the unitary operator corresponding to the jth I_f gate, acting on some sequence of n of the r qubits,
- U_0 represents all the unitary gates applied prior to $I_f^{(1)}$,
- U_m represents all the unitary gates applied after $I_f^{(m)}$, and
- for all $0 < j < m$, U_j represents all the unitary gates applied strictly in between $I_f^{(j)}$ and $I_f^{(j+1)}$.

None of the unitary operators $U_0, \ldots, U_m$ depend on f.

For any $w \in \{0,1\}^n$ and $1 \leqslant j \leqslant m$, we let $I_w^{(j)}$ be $I_f^{(j)}$ where f is such that $f(w) = 1$ and $f(z) = 0$ for all $z \in \{0,1\}^n - \{w\}$. That is, $I_w^{(j)} = (I - 2|w\rangle\langle w|) \otimes I$, where the first operator applies to the qubits involved in the jth I_f gate, and the identity applies to the other qubits.

First we run C with each I_f gate replaced with the identity I (or if you like, I_z where z is the constant 0 function). That is, we run C with no targets. For all $0 \leqslant j \leqslant m$, let $\left|\psi^{(j)}\right\rangle$ be the state of the r qubits immediately after the application of U_j. That is,

$$\left|\psi^{(j)}\right\rangle = U_j I_z^{(j)} U_{j-1} \cdots U_1 I_z^{(1)} U_0 |0\rangle = U_j U_{j-1} \cdots U_1 U_0 |0\rangle.$$

In particular $\left|\psi^{(m)}\right\rangle$ is the final state. For $0 \leqslant j < m$ we can factor $\left|\psi^{(j)}\right\rangle$ uniquely as

$$\left|\psi^{(j)}\right\rangle = \sum_{x \in \{0,1\}^n} |x\rangle \left|\beta_x^{(j)}\right\rangle,$$

where the first ket in each term represents a basis state of the n qubits entering the $(j+1)$st I_f gate, and the second ket is a (not necessarily unit) vector representing the other $r - n$ qubits. Likewise, we uniquely factor $\left|\psi^{(m)}\right\rangle$ as

$$\left|\psi^{(m)}\right\rangle = \sum_{x \in \{0,1\}^n} |x\rangle \left|\beta_x^{(m)}\right\rangle,$$

where here the first ket in each term represents a basis state of the n qubits that are about to be measured, and the second ket is a vector representing the $r - n$ unmeasured qubits.

Since $\left|\psi^{(j)}\right\rangle$ is a state, we have, for all $0 \leqslant j \leqslant m$,

$$1 = \left\langle \psi^{(j)} \middle| \psi^{(j)} \right\rangle = \sum_{x \in \{0,1\}^n} \left\langle \beta_x^{(j)} \middle| \beta_x^{(j)} \right\rangle = \sum_x \left\| \left| \beta_x^{(j)} \right\rangle \right\|^2. \tag{82}$$

Let $w \in \{0,1\}^n$ be arbitrary. Now we run C again with I_w gates. For $0 \leqslant j \leqslant m$, define

$$\left|\varphi_w^{(j)}\right\rangle := U_j I_w^{(j)} U_{j-1} \cdots U_1 I_w^{(1)} U_0 |0\rangle$$

to be the state of the circuit just after the application of U_j. We claim that there are many values of w for which $|\varphi_w^{(j)}\rangle$ does not differ too much from $\left|\psi^{(j)}\right\rangle$, for any $1 \leqslant j \leqslant m$. For each $0 \leqslant j \leqslant m$, define the error vector

$$\left|\eta_w^{(j)}\right\rangle := \left|\varphi_w^{(j)}\right\rangle - \left|\psi^{(j)}\right\rangle$$

We want to show that enough of the vectors $|\eta_w^{(j)}\rangle$ have small norm. For each j, define

$$D^{(j)} := \sum_{w\in\{0,1\}^n} \left\| \left|\eta_w^{(j)}\right\rangle \right\|^2.$$

Claim 21.1 *$D^{(j)} \leqslant 4j^2$ for all $0 \leqslant j \leqslant m$.*

Proof. We proceed by induction on j. For $j = 0$, we have $|\varphi_w^{(0)}\rangle = U_0|0\rangle = |\psi^{(0)}\rangle$ and thus $|\eta_w^{(0)}\rangle = 0$ for all w, and so the claim clearly holds. Now for the inductive case where $0 \leqslant j < m$, we want to express $|\eta_w^{(j+1)}\rangle$ in terms of $|\eta_w^{(j)}\rangle$. We have, for all w,

$$\begin{aligned}
\left|\varphi_w^{(j+1)}\right\rangle &= U_{j+1} I_w^{(j+1)} \left|\varphi_w^{(j)}\right\rangle \\
&= U_{j+1} I_w^{(j+1)} \left(\left|\psi^{(j)}\right\rangle + \left|\eta_w^{(j)}\right\rangle\right) \\
&= U_{j+1} I_w^{(j+1)} \left(\sum_{x\in\{0,1\}^n} |x\rangle \left|\beta_x^{(j)}\right\rangle\right) + U_{j+1} I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle \\
&= U_{j+1} \left(\sum_x (I_w|x\rangle) \otimes \left|\beta_x^{(j)}\right\rangle\right) + U_{j+1} I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle \\
&= U_{j+1} \left(\sum_x (|x\rangle - 2|w\rangle\langle w|x\rangle) \otimes \left|\beta_x^{(j)}\right\rangle\right) + U_{j+1} I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle \\
&= U_{j+1} \left|\psi^{(j)}\right\rangle - 2U_{j+1}|w\rangle \left|\beta_w^{(j)}\right\rangle + U_{j+1} I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle \\
&= \left|\psi^{(j+1)}\right\rangle - 2U_{j+1}|w\rangle \left|\beta_w^{(j)}\right\rangle + U_{j+1} I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle.
\end{aligned}$$

Subtracting, we get

$$\left|\eta_w^{(j+1)}\right\rangle = \left|\varphi_w^{(j+1)}\right\rangle - \left|\psi^{(j+1)}\right\rangle = U_{j+1}\left(I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle - 2|w\rangle \left|\beta_w^{(j)}\right\rangle\right),$$

whence

$$\left\| \left|\eta_w^{(j+1)}\right\rangle \right\|^2 = \left\| I_w^{(j+1)} \left|\eta_w^{(j)}\right\rangle - 2|w\rangle \left|\beta_w^{(j)}\right\rangle \right\|^2 \leqslant \left(\left\| \left|\eta_w^{(j)}\right\rangle \right\| + 2 \left\| \left|\beta_w^{(j)}\right\rangle \right\| \right)^2.$$

Expanding and summing over $w \in \{0,1\}^n$, we have

$$\begin{aligned}
D^{(j+1)} &\leqslant D^{(j)} + 4 \sum_{w\in\{0,1\}^n} \left\| \left|\eta_w^{(j)}\right\rangle \right\| \cdot \left\| \left|\beta_w^{(j)}\right\rangle \right\| + 4 \sum_{w\in\{0,1\}^n} \left\| \left|\beta_w^{(j)}\right\rangle \right\|^2 \\
&= D^{(j)} + 4\langle \kappa, \lambda \rangle + 4,
\end{aligned}$$

where we have used Equation (82) for the last term, and where κ and λ are 2^n-dimensional column vectors whose entries, indexed by w, are $\||\eta_w^{(j)}\rangle\|$ and $\||\beta_w^{(j)}\rangle\|$, respectively. We can apply Cauchy-Schwarz to $\langle \kappa, \lambda \rangle$:

$$\langle \kappa, \lambda \rangle = |\langle \kappa, \lambda \rangle| \leqslant \|\kappa\| \cdot \|\lambda\| = \left(\sum_w \left\| \left|\eta_w^{(j)}\right\rangle \right\|^2 \right)^{1/2} \left(\sum_w \left\| \left|\beta_w^{(j)}\right\rangle \right\|^2 \right)^{1/2} = \sqrt{D^{(j)}} \cdot 1 = \sqrt{D^{(j)}},$$

using (82) again. Plugging this in above and using the inductive hypothesis, we have

$$D^{(j+1)} \leqslant D^{(j)} + 4\sqrt{D^{(j)}} + 4 \leqslant 4j^2 + 8j + 4 = 4(j+1)^2,$$

which proves the claim. □

Now for $j = m$, the claim asserts that $\sum_{w\in\{0,1\}^n} \left\||\eta_w^{(m)}\rangle\right\|^2 \leqslant 4m^2$. This implies that $\left\||\eta_w^{(m)}\rangle\right\|^2 > 4m^2/2^{n-1}$ for less than 2^{n-1} many w, *i.e.*, for more than half of the w, we have $\left\||\eta_w^{(m)}\rangle\right\|^2 \leqslant 4m^2/2^{n-1}$. Using a similar argument with Equation (82), we must have $\left\||\beta_w^{(m)}\rangle\right\|^2 \leqslant 1/2^{n-1}$ for more than half of the w. Thus there is some $w \in \{0,1\}^n$ such that both of these inequalities hold. Fix such a w. The final state of the circuit when run with target w is $|\varphi_w^{(m)}\rangle$, and we can factor it as

$$\left|\varphi_w^{(m)}\right\rangle = \sum_{x\in\{0,1\}^n} |x\rangle\left|\gamma_x^{(m)}\right\rangle,$$

where (as with $|\psi^{(m)}\rangle$) the first ket represents the n qubits that are about to be measured, and the second ket represents the other qubits (and is not necessarily a unit vector). The probability of seeing w as the outcome of the measurement when the state is $|\varphi_w^{(m)}\rangle$ is then $\Pr[w] = \left\||\gamma_w^{(m)}\rangle\right\|^2$, but this value is quite small, provided m is not too large:

$$\begin{aligned}
\left\|\left|\gamma_w^{(m)}\right\rangle\right\| &= \left\|\left|\gamma_w^{(m)}\right\rangle - \left|\beta_w^{(m)}\right\rangle + \left|\beta_w^{(m)}\right\rangle\right\| \\
&\leqslant \left\|\left|\gamma_w^{(m)}\right\rangle - \left|\beta_w^{(m)}\right\rangle\right\| + \frac{\sqrt{2}}{2^{n/2}} \\
&= \left\||w\rangle \otimes \left(\left|\gamma_w^{(m)}\right\rangle - \left|\beta_w^{(m)}\right\rangle\right)\right\| + \frac{\sqrt{2}}{2^{n/2}} \\
&= \left\|(|w\rangle\langle w| \otimes I)\left(\left|\varphi_w^{(m)}\right\rangle - \left|\psi^{(m)}\right\rangle\right)\right\| + \frac{\sqrt{2}}{2^{n/2}} \\
&= \left\|(|w\rangle\langle w| \otimes I)\left|\eta_w^{(m)}\right\rangle\right\| + \frac{\sqrt{2}}{2^{n/2}} \\
&\leqslant \left\|\left|\eta_w^{(m)}\right\rangle\right\| + \frac{\sqrt{2}}{2^{n/2}} \\
&\leqslant \frac{(2m+1)\sqrt{2}}{2^{n/2}}.
\end{aligned}$$

And so we get that $\Pr[w] \leqslant (2m+1)^2/2^{n-1} = O(m^2/2^n)$. So finally, if $m = o(2^{n/2})$, we have $\Pr[w] = o(1)$, *i.e.*, $\Pr[w]$ approaches zero as n gets large, and the circuit likely won't find w.

22 Week 11: Quantum cryptography

Quantum Cryptographic Key Exchange. If Alice and Bob share knowledge of a secret string r of random bits, then Alice can send a message m with the same number of bits as r to Bob over a channel subject to eavesdropping with *perfect secrecy*, *i.e.*, no third party (Eve), monitoring the channel with no knowledge of r, can gain any knowledge about m whatsoever. This scheme, known as a *one-time pad*, works as follows:

1. Alice computes $c = m \oplus r$, the bitwise exclusive OR of m and r. The message m is called the *cleartext* or *plaintext*, and c is called the *ciphertext*.
2. Alice transmits the ciphertext c to Bob over the channel, which we'll assume is publically readable, e.g., a newspaper or an internet bulletin board.
3. Bob gets c and computes $m = c \oplus r$, thus recovering the cleartext m.

All Eve sees is $c = m \oplus r$, and since she doesn't know r which is assumed to be uniformly random, the bits of c look completely random to her—all possible c's are equally likely if all possible r's are equally likely. Hence the perfect secrecy.

It's called a one-time pad for a reason: r cannot be reused to send another message. Suppose Alice sends another message m' using the same r to transmit $c' = m' \oplus r$. Then Eve can compute

$$c \oplus c' = (m \oplus r) \oplus (m' \oplus r) = m \oplus m' \oplus r \oplus r = m \oplus m'.$$

If m and m' are both uncompressed files of English text, then they have enough redundancy that Eve can gain some knowledge of m and m' from their XOR, and likely can even decipher both m and m' uniquely from $m \oplus m'$ if the messages are long enough.

If r is short, say, only a few hundred bits long, then Alice can only transmit that amount of bits in her message with a one-time pad. It is more practical instead for Alice and Bob to use r as the key to some symmetric cipher by which they can communicate longer messages. Some commonly used ciphers for electronic communications include the Advanced Encryption Standard (AES, a.k.a. Rijndael), Blowfish, and 3DES. These ciphers are called *symmetric* because the same key r is used by Alice to encrypt and by Bob to decrypt. These ciphers do not provide perfect secrecy in the theoretical sense, but they are widely believed to be infeasible to crack.

We get back to the question of how Alice and Bob manage to share r securely in the first place. If they spend any time together in a physically secure room, they can flip coins and generate an r. In practice, though, it is not possible for Alice and Bob to ever be together; they may not even know each other (for example, Alice buys a book online from Bob, who is Barnes and Noble). This is the problem of *key exchange*, and it is currently handled using some kind of public key cryptography such as RSA, Diffie-Hellman, or El-Gamal. I won't go into how public key crypto works here, except to say that it relies for its security on the difficulty of performing certain number-theoretic tasks, such as factoring (RSA) and computing discrete logarithms (Diffie-Hellman, El-Gamal). If quantum computers are ever physically realized, then Shor's algorithms for factoring and discrete log could break current public key schemes.

A key-exchange protocol using quantum mechanics was proposed in 1984 by Charles Bennett and Gilles Brassard. In this protocol, known as *BB84*, Alice sends a sequence of qubits to Bob

across an insecure quantum channel, subject to eavesdropping/tampering by Eve. Alice and Bob then perform a series of checks, communicate through a public, nonquantum channel, and in the end they share some secret random bits. The security of the protocal relies only on the laws of physics and the faithfulness of the implementation, and not on the assumed difficulty of certain tasks like factoring large numbers. The key intuition is that in quantum mechanics, measuring a quantum system may unavoidably alter the system being measured. If Eve wants to get information about the qubits being sent from Alice to Bob, she must perform a measurement, which will disrupt the qubits enough to be detected by Alice and Bob with high probability. For brevity, I will only describe the basic, simplistic, idealized, and unoptimized protocol here. There are a number of technical issues (such as noise) that I won't go into. A quick tutorial on quantum cryptography by Jamie Ford at Dartmouth College can be found at `https://www.cs.dartmouth.edu/~jford/crypto.html` (this link now appears to be broken). There is an on-line simulation of BB84 by Frederick Henle at `http://fredhenle.net/bb84/demo.php`. An extensive (though now somewhat outdated) bibliography of quantum cryptography papers, started(?) by Gilles Brassard (Université de Montréal) and maintained(?) by Claude Crépeau at McGill University, is at `https://www.cs.mcgill.ca/~crepeau/CRYPTO/Biblio-QC.html`.

In the BB84 protocol, it is assumed that Alice and Bob share an insecure quantum channel, which Alice will use to send qubits to Bob, and a classical information channel (such as a newspaper, phone, or electronic bulletin board) that is public (anyone can monitor it) but *reliable*, in the sense that any message that Alice and Bob send to each other along this channel reaches the recipient without alteration, and it is impossible for a third party to send a message to Alice or Bob pretending to be the other (*i.e.*, it is forgery proof). The description of BB84 needs the following:

Definition 22.1 Let $\mathcal{H}$ be an n-dimensional Hilbert space, and let $\mathcal{B} = \{b_1, \ldots, b_n\}$ and $\mathcal{C} = \{c_1, \ldots, c_n\}$ be two orthonormal bases for $\mathcal{H}$. We say that $\mathcal{B}$ and $\mathcal{C}$ are *mutually unbiased*, or *complementary*, if $|\langle b_1|c_j\rangle| = 1/\sqrt{n}$ for all $1 \leqslant i, j \leqslant n$. A collection $\mathcal{B}_1, \ldots, \mathcal{B}_k$ of orthonormal bases for $\mathcal{H}$ is *mutually unbiased* if each pair of bases in the collection is mutually unbiased.

The geometrical intuition is that $\mathcal{B}$ and $\mathcal{C}$ are mutually unbiased iff the "angle" between any member of $\mathcal{B}$ and any member of $\mathcal{C}$ is always the same, up to a phase factor.

Exercise 22.2 Show that if $\mathcal{B}$ and $\mathcal{C}$ are two orthonormal bases of an n-dimensional Hilbert space such that $|\langle b, c\rangle| = |\langle b', c'\rangle|$ for any $b, b' \in \mathcal{B}$ and $c, c' \in \mathcal{C}$, then $1/\sqrt{n}$ is the common value of $|\langle b, c\rangle|$ for any $b \in \mathcal{B}$ and $c \in \mathcal{C}$. [Hint: Express a vector from $\mathcal{C}$ as a linear combination of vectors from $\mathcal{B}$. What can you say about the coefficients?]

A d-dimensional Hilbert space cannot have a mutually unbiased collection of more than $d + 1$ orthonormal bases. If d is a power of a prime number, then $d + 1$ mutually unbiased bases can be constructed, but it is an open problem to determine how many mutually unbiased bases there can be when d is not a prime power. Even the case where $d = 6$ is open. Anyway, for the one-qubit case where $d = 2$, the bases $\{|+x\rangle, |-x\rangle\}$, $\{|+y\rangle, |-y\rangle\}$, and $\{|+z\rangle, |-z\rangle\}$ are mutually unbiased. (Other collections of three mutually unbiased bases can be obtained from these three by applying some unitary operator U to every vector (the same for all the vectors). Applying U does not change the inner product of any pair of vectors.) BB84 uses two of these three, say, $\{|+z\rangle, |-z\rangle\}$ and $\{|+x\rangle, |-x\rangle\}$. We'll denote the first of these by $\updownarrow$, consisting of spin-up ($\uparrow$) and spin-down ($\downarrow$) states, and the

second by $\leftrightarrow$, consisting of spin-right ($\rightarrow$) and spin-left ($\leftarrow$) states. The two vectors of each basis encode the two possible bit values: in the $\updownarrow$ basis, $\uparrow$ encodes 0 and $\downarrow$ encodes 1; in the $\leftrightarrow$ basis, $\rightarrow$ encodes 0 and $\leftarrow$ encodes 1. Here is the protocol:

Sending qubits. Alice and Bob repeat the following for j running from 1 to n, where n is some large number. The random choices made at one iteration are independent of those made at other iterations.

1. Alice chooses a bit $b_j \in \{0,1\}$ uniformly at random. She also chooses $\mathcal{B}_j$ to be one of the bases $\updownarrow$ or $\leftrightarrow$ uniformly at random, independent of b_j. She prepares a qubit in a state $|q_j\rangle$ encoding the bit b_j in the basis $\mathcal{B}_j$ (*i.e.*, $|q_j\rangle$ is either $\uparrow$ or $\rightarrow$ for $b_j = 0$, and either $\downarrow$ or $\leftarrow$ for $b_j = 1$), and sends the qubit $|q_j\rangle$ to Bob across the quantum channel.
2. Bob receives the qubit sent from Alice, chooses a basis $\mathcal{C}_j$ from $\{\updownarrow, \leftrightarrow\}$ uniformly at random, and measures the qubit projectively using $\mathcal{C}_j$, obtaining a bit value c_j according to the same encoding scheme described above.

This ends the quantum part of the protocol. All further communication between Alice and Bob is classical and uses the public, classical channel.

Discarding uncorrelated bits. Note that if the quantum channel faithfully transmits all of Alice's qubits to Bob unaltered, then $b_j = c_j$ with certainty whenever Alice's basis was the same as Bob's, *i.e.*, whenever $\mathcal{B}_j = \mathcal{C}_j$; otherwise b_j and c_j are completely uncorrelated (because $\updownarrow$ and $\leftrightarrow$ are mutually unbiased).

1. For each $1 \leqslant j \leqslant n$, Bob tells Alice the basis $\mathcal{C}_j$ he used to measure c_j.
2. Alice replies to Bob with the set $C = \{j \in \{1, \ldots, n\} : \mathcal{B}_j = \mathcal{C}_j\}$ (C stands for "correlated"). Let k be the size of C. Note that k is expected to be about $n/2$, because each $\mathcal{B}_j$ and $\mathcal{C}_j$ were chosen independently.
3. Alice and Bob discard the results of all trials where $\mathcal{B}_j \neq \mathcal{C}_j$. Alice retains the bits b_j and Bob retains the bits c_j, for all $j \in C$. If the quantum channel was not tampered with, then $b_j = c_j$ for all $j \in C$.

Security check.
1. Alice chooses a subset $S \subseteq C$ uniformly at random (S stands for "security check"). For example, she decides to put j into S with probability $1/2$ independently for each $j \in C$. The set S is expected to have size about $k/2$.
2. Alice sends S to Bob along with the value of b_j for each $j \in S$.
3. Bob checks whether $b_j = c_j$ for every $j \in S$. If so, he tells Alice that they can accept the protocol, in which case, Alice and Bob respectively discard the bits b_j and c_j where $j \in S$ and retain the rest of the bits b_j and c_j for $j \in C - S$ (about $k/2$ or about $n/4$ many bits). On the other hand, if there are *any* discrepancies, then Bob tells Alice that they should reject the protocol, in which case, all bits are discarded and they start over with an entirely new run of the protocol.

Note that if the quantum channel is not tampered with, then Alice and Bob will accept the protocol. Also notice that any third party monitoring the classical communication between Alice and Bob knows nothing of the bits that Alice and Bob eventually retain. We'll explain why there is a good

chance that Eve will be caught and the protocol rejected if she tries to eavesdrop on the quantum channel during the initial qubit communication.

For technical simplicity, we will assume that there is only one way that Eve can eavesdrop on the quantum channel: she can choose to measure some qubit in either of the bases $\updownarrow$ or $\leftrightarrow$, then send along to Bob some qubit that she prepares based on her measurement. This is not a general proof of security then, because Eve could do other things: measure a qubit in some arbitrary basis, or even couple the qubit to another quantum system, let the combined system evolve, make a measurement in the combined system, then send along some qubit to Bob based on that. She could even make correlated measurements involving several of the sent qubits together. It takes some work to show that Eve's chances are being caught are not significantly reduced by these more general attacks, and we won't show the more general proof here.

If Eve happens to measure a qubit $\left|q_j\right\rangle$ in the same basis $\mathcal{B}_j$ that Alice used, then this is very good for Eve: She knows the encoded bit with certainty, and the post-measurement state is still $\left|q_j\right\rangle$, *i.e.*, Eve did not alter it. So she can simply retransmit the post-measurement qubit to Bob. In this case, if $j \in S$, then this qubit won't provide any evidence of tampering; if $j \in C - S$, then Eve knows one of the "secret" bits that Alice and Bob share, assuming they accept the protocol.

With probability $1/2$, however, Eve measures $\left|q_j\right\rangle$ in the wrong basis $\mathcal{B}'_j$—the one other than $\mathcal{B}_j$. In this case, she gets a bit value uncorrelated with b_j, but even worse (for Eve), her measurement alters the qubit so as to lose any information about b_j. She has to send a qubit to Bob, and at this point she cannot tell that she has chosen the wrong basis, so the best she can do is what she did before: resend the post-measurement qubit to Bob. If $j \in C$, then Bob will measure Eve's altered qubit $\left|r_j\right\rangle$ using $\mathcal{B}_j$, and since $\left|r_j\right\rangle$ is in the basis $\mathcal{B}'_j$, which is mutually unbiased with $\mathcal{B}_j$, Bob's result c_j will be completely random and uncorrelated with Alice's b_j. If $j \in S$ and $b_j \neq c_j$, then Eve is caught and the protocol is rejected.

To summarize, for each qubit $\left|q_j\right\rangle$ that Eve decides to eavesdrop on, Eve will get caught measuring the qubit if and only if

- she chooses the wrong basis (the one other than $\mathcal{B}_j$), and
- $j \in C$ (*i.e.*, Bob uses $\mathcal{B}_j$ to do his measurement and the bit is not discarded as uncorrelated), and
- Bob measures a value $c_j \neq b_j$, and
- $j \in S$ (*i.e.*, this is one of the bits Alice and Bob use for the security check).

Each of these four things happens with probability $1/2$, conditioned on the event that the things above it all happened. This makes the chances of Eve being caught on behalf of this qubit to be $(1/2)^4 = 1/16$. If Eve decides to eavesdrop on qubits $\left|q_{j_1}\right\rangle, \ldots, \left|q_{j_\ell}\right\rangle$ for $1 \leqslant j_1 < \cdots < j_\ell \leqslant n$, then each of these gives her a $1/16$ chance of being caught, independently of the others. The probability of her *not* being caught is then

$$\left(1 - \frac{1}{16}\right)^\ell < e^{-\ell/16},$$

which decreases exponentially in ℓ and is less than $1/e$ for $\ell \geqslant 16$. So Eve cannot eavesdrop on more than 16 qubits without a high probability of being caught. If $n \gg 16$, this is a negligible fraction of the roughly $n/4$ bits retained by Alice and Bob if they accept the protocol.

Exercise 22.3 Suppose that instead of the security check given above, Alice and Bob decide to do the following alternate security check:

1. Alice and Bob each compute the parities $b = \bigoplus_{j \in C} b_j$ and $c = \bigoplus_{j \in C} c_j$ of their current respective qubits.
2. They compare b and c over the public channel.
3. If $b \neq c$, then Alice and Bob reject the protocol and start over. Otherwise, they agree on some $j_0 \in C$ (it doesn't matter which), discard b_{j_0} and c_{j_0}, and retain the rest of the bits b_j and c_j for $j \in C - \{j_0\}$ as their shared secret, accepting the protocol. [If they didn't discard one of the bits, then someone monitoring the public channel would know the parity of Alice's and Bob's shared bits. Discarding a bit removes this information.]

How many bits on average do Alice and Bob retain in this altered protocol, assuming they accept it? What are Eve's chances of being caught if she eavesdrops on ℓ of the qubits, where $\ell > 0$?

In practice, polarized photons are used as qubits for the quantum communication phase. Alice may generate these photons by a light-emitting diode (LED) and can send them to Bob through fiber optic cable. Photon polarization has a two-dimensional state space and so can serve as a qubit. The three standard mutually unbiased bases for photon polarization (each given with its two possible states) are:

- rectilinear (horizontal, vertical),
- diagonal (northeast-southwest, northwest-southeast), and
- circular (clockwise, counterclockwise).

Photons have the advantage that their polarization is easy to measure and is insensitive to certain common sources of noise, e.g., stray electric and magnetic fields.

One technical problem is making sure that only one photon is sent at a time. Alice sends a pulse of electric current through the LED, which emits light in a burst of coherent photons with intensity (expected number of photons) proportional to the strength of the current. If more than one photon is sent at a time (in identical quantum states), then Eve could conceivably catch one of the photons and measure it, letting the other photon(s) go through to Bob as if nothing had been tampered with. To reduce the probability of a multiphoton burst, the current Alice sends through the LED must be exceedingly weak: about one tenth the energy of a single photon, say. Then the expected number of photons sent each time is about $1/10$. This means that about nine times out of ten, no photons are emitted at all. If $\lambda > 0$ is the ratio of the current energy divided by the energy of a single photon (in this example, $\lambda = 0.1$), then the number of photons emitted in any given burst satisfies a Poisson distribution with mean λ(?); that is, the probability that k photons are emitted is

$$f(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!},$$

where k is any nonnegative integer. If $\lambda = 0.1$, then $f(k, 0) = e^{-\lambda} \doteq 0.9$, which is the probability that the LED emits no photons. The probability of getting a single photon is $f(1, \lambda) = e^{-\lambda}\lambda \doteq 0.09 \doteq 0.1$.

The probability of two emitted photons is $f(2,\lambda) = e^{-\lambda}\lambda^2/2 \doteq 0.005$, or about one twentieth the probability of a single photon. More photons occur with rapidly diminishing probability. So if we ignore the times when no photons are emitted (Bob tells Alice that he did not receive a photon), the chances of multiple photons is small—about 1/20. The smaller λ is, the smaller this probability will be, but the trade-off is that we have to wait longer for a single photon.

Of course, the quantum channel could also be subject to random, nonmalicious noise, which would cause discrepancies between Alice's and Bob's bits. One subtlety is to make the protocol tolerate a certain amount of noise but still detect malicious tampering with high probability.

23 Week 11: Basic quantum information

We now start our discussion of quantum information. One of the major uses of quantum information theory is to analyze how noise can disrupt a quantum computation and how to make the computation resistant to it. The textbook discusses quantum information in earnest in Chapters 8–12, with quantum error correction in Chapter 10 and quantum information theory in Chapter 12. Quantum information is one of the textbook's real strong points, and I will assume you will read starting with Chapter 8. The lectures will fill in some background and reiterate points in the text. In the next few topics, we will be using the density operator formalism almost exclusively. We really have no choice about this once we generalize our notion of "state" to include mixed states.

Norms of Operators. Recall the definition of the Hilbert-Schmidt inner product on $\mathcal{L}(\mathcal{H})$ (Equation (11)). This suggests another way to define the norm of an operator:

$$\|A\|_2 := \langle A, A\rangle^{1/2} = (\mathrm{tr}(|A|^2))^{1/2}.$$

This norm, known as the *Euclidean norm*, the L_2*-norm*, or the *Hilbert-Schmidt norm*, satisfies all ten of the properties satisfied by the operator norm of Definition 18.4 except property 4; in fact, $\|I\|_2 = \sqrt{n}$, where n is the dimension of $\mathcal{H}$. The Euclidean norm is one of a parameterized family of norms defined on operators. For any real $p \geqslant 1$, define the L_p*-norm* (also called the *Schatten* p*-norm*) of an operator A to be

$$\|A\|_p := (\mathrm{tr}(|A|^p))^{1/p} = \left(\sum_{j=1}^{n} s_j^p\right)^{1/p}, \tag{83}$$

where $s_1, \ldots, s_n \geqslant 0$ are the eigenvalues of $|A|$, which are called the *singular values* of A. (If p is not an integer, then technically, we have not yet defined $|A|^p$, because $|A|$ is an operator. For now, you can ignore the middle expression in the equation above and use the right-hand side for the definition of $\|A\|_p$.) For $p = 2$, we get the Euclidean norm. The L_1 norm $\|A\|_1 = \mathrm{tr}\,|A|$ is also called the *trace norm* and is often useful. In addition, we could define the L_∞ norm

$$\|A\|_\infty = \lim_{p\to\infty} \|A\|_p = \max(s_1, \ldots, s_n),$$

but this is precisely the operator norm $\|A\|$ of Definition 18.4, because as p gets large, the largest term in the sum in (83) starts to dominate.

Exercise 23.1 Show that if A is an operator on an n-dimensional space, and $1 \leqslant p \leqslant q$ are real numbers, then $\|A\|_p \geqslant \|A\|_q$. Also show that $n\|A\| \geqslant \|A\|_1$. Thus all these norms are within a factor of n of each other. What is $\|I\|_p$? [Hint: For the first part, fix $s_1, \ldots, s_n \geqslant 0$ and differentiate the expression $\left(\sum_{j=1}^n s_j^p\right)^{1/p}$ with respect to p, and show that the derivative is always negative or zero.]

POVMs. Let S be a physical system with state space $\mathcal{H}_S$. Often, we want to get some classical information about the current state of S. We can perform a projective measurement on $\mathcal{H}_S$, obtaining various possible outcomes with various probabilites. This is not the only way to get information about the state of S, however. We could instead couple the system S with another system T in some known state in the state space $\mathcal{H}_T$, let the combined system ST evolve for a while, then make a projective measurement of the combined system, *i.e.*, on the space $\mathcal{H}_S \otimes \mathcal{H}_T$. This approach is more general and can get information that cannot be obtained by a projective measurement on $\mathcal{H}_S$ itself.

Recall that mathematically, a projective measurement on a Hilbert space $\mathcal{H}$ corresponds to a complete set of orthogonal projectors $\{P_j : j \in \mathcal{I}\}$, where $\mathcal{I}$ is the set of possible outcomes. We'll now relax this restriction a bit.

Definition 23.2 Let $\mathcal{H}$ be a Hilbert space. A *positive operator-valued measure*, or *POVM* on $\mathcal{H}$ is a set $\mathcal{M} = \{M_j : j \in \mathcal{I}\}$ where $\mathcal{I}$ is some finite or countably infinite set (the possible outcomes), each $M_j \geqslant 0$ is a positive operator in $\mathcal{L}(\mathcal{H})$, and

$$\sum_{j \in \mathcal{I}} M_j = I,$$

the identity operator on $\mathcal{H}$. If $\rho \in \mathcal{L}(\mathcal{H})$ is a state, then measuring ρ with respect to $\mathcal{M}$ yields outcome $j \in \mathcal{I}$ with probability $\Pr[j] := \langle M_j, \rho \rangle = \operatorname{tr}(M_j \rho)$.

We need to check that the $\Pr[j]$ really form a probability distribution. Fix a state $\rho \in \mathcal{L}(\mathcal{H})$. For each $j \in \mathcal{I}$, we have $\Pr[j] = \langle M_j, \rho \rangle \geqslant 0$ by Theorem 9.31, since M_j and ρ are both positive operators. Furthermore,

$$\sum_{j \in \mathcal{I}} \Pr[j] = \sum_j \langle M_j, \rho \rangle = \left\langle \sum_j M_j,\ \rho \right\rangle = \langle I, \rho \rangle = \operatorname{tr} \rho = 1 \,.$$

So the $\Pr[j]$ do form a probability distribution. It is important to note that the *only* properties of ρ that we used here are that $\rho \geqslant 0$ and that $\operatorname{tr} \rho = 1$. This is important, because we are about the expand our definition of "state" to include mixed states, which may no longer be projection operators, but are still positive and have unit trace.

Notice that for a POVM, we don't specify the post-measurement state. This is OK—quite often, we don't care what the post-measurement state is; we only care about the outcomes and their statistics, and a POVM provides the most general means of measuring a quantum system if we don't care about the state after the measurement. We'll show in a bit that a POVM is equivalent

to what we described above: coupling the system to another system, letting the combined system evolve, then making a projective measurement on the combined system.

A projective measurement on $\mathcal{H}$ is just a special case of a POVM where the M_j form a complete set of orthogonal projectors. To see this, we refer back to Equation (15): $\Pr[k] = \langle\psi|P_k|\psi\rangle$, where $\Pr[k]$ is the probability of outcome k when measuring the system in state $|\psi\rangle$, and P_k is the corresponding projector. Letting $\rho := |\psi\rangle\langle\psi|$ and treating the scalar $\langle\psi|P_k|\psi\rangle$ as a 1×1 matrix, we get

$$\Pr[k] = \langle\psi|P_k|\psi\rangle = \mathrm{tr}\,\langle\psi|P_k|\psi\rangle = \mathrm{tr}\,(P_k|\psi\rangle\langle\psi|) = \mathrm{tr}(P_k\rho) = \langle P_k, \rho\rangle\,,$$

which accords with Definition 23.2.

Exercise 23.3 Consider the following three-outcome POVM:

$$M_1 = \frac{1}{4}\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix},\ M_2 = \frac{1}{4}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix},\ M_3 = \frac{1}{4}\begin{bmatrix} 1 & 1-i \\ 1+i & 2 \end{bmatrix}.$$

Let $\rho = |\psi\rangle\langle\psi|$, where

$$|\psi\rangle = \frac{1}{5}\begin{bmatrix} 4 \\ -3i \end{bmatrix}.$$

What is the probability of each of the three outcomes when ρ is measured using the POVM above? Find a unit vector $|\varphi\rangle$ such that when $|\varphi\rangle\langle\varphi|$ is measured with the POVM, the second outcome occurs with probability 0. (Challenging part) Prove that the first outcome occurs with positive probability no matter what the state is. What is the essential property of M_1 that makes this true? Clearly, M_2 does not share this property. Does M_3? [Hint: When computing the probabilities above, you can save yourself some calculation by using the fact that $\langle M_j, \rho\rangle = \langle\psi|M_j|\psi\rangle$ for all j.]

Mixed States.

Definition 23.4 Let $A_1, \ldots, A_k$ be scalars, vectors, operators, matrices, etc., all of the same type. A *convex linear combination* of $A_1, \ldots, A_k$ is a value of the form

$$\sum_{i=1}^{k} p_i A_i,$$

where the p_i are real scalars, each $p_i \geqslant 0$, and $\sum_{i=1}^{k} p_i = 1$. In other words, $p_1, \ldots, p_k$ form a probability distribution.

Suppose Alice has a lab where she can prepare several states $\rho_1 = |\psi_1\rangle\langle\psi_1|, \ldots, \rho_k = |\psi_k\rangle\langle\psi_k| \in \mathcal{L}(\mathcal{H})$, and she flips coins and decides to prepare a state σ chosen at random from this set, where each ρ_i is chosen with probability p_i. She then sends the state σ she prepared to Bob, without telling him what it is. What can Bob find out about the state σ that Alice sent her? He can, most generally, perform a measurement corresponding to some POVM $\{M_j : j \in \mathcal{J}\}$. The probability of obtaining any outcome j, taken over both the POVM and Alice's random choice is then

$$\Pr[j] = \sum_{i=1}^{k} \Pr[j \mid \sigma = \rho_i] \cdot \Pr[\sigma = \rho_i] = \sum_i \langle M_j, \rho_i\rangle p_i = \langle M_j, \rho\rangle,$$

where $\rho = \sum_{i=1}^{k} p_i \rho_i$ is a convex combination of the ρ_i with the associated probabilities. So all Bob can ever determine physically about Alice's σ is given by the single operator ρ, which is called a *mixed state*. By definition, a mixed state is any nontrivial convex linear combination of one-dimensional projectors. By "nontrivial" we mean that all probabilities are strictly less than 1, or equivalently, there are at least two probabilities that are nonzero. Mathematically, a mixed state behaves in many ways much like a state of the form $|\psi\rangle\langle\psi|$ for some unit vector $|\psi\rangle$ (*i.e.*, a one-dimensional projector). It represents the state of a quantum system about which we have incomplete information, or which we are not describing completely. Completely described states, which up until now we have been dealing with exclusively, are of the form $|\psi\rangle\langle\psi|$ for unit vectors $|\psi\rangle$. From now on we will call these latter states *pure states*, and when we use the word "state" unqualified, we will mean either a pure or mixed state. Both kinds of states are convex combinations of pure states, trivial or otherwise. A mixed state is then some nontrivial probabilistic mixture, or weighted average, of pure states.

Why are mixed states important? When we consider a quantum system that is *not* isolated from its environment (which we must do when we consider quantum errors and decoherence), then some information about the state of the system "bleeds" out into the environment, leaving the system in a partially unknown state—even if the system started out in a pure state. We model an incompletely known quantum state as a mixed state.

Let's verify that if ρ is any state (say, $\rho = \sum_{i=1}^{k} p_i \rho_i$, where the p_i form a probability distribution and the ρ_i are all pure states), we have $\rho \geqslant 0$ and $\operatorname{tr} \rho = 1$. For positivity, let v be any vector. Then

$$v^* \rho v = \sum_{i=1}^{k} p_i \, v^* \rho_i v \geqslant 0,$$

because all the ρ_i are positive operators. Thus $\rho \geqslant 0$. By linearity of the trace, we have

$$\operatorname{tr} \rho = \sum_{i=1} p_i \operatorname{tr} \rho_i = \sum_i p_i = 1,$$

becase all the ρ_i have unit trace. The next proposition says that the converse of this is also true.

Proposition 23.5 *If $\rho \in \mathcal{L}(\mathcal{H})$ is such that $\rho \geqslant 0$ and $\operatorname{tr} \rho = 1$, then ρ is a convex linear combination of one-dimensional projectors that project onto mutually orthogonal subspaces.*

Proof. Suppose $\rho \geqslant 0$ and $\operatorname{tr} \rho = 1$. Since ρ is normal, it has an eigenbasis $\{|\psi_1\rangle, \ldots, |\psi_n\rangle\}$. With respect to this eigenbasis, ρ is represented by the matrix $\operatorname{diag}(p_1, \ldots, p_n)$ for some $p_1, \ldots, p_n$ and so $\rho = \sum_{i=1}^{n} p_i |\psi_i\rangle\langle\psi_i|$. Since $\rho \geqslant 0$, all the p_i are nonnegative real, and further $1 = \operatorname{tr} \rho = \sum_i p_i$. So ρ is a convex combination of $|\psi_1\rangle\langle\psi_1|, \ldots, |\psi_n\rangle\langle\psi_n|$, which project onto mutually orthogonal, one-dimensional subspaces. □

Thus we get the following two corollaries:

Corollary 23.6 *An operator $\rho \in \mathcal{L}(\mathcal{H})$ is a state (i.e., a convex combination of one-dimensional projectors) if and only if ρ is positive and has unit trace.*

Corollary 23.7 *An operator $\rho \in \mathcal{L}(\mathcal{H})$ is a state if and only if ρ is normal and its eigenvalues form a probability distribution.*

Measuring a mixed state with a POVM has exactly the same mathematical form as with a pure state. Recall that the only two properties of the state ρ we used to show that the measurement makes sense is that $\rho \geqslant 0$ and $\operatorname{tr}\rho = 1$, both of which are true of any mixed state. Similarly, unitary time evolution of a mixed state has exactly the same mathematical form as with a pure state. Indeed, if $\rho = \sum_i p_i \rho_i$ is some mixture of pure states, then evolving ρ via a unitary operator U should be equivalent to evolving each ρ_i by U and taking the same mixture of the results. By linearity, this gives

$$\rho = \sum_i p_i \rho_i \stackrel{U}{\longmapsto} \sum_i p_i (U\rho_i U^*) = U\left(\sum_i p_i \rho_i\right) U^* = U\rho U^*.$$

Finally, we won't bother proving it, but Equations (29) and (30), which describe projective measurements, are equally valid for mixed states ρ as well as for pure states.

Different probability distributions of pure states can yield the same state, but if they do, they are physically indistinguishable, that is, no physical experiment can tell one distribution from the other with positive probability. However, for any state ρ, there is a *preferred* mix of pure states that yields ρ, namely, the "eigenstates" $|\psi_1\rangle\langle\psi_1|, \ldots, |\psi_n\rangle\langle\psi_n|$ used in the proof of Proposition 23.5, with their respective eigenvalues as probabilities. The states are distinguished by the fact that they are pairwise orthogonal. We will call this preferred probability distribution the *eigenvalue distribution* of ρ.

It's time for an example. Alice may send Bob a single qubit in state $|0\rangle$ with probability 1/2 and state $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$ with probability 1/2. The resulting mixed state is

$$\rho = \frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|+\rangle\langle +| = \frac{|0\rangle\langle 0|}{2} + \frac{|0\rangle\langle 0| + |0\rangle\langle 1| + |1\rangle\langle 0| + |1\rangle\langle 1|}{4} = \frac{1}{4}\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}.$$

Let's find the eigenvalue distribution of ρ. One can easily check that an eigenbasis of this ρ consists of states

$$\begin{aligned} |\psi_1\rangle &= \frac{1}{\sqrt{4-2\sqrt{2}}}\begin{bmatrix} 1 \\ \sqrt{2}-1 \end{bmatrix} \text{ with eigenvalue } p_1 = (2+\sqrt{2})/4, \\ |\psi_2\rangle &= \frac{1}{\sqrt{4-2\sqrt{2}}}\begin{bmatrix} \sqrt{2}-1 \\ -1 \end{bmatrix} \text{ with eigenvalue } p_2 = (2-\sqrt{2})/4. \end{aligned}$$

Thus $\rho = p_1|\psi_1\rangle\langle\psi_1| + p_2|\psi_2\rangle\langle\psi_2|$. So if Carol prepares $|\psi_1\rangle\langle\psi_1|$ with probability $p_1 = (2+\sqrt{2})/4$ and $|\psi_2\rangle\langle\psi_2|$ with probability $p_2 = (2-\sqrt{2})/4$, then ships her state to Bob, then Bob (who doesn't see who the sender is) can't tell with any advantage over guessing who sent him the state.

Exercise 23.8 Do a similar analysis as that above, this time assuming Alice sends $(4|0\rangle+3|1\rangle)(4\langle 0|+3\langle 1|)/25$ with probability 1/2 and $(4|0\rangle-3|1\rangle)(4\langle 0|-3\langle 1|)/25$ with probability 1/2.

Exercise 23.9 Prove that any convex combination of states (pure or mixed) is a state.

One-Qubit States and the Bloch Sphere. Recall that we have a nice geometrical representation of one-qubit pure states: for each one-qubit pure state ρ there correspond unique $x, y, z \in \mathbb{R}$ such that $x^2 + y^2 + z^2 = 1$ and $\rho = (I + xX + yY + zZ)/2$, and conversely, for any point (x, y, z) on the unit sphere in $\mathbb{R}^3$ (Bloch sphere), the operator $(I + xX + yY + zZ)/2$ is a one-qubit pure state.

Can we characterize general one-qubit states in a similarly geometrical way? Yes. Let $\rho = \sum_{i=1}^{k} p_i \rho_i$ be any one-qubit state, where the ρ_i are one-qubit pure states and the p_i form a probability distribution as usual. For $1 \leqslant i \leqslant k$, let (x_i, y_i, z_i) be the point on the Bloch sphere such that $\rho_i = (I + x_i X + y_i Y + z_i Z)/2$. Then by linearity we have

$$\rho = \sum_{i=1}^{k} p_i \rho_i = \sum_i p_i \left(\frac{I + x_i X + y_i Y + z_i Z}{2} \right) = \frac{I + xX + yY + zZ}{2},$$

where $(x, y, z) := \sum_{i=1}^{k} p_i (x_i, y_i, z_i) \in \mathbb{R}^3$. That is, ρ corresponds geometrically to the point $(x, y, z) \in \mathbb{R}^3$ that is the convex combination of all the points (x_i, y_i, z_i), weighted by the same probabilities p_i used to weight ρ in terms of the ρ_i. We note that

$$\sqrt{x^2 + y^2 + z^2} = \|(x, y, z)\| \leqslant \sum_{i=1}^{k} p_i \|(x_i, y_i, z_i)\| = \sum_i p_i = 1,$$

and the inequality is strict iff there are at least two distinct points (x_i, y_i, z_i) on the sphere with $p_i > 0$. This means that the point (x, y, z) is somewhere on or inside the Bloch sphere. The surface points of the Bloch sphere correspond to the pure states, and the points in the interior correspond to mixed states. A one-qubit unitary U rotates a mixed state ρ in the interior just as it does points on the surface of the sphere (it rotates all of $\mathbb{R}^3$, in fact).

We can get some important facts about ρ based on its geometry. For example, if $\rho = (I + xX + yY + zZ)/2$, then let $r = \|(x, y, z)\| = (x^2 + y^2 + z^2)^{1/2} \leqslant 1$ be the distance from (x, y, z) to the origin. Then the eigenvalues of ρ are $(1 \pm r)/2$, and if $r > 0$, the corresponding eigenvectors are the states corresponding to the antipodal points $\pm(x, y, z)/r$ on the surface of the sphere. ($(I + (x/r)X + (y/r)Y + (z/r)Z)/2$ has eigenvalue $(1 + r)/2$, while $(I - (x/r)X - (y/r)Y - (z/r)Z)/2$ has eigenvalue $(1 - r)/2$, which are the two probabilities in the eigenvalue distribution of ρ.) These are the points where the line through $(0, 0, 0)$ and (x, y, z) intersects the surface of the sphere. (If $r = 0$, then (x, y, z) is the origin, $\rho = I/2$, and every nonzero vector is an eigenvector with eigenvalue $1/2$.)

Exercise 23.10 Prove all the assertions in the paragraph above. [Hint: You could certainly compute the eigenvectors and eigenvalues of ρ by brute force if you had to. Alternatively, you might note that if you let $\rho_1 = |\psi_1\rangle\langle\psi_1| = (I + (x/r)X + (y/r)Y + (z/r)Z)/2$ and $\rho_2 = |\psi_2\rangle\langle\psi_2| = (I - (x/r)X - (y/r)Y - (z/r)Z)/2$, then $\langle\psi_1|\psi_2\rangle = 0$ because the corresponding points are antipodal, and further, ρ is a convex combination of ρ_1 and ρ_2. What are the coefficients of this combination in terms of r? What does the matrix of ρ look like in the $\{|\psi_1\rangle, |\psi_2\rangle\}$ basis?]

24 Week 12: Quantum channels (quantum operations)

The Partial Trace. We sometimes have a system T that we are interested in coupling with another system S that we are *not* interested in, producing an entangled state in the combined system ST. This happens, for example, when a quantum computation (system T), rather than proceeding in perfect isolation, gets corrupted by an unintended interaction with its environment (system S), *e.g.*, a cosmic ray hitting the quantum computing device. Since we only care about system T, does it make sense to ask, "what is the state of T?" even though it is entangled with S? The partial trace operator lets us do just that.

Let $\mathcal{H}_S$ and $\mathcal{H}_T$ be Hilbert spaces. There is a unique linear map $\mathrm{tr}_S : \mathcal{L}(\mathcal{H}_S \otimes \mathcal{H}_T) \to \mathcal{L}(\mathcal{H}_T)$ such that for every $A \in \mathcal{L}(\mathcal{H}_S)$ and $B \in \mathcal{L}(\mathcal{H}_T)$,

$$\mathrm{tr}_S(A \otimes B) = (\mathrm{tr}\, A)B. \tag{84}$$

The map tr_S is an example of a *partial trace*. When we apply tr_S, we often say that we are *tracing out* the system S. There can be only one linear map satisfying (84), because $\mathcal{L}(\mathcal{H}_S \otimes \mathcal{H}_T) \cong \mathcal{L}(\mathcal{H}_S) \otimes \mathcal{L}(\mathcal{H}_T)$ is spanned by tensor products of operators. Suppose $\mathcal{H}_S$ has dimension m and $\mathcal{H}_T$ has dimension n. Suppose some operator $C \in \mathcal{L}(\mathcal{H}_S \otimes \mathcal{H}_T)$ is written in block matrix form with respect to some product basis:

$$C = \left[\begin{array}{c|c|c|c} B_{11} & B_{12} & \cdots & B_{1m} \\ \hline B_{21} & B_{22} & \cdots & B_{2m} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline B_{m1} & B_{m2} & \cdots & B_{mm} \end{array}\right],$$

where each block B_{ij} is an $n \times n$ matrix. Then we can also write C uniquely as

$$C = \sum_{i,j=1}^{m} E_{ij} \otimes B_{ij},$$

where E_{ij} is the $m \times m$ matrix whose (i,j)th entry is 1 and all other entries 0. The partial trace of C is then given in matrix form as

$$\mathrm{tr}_S(C) = \sum_{i,j} (\mathrm{tr}\, E_{ij}) B_{ij} = \sum_{i=1}^{m} B_{ii},$$

which is the sum of all the diagonal blocks of C and is an $n \times n$ matrix.

We may alternatively trace out the system T via the unique linear map $\mathrm{tr}_T : \mathcal{L}(\mathcal{H}_S \otimes \mathcal{H}_T) \to \mathcal{L}(\mathcal{H}_T)$ that satisfies

$$\mathrm{tr}_T(A \otimes B) = (\mathrm{tr}\, B)A$$

for any $A \in \mathcal{L}(\mathcal{H}_S)$ and $B \in \mathcal{L}(\mathcal{H}_T)$. If C is as above, then

$$\mathrm{tr}_T(C) = \sum_{i,j=1}^{m} (\mathrm{tr}\, B_{ij}) E_{ij} = \begin{bmatrix} \mathrm{tr}\, B_{11} & \mathrm{tr}\, B_{12} & \cdots & \mathrm{tr}\, B_{1m} \\ \mathrm{tr}\, B_{21} & \mathrm{tr}\, B_{22} & \cdots & \mathrm{tr}\, B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{tr}\, B_{m1} & \mathrm{tr}\, B_{m2} & \cdots & \mathrm{tr}\, B_{mm} \end{bmatrix},$$

which is an $m \times m$ matrix.

The partial trace operator extends in a similar way to combinations of several systems at once. Intuitively, tracing out a system is a bit like averaging over the system. In tensor algebra, the partial trace operators and the (total) trace operator are called *contractions*.

If system ST is in some separable (*i.e.*, tensor product) state $\rho = \rho_S \otimes \rho_T \in \mathcal{L}(\mathcal{H}_S \otimes \mathcal{H}_T)$, where $\rho_S \in \mathcal{L}(\mathcal{H}_S)$ and $\rho_T \in \mathcal{L}(\mathcal{H}_T)$ are states in S and T, respectively, then $\mathrm{tr}_S\, \rho = (\mathrm{tr}\, \rho_S)\rho_T = \rho_T$ and $\mathrm{tr}_T\, \rho = (\mathrm{tr}\, \rho_T)\rho_S = \rho_S$. So we can say unequivocably that the system S is in state $\mathrm{tr}_T\, \rho$ and the system T is in state $\mathrm{tr}_S\, \rho$. If ρ is entangled, then we can still say that system S is in state $\mathrm{tr}_T\, \rho$ and T is in $\mathrm{tr}_S\, \rho$, but now these two states (called *reduced* states) are mixed states, even if ρ itself is a pure state. Thus by tracing out one or the other system, we will lose some information about the state of the remaining system if the original combined state was entangled.

Open Systems and Quantum Channels. A closed quantum system is one that does not interact with the outside world. Closed systems evolve unitarily. An open quantum system does couple with one or more other systems (collectively called the *environment*) that we wish to ignore. By considering open systems, we will obtain a powerful formalism for describing what can happen to a quantum system that may interact with its environment. This formalism, the formalism of quantum channels (sometimes called quantum operations) is general enough to encompass both unitary evolution and measurements. A quantum channel is a certain linear map that maps states in one Hilbert space to states in another. All physical processes, including unitary evolution, measurements, etc., or any combination of these, are modeled mathematically as quantum channels.

Definition 24.1 For Hilbert spaces $\mathcal{H}$ and $\mathcal{J}$, we let $\mathcal{T}(\mathcal{H}, \mathcal{J})$ denote the (Hilbert) space $\mathcal{L}(\mathcal{L}(\mathcal{H}), \mathcal{L}(\mathcal{J}))$ of all linear maps from $\mathcal{L}(\mathcal{H})$ into $\mathcal{L}(\mathcal{J})$. We write $\mathcal{T}(\mathcal{H})$ to mean $\mathcal{T}(\mathcal{H}, \mathcal{H})$. A map $\Phi \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ is sometimes called a *superoperator*.

Since a quantum state is an operator over a Hilbert space, all quantum channels are superoperators, mapping states of one Hilbert space $\mathcal{H}$ to states of another (or the same) Hilbert space $\mathcal{J}$. Not all superoperators are quantum channels, however. As we will optionally see below, there are some simple conditions on a superoperator Φ that makes it a quantum channel. One such condition is that Φ must be trace-preserving, which is necessary so that Φ maps states to states, all of which have unit trace. The other condition is that Φ be *completely positive*, a concept we will discuss later. At first, we will only consider the case where $\mathcal{H} = \mathcal{J}$, *i.e.*, where the channel maps states to states in the same system, but later we will see how to generalize to arbitrary $\mathcal{H}$ and $\mathcal{J}$.

Throughout this section, we will avoid Dirac notation, as it tends to get in the way.

There are a number of different, equivalent ways of representing a quantum channel. We consider two here: the *coupled-systems representation* (aka the *Stinespring representation*), where we include the environment then trace it out, and the *operator-sum representation* (aka the representation by *Kraus operators*), where we simply apply operators to states in the system without mentioning the environment. We'll show that these two views are equivalent. The coupled-systems view is more physically intuitive, while the operator-sum view is more mathematically convenient.

We now formally describe a quantum channel $\mathcal{E}$ on some system S according to the coupled-systems view. We imagine S (state space $\mathcal{H}_S$ of dimension n) in some state ρ. We now consider another system T (whose state space $\mathcal{H}_T$ has dimension N), in some known or prepared pure state σ, initially isolated from system S. The combined state of TS is then $\sigma \otimes \rho$ initially.[21] We now couple T and S together and let the combined system TS evolve according to some unitary operator $U \in \mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$, resulting in the state $U(\sigma \otimes \rho)U^*$. We now "forget" the system T by tracing it out, obtaining the final state

$$\rho' = \mathcal{E}(\rho) := \mathrm{tr}_T(U(\sigma \otimes \rho)U^*) \in \mathcal{L}(\mathcal{H}_S). \tag{85}$$

More generally, for any $X \in \mathcal{L}(\mathcal{H})$, we define

$$X' = \mathcal{E}(X) := \mathrm{tr}_T(U(\sigma \otimes X)U^*) \in \mathcal{L}(\mathcal{H}_S). \tag{86}$$

Because all the components making up the definition of $\mathcal{E}$ in (85) are linear maps, $\mathcal{E}$ itself is linear, mapping $\mathcal{L}(\mathcal{H}_S)$ into $\mathcal{L}(\mathcal{H}_S)$, and thus $\mathcal{E} \in \mathcal{T}(\mathcal{H}_S)$ is a superoperator. $\mathcal{E}$ depends implicitly on the system T, its initial state σ, and U.

At first blush, the operator sum formulation of the quantum channel $\mathcal{E}$ looks completely different. We pick some finite collection of operators $K_1, \ldots, K_N \in \mathcal{L}(\mathcal{H}_S)$ (for some $N \geqslant 1$) that are completely arbitrary except that we must have

$$\sum_{j=1}^{N} K_j^* K_j = I\,. \tag{87}$$

We then define, for any $X \in \mathcal{L}(\mathcal{H}_S)$,

$$X' = \mathcal{E}(X) := \sum_{j=1}^{N} K_j X K_j^* \in \mathcal{L}(\mathcal{H}_S). \tag{88}$$

Defined this way, the map $\mathcal{E}$ is evidently linear from $\mathcal{L}(\mathcal{H}_S)$ into itself, and so $\mathcal{E} \in \mathcal{T}(\mathcal{H})$ is a superoperator, and it depends implicitly on the choice of $K_1, \ldots, K_N$, which are called *Kraus operators*. We'll show in a minute that the two definitions of $\mathcal{E}$ just described are equivalent.

Exercise 24.2 Verify that if ρ is a state (*i.e.*, $\rho \geqslant 0$ and $\mathrm{tr}\,\rho = 1$), then the operator $\rho' = \mathcal{E}(\rho)$ defined by (88) is also a state.

The next exercise shows that quantum channels include unitary evolution.

Exercise 24.3 Show that unitary evolution of the system S through a unitary operator $U \in \mathcal{L}(\mathcal{H}_S)$ is a legitimate quantum channel. Argue with respect to both views of quantum channels.

[21] The textbook puts the auxilliary system T on the right, whereas we put it on the left. The two ways are equivalent, but ours will be more consistent with a block matrix representation we'll use later when we prove equivalence of the two views.

For another example, suppose we make a projective measurement on the system S in state ρ—using some complete set $\{P_1, \ldots, P_k\}$ of orthogonal projectors in $\mathcal{L}(\mathcal{H}_S)$—but we don't bother to look at what the outcome of the measurement is. Then for all we know, the post-measurement state of S will be a mixture of the post-measurement states corresponding to all the possible outcomes, weighted by their probabilities. That is, using Equation (30), the state of S after this "information-free" measurement should be[22]

$$\rho' = \sum_{j=1}^{k} \Pr[j] \frac{P_j \rho P_j}{\Pr[j]} = \sum_j P_j \rho P_j^*.$$

This looks like the operator sum representation of a quantum channel (Equation (88)), and indeed we have

$$I = \sum_{j=1}^{k} P_j = \sum_j P_j^2 = \sum_j P_j^* P_j,$$

because the P_j form a complete set of projectors. Thus $P_1, \ldots, P_k$ satisfy Equation (87) to be Kraus operators, and this information-free measurement is a quantum channel.

Equivalence of the Coupled-Systems and Operator-Sum Representations. First we'll show that every quantum channel defined by the coupled system definition has an operator sum representation. Suppose that $\mathcal{E} \in \mathcal{T}(\mathcal{H}_S)$ is defined so that for all $X \in \mathcal{L}(\mathcal{H}_S)$,

$$\mathcal{E}(X) = \mathrm{tr}_T(U(\sigma \otimes X)U^*),$$

where T is a system with state space $\mathcal{H}_T$, $\sigma \in \mathcal{L}(\mathcal{H}_T)$ is a pure state (1-dimensional projector), and $U \in \mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$ is unitary. Let $n = \dim(\mathcal{H}_S)$ and let $N = \dim(\mathcal{H}_T)$. We'll pick a product basis for $\mathcal{H}_T \otimes \mathcal{H}_S$ so that we can work directly with matrices. Let $\{e_1, \ldots, e_N\}$ be an orthonormal basis for $\mathcal{H}_T$ and let $\{f_1, \ldots, f_n\}$ be an orthonormal basis for $\mathcal{H}_S$. We can choose these bases arbitrarily, so we'll assume that $\sigma = e_1 e_1^*$, *i.e.*, σ projects onto the 1-dimensional subspace spanned by e_1. With respect to the product basis $\{e_i \otimes f_j : 1 \leqslant i \leqslant N \;\&\; 1 \leqslant j \leqslant n\}$, the operator U can be written uniquely in block matrix form as

$$U = \sum_{a,b=1}^{N} E_{ab} \otimes B_{ab},$$

where each B_{ab} is an $n \times n$ matrix, and each $E_{ab} := e_a e_b^*$ is the $N \times N$ matrix whose (a,b)th entry is 1 and all the other entries are 0. Noting that $E_{ab}E_{cd} = e_a e_b^* e_c e_d^* = \langle e_b, e_c\rangle e_a e_d^* = \delta_{bc} E_{ad}$ and $E_{cd}^* = E_{dc}$, we have

$$\begin{aligned} U(e_1 e_1^* \otimes X)U^* &= \sum_{a,b,c,d=1}^{N} (E_{ab} \otimes B_{ab})(E_{11} \otimes X)(E_{cd} \otimes B_{cd})^* && (89)\\ &= \sum_{a,b,c,d} E_{ab}E_{11}E_{dc} \otimes B_{ab}XB_{cd}^* && (90)\\ &= \sum_{a,c} E_{ac} \otimes B_{a1}XB_{c1}^* \,. && (91) \end{aligned}$$

[22] A minor technical point: To be well-defined, the first sum in the next equation is really only over those j for which $\Pr[j] > 0$. However, the second sum is over all j. The two sums are still equal, because if $\Pr[j] = \mathrm{tr}(P_j \rho P_j^*) = 0$ for some j, then $P_j \rho P_j^* = 0$ by Exercise 9.28.

Tracing out the first component of each tensor product, and using the fact that $\operatorname{tr} E_{ac} = \delta_{ac}$, we get

$$\mathcal{E}(X) = \operatorname{tr}_T(U(e_1 e_1^* \otimes X)U^*) = \sum_{a,c=1}^{N} (\operatorname{tr} E_{ac}) B_{a1} X B_{c1}^* = \sum_{a=1}^{N} B_{a1} X B_{a1}^* \,, \tag{92}$$

which has the form of (88) if we let $K_a := B_{a1}$. We're done if (87) holds. Let I_T and I_S be the identity operators in $\mathcal{L}(\mathcal{H}_T)$ and $\mathcal{L}(\mathcal{H}_S)$, respectively, and define $I_{TS} := I_T \otimes I_S$, which is the identity on $\mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$. Since U is unitary, we have

$$\begin{aligned} I_{TS} \;=\; U^*U \;&=\; \sum_{a,b,c,d=1}^{N} (E_{ab} \otimes B_{ab})^* (E_{cd} \otimes B_{cd}) \\ &=\; \sum_{a,b,c,d} E_{ba} E_{cd} \otimes B_{ab}^* B_{cd} \\ &=\; \sum_{a,b,d} E_{bd} \otimes B_{ab}^* B_{ad} \\ &=\; \sum_{b,d=1}^{N} E_{bd} \otimes \left(\sum_{a} B_{ab}^* B_{ad} \right). \end{aligned}$$

We want to isolate what is in the parentheses for $b = d = 1$ and show that it is the identity I_S. We can do this by multiplying both sides on the left and right with the Hermitean operator $E_{11} \otimes I_S$ and then tracing out T. For the left-hand side, we get

$$\operatorname{tr}_T((E_{11} \otimes I_S) I_{TS} (E_{11} \otimes I_S)) = \operatorname{tr}_T((E_{11} \otimes I_S)(I_T \otimes I_S)(E_{11} \otimes I_S)) = \operatorname{tr}_T(E_{11} \otimes I_S) = (\operatorname{tr} E_{11}) I_S = I_S.$$

For the right-hand side, using linearity of tr_T, we get

$$\begin{aligned} &\sum_{b,d=1}^{N} \operatorname{tr}_T \left((E_{11} \otimes I_S) \left(E_{bd} \otimes \left(\sum_{a=1}^{N} B_{ab}^* B_{ad} \right) \right) (E_{11} \otimes I_S) \right) \\ &=\; \sum_{b,d=1}^{N} \operatorname{tr}_T \left(E_{11} E_{bd} E_{11} \otimes \sum_{a} B_{ab}^* B_{ad} \right) \\ &=\; \sum_{b,d=1}^{N} \operatorname{tr}_T \left(\delta_{1b} \delta_{d1} E_{11} \otimes \sum_{a} B_{ab}^* B_{ad} \right) \\ &=\; \sum_{b,d=1}^{N} \delta_{1b} \delta_{d1} \operatorname{tr}_T \left(E_{11} \otimes \sum_{a} B_{ab}^* B_{ad} \right) \\ &=\; \operatorname{tr}_T \left(E_{11} \otimes \sum_{a} B_{a1}^* B_{a1} \right) \\ &=\; (\operatorname{tr} E_{11}) \sum_{a} B_{a1}^* B_{a1} = \sum_{a} B_{a1}^* B_{a1} \end{aligned}$$

as intended. Equating both sides, we then have

$$\sum_{a=1}^{N} B_{a1}^* B_{a1} = I_S \,,$$

which means that (87) holds for Kraus operators $B_{11}, \ldots, B_{N1}$, and we have a legitimate operator sum representation of $\mathcal{E}$.

We'll now show the other direction. Suppose we are given an operator sum representation of $\mathcal{E}$ in the form of some collection $K_1, \ldots, K_N$ of Kraus operators such that $\sum_{j=1}^{N} K_j^* K_j = I_S$. That is $\mathcal{E}(X) = \sum_{a=1}^{N} K_a X K_a^*$ for all $X \in \mathcal{L}(\mathcal{H}_S)$. We want to find a coupled-systems representation of $\mathcal{E}$. As before, we will fix some orthonormal basis $\{f_j\}_{1 \leqslant j \leqslant n}$ of $\mathcal{H}_S$, so that we can talk about matrices instead of operators. Define K to be the $nN \times n$ matrix

$$K = \left[\begin{array}{c} K_1 \\ \hline \vdots \\ \hline K_N \end{array}\right].$$

The condition that $\sum_{j=1}^{N} K_j^* K_j = I$ can be written in block matrix form as

$$K^* K = \left[\begin{array}{c|c|c|c} K_1^* & \cdots & K_N^* \end{array}\right] \left[\begin{array}{c} K_1 \\ \hline \vdots \\ \hline K_N \end{array}\right] = I_S \,. \tag{93}$$

Here we are multiplying an $n \times nN$ matrix on the left and an $nN \times n$ matrix on the right to get the $n \times n$ identity matrix. Consider the columns of K as nN-dimensional column vectors. Equation (93) is equivalent to saying that the columns of K form an orthonormal set. By Gram-Schmidt, we can take these column vectors as the first n vectors in an orthonormal basis for $\mathbb{C}^{nN}$. We assemble these basis vectors as the columns of an $nN \times nN$ matrix U written in block form by

$$U = \left[\begin{array}{c|c|c|c} B_{11} & B_{12} & \cdots & B_{1N} \\ \hline B_{21} & B_{22} & \cdots & B_{2N} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline B_{N1} & B_{N2} & \cdots & B_{NN} \end{array}\right] = \sum_{a,b=1}^{N} E_{ab} \otimes B_{ab} \,,$$

where each B_{ab} is an $n \times n$ matrix, and the first n columns of U form K, *i.e.*, $K_a = B_{a1}$ for $1 \leqslant a \leqslant N$. The orthonormality of the columns of U is equivalent to the equation $U^* U = I$, and so U is unitary. Now let $\mathcal{H}_T$ be any N-dimensional Hilbert space, and fix an orthonormal basis $\{e_i\}_{1 \leqslant i \leqslant N}$ for $\mathcal{H}_T$. Then with respect to the product basis, U can be considered a unitary operator in $\mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$, and so now we follow the string of equations of (92) to see that $\mathcal{E}(X) = \mathrm{tr}_T(U(e_1 e_1^* \otimes X)U^*)$ for any $X \in \mathcal{L}(\mathcal{H}_S)$. Indeed, for all $X \in \mathcal{L}(\mathcal{H}_S)$ we have

$$\begin{aligned} \mathrm{tr}_T(U(e_1 e_1^* \otimes X)U^*) &= \sum_{a,b,c,d=1}^{N} \mathrm{tr}_T\left((E_{ab} \otimes B_{ab})(e_1 e_1^* \otimes X)(E_{cd} \otimes B_{cd})^*\right) \\ &= \sum_{a,b,c,d} \mathrm{tr}_T\left((E_{ab} \otimes B_{ab})(E_{11} \otimes X)(E_{dc} \otimes B_{cd}^*)\right) \\ &= \sum_{a,b,c,d} \mathrm{tr}_T\left(E_{ab} E_{11} E_{dc} \otimes B_{ab} X B_{cd}^*\right) \\ &= \sum_{a,b,c,d} \mathrm{tr}_T\left(\delta_{b1}\delta_{1d} E_{ac} \otimes B_{ab} X B_{cd}^*\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{a,c} \mathrm{tr}_T\,(E_{ac} \otimes B_{a1} X B_{c1}^*) \\
&= \sum_{a,c} (\mathrm{tr}\, E_{ac}) B_{a1} X B_{c1}^* \\
&= \sum_{a,c} \delta_{ac}\; B_{a1} X B_{c1}^* \\
&= \sum_{a=1}^{N} B_{a1} X B_{a1}^* = \sum_{a=1}^{N} K_a X K_a^* = \mathcal{E}(X)\ .
\end{aligned}
$$

This is then a coupled-systems representation of $\mathcal{E}$.

A Normal Form for the Kraus Operators. The choices of Kraus operators in the operator-sum representation (88) of an quantum channel $\mathcal{E}$ are not unique. Neither is the form of the unitary U in the coupled-systems representation of (85). The freedom in the coupled systems case can be seen as follows: Suppose $A \in \mathcal{L}(\mathcal{H}_T)$ and $B \in \mathcal{L}(\mathcal{H}_S)$ are any operators, and suppose $V \in \mathcal{L}(\mathcal{H}_T)$ is unitary. Then

$$\mathrm{tr}_T((V \otimes I)(A \otimes B)(V \otimes I)^*) = \mathrm{tr}_T(VAV^* \otimes B) = (\mathrm{tr}(VAV^*))B = (\mathrm{tr}\, A)B = \mathrm{tr}_T(A \otimes B).$$

Since tr_T is linear, this extends to

$$\mathrm{tr}_T[(V \otimes I)C(V \otimes I)^*] = \mathrm{tr}_T C$$

for every $C \in \mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$. In other words, if we eventually trace out the environment T, then it doesn't matter if we evolve T's state unitarily or not. Let's conjugate Equations (89–91) by $V \otimes I$, where V is an $N \times N$ unitary matrix and I is the $n \times n$ identity matrix. Noting that $V \otimes I = \sum_{a,b=1}^{N} [V]_{ab} E_{ab} \otimes I$, we get

$$
\begin{aligned}
&(V \otimes I)U(e_1 e_1^* \otimes X)U^*(V \otimes I)^* \\
&= \sum_{a,b,c,d,e,f,g,h=1}^{N} ([V]_{ef} E_{ef} \otimes I)(E_{ab} \otimes B_{ab})(E_{11} \otimes X)(E_{cd} \otimes B_{cd})^*([V]_{gh} E_{gh} \otimes I)^* \\
&= \sum_{a,\ldots,h} [V]_{ef} [V]_{gh}^* (E_{ef} E_{ab} E_{11} E_{dc} E_{hg}) \otimes (B_{ab} X B_{cd}^*) \\
&= \sum_{a,c,e,g} [V]_{ea} [V]_{gc}^* E_{eg} \otimes B_{a1} X B_{c1}^*\ .
\end{aligned}
$$

Tracing out T, we have

$$
\begin{aligned}
\mathcal{E}(X) &= \mathrm{tr}_T(U(e_1 e_1^* \otimes X)U^*) \\
&= \mathrm{tr}_T((V \otimes I)U(e_1 e_1^* \otimes X)U^*(V \otimes I)^*) \\
&= \sum_{a,c,e,g} [V]_{ea} [V]_{gc}^*\; \mathrm{tr}_T(E_{eg} \otimes B_{a1} X B_{c1}^*) \\
&= \sum_{a,c,e,g} [V]_{ea} [V]_{gc}^* (\mathrm{tr}\, E_{eg}) B_{a1} X B_{c1}^*
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{a,c,e} [V]_{ea} [V]^*_{ec} B_{a1} X B^*_{c1} \\
&= \sum_{e} \left(\sum_{a} [V]_{ea} B_{a1} \right) X \left(\sum_{c} [V]^*_{ec} B_{c1} \right) \\
&= \sum_{e} \widetilde{K}_e X \widetilde{K}^*_e ,
\end{aligned}
$$

where

$$
\widetilde{K}_e := \sum_{a=1}^{N} [V]_{ea} B_{a1} = \sum_{a=1}^{N} [V]_{ea} K_a \tag{94}
$$

for all $1 \leqslant e \leqslant N$. So these equations give us the effect of V on the Kraus operators.

Exercise 24.4 Show by direct calculation that if $K_1, \ldots, K_N \in \mathcal{L}(\mathcal{H}_S)$ are operators such that $\sum_{j=1}^{N} K_j^* K_j = I$, and for all $1 \leqslant j \leqslant N$ we define $\widetilde{K}_j := \sum_{a=1}^{N} [V]_{ja} K_a$ for some fixed $N \times N$ unitary matrix V, then

$$
\sum_{j=1}^{N} \widetilde{K}_j^* \widetilde{K}_j = I,
$$

and for every $X \in \mathcal{L}(\mathcal{H}_S)$,

$$
\sum_{j=1}^{N} \widetilde{K}_j X \widetilde{K}_j^* = \sum_{j=1}^{N} K_j X K_j^* .
$$

So we are allowed to choose V to be any unitary matrix we want without affecting the quantum channel. We'll pick a specific V as follows: Given any set of Kraus operators $K_1, \ldots, K_N$, let T be the $N \times N$ matrix whose (i,j)th entry is

$$
[T]_{ij} := \langle K_j, K_i \rangle = \mathrm{tr}(K_j^* K_i),
$$

for $1 \leqslant i, j \leqslant N$. Note that $[T]_{ij} = \langle K_j, K_i \rangle = \langle K_i, K_j \rangle^* = [T]^*_{ji}$, and so T is a Hermitean matrix. Since T is normal, we can choose V so that VTV^* is a diagonal matrix. Now defining $\widetilde{K}_j := \sum_{a=1}^{N} [V]_{ja} K_a$ as in (94), we get, for all $1 \leqslant i, j \leqslant N$,

$$
\begin{aligned}
\left\langle \widetilde{K}_i , \widetilde{K}_j \right\rangle &= \sum_{a,b=1}^{N} [V]^*_{ia} [V]_{jb} \langle K_a, K_b \rangle \\
&= \sum_{a,b} [V]_{jb} [T]_{ba} [V^*]_{ai} \\
&= [VTV^*]_{ji} \\
&= \lambda_j \delta_{ij}
\end{aligned}
$$

for some values $\lambda_1, \ldots, \lambda_N$, because VTV^* is diagonal. Thus the $\widetilde{K}_j$ are pairwise orthogonal with respect to the Hilbert-Schmidt inner product, and hence linearly independent. Since the $\widetilde{K}_j$ occupy an n^2-dimensional space, there can only be at most n^2 many of them that are nonzero.

Hence we have a normal form for the operator-sum representation of a quantum channel: Any quantum channel on an n-dimensional state space may be represented by $N \leqslant n^2$ many Kraus operators that are pairwise orthogonal.

Exercise 24.5 Explain why the values $\lambda, \ldots, \lambda_N$ above are all nonnegative reals.

Quantum Channels Between Different Hilbert Spaces. We have restricted our attention to quantum channels of the form $\mathcal{E} : \mathcal{L}(\mathcal{H}) \to \mathcal{L}(\mathcal{H})$, that is, linear maps that map operators of a space to operators of the same space. This restriction is unnecessary, and it is easy to imagine quantum channels mapping states in one space to states in another. The partial trace operator itself is a good example of such a thing. The operator-sum view is the easiest way to characterize these more general quantum channels. We will satisfy ourselves with the following general definition, without going into the details of why it is the best one. It certainly coincides with our previous view in the case where the two spaces are the same.

Definition 24.6 Let $\mathcal{H}$ and $\mathcal{J}$ be Hilbert spaces. A *quantum channel* from $\mathcal{H}$ to $\mathcal{J}$ is a superoperator $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ such that there exists an integer $N > 0$ and linear maps $K_j : \mathcal{H} \to \mathcal{J}$ for $1 \leqslant j \leqslant N$ satisfying the completeness condition $\sum_{j=1}^{N} K_j^* K_j = I_{\mathcal{H}}$, such that for every $X \in \mathcal{L}(\mathcal{H})$,

$$\mathcal{E}(X) = \sum_{j=1}^{N} K_j X K_j^* \, .$$

Here, $I_{\mathcal{H}}$ denotes the identity map on $\mathcal{H}$. As in the special case where $\mathcal{H} = \mathcal{J}$, the K_j are known as *Kraus operators*.

Recall that for any linear map $K : \mathcal{H} \to \mathcal{J}$, the adjoint K^* is a uniquely defined linear map from $\mathcal{J}$ to $\mathcal{H}$. $\mathcal{E}$ maps positive operators to positive operators, and the completeness condition guarantees that $\mathcal{E}$ is trace-preserving.

General Measurements. I didn't lecture on this in class. The textbook makes a rather significant logical mistake in its discussion of quantum measurement, starting in Section 2.2.3 but carrying over into Chapter 8. Except for alerting you to this mistake, this material is optional.

Postulate 3 on pages 84–85 (reformulated in terms of density operators on page 102) describes general measurements where some classical information may be obtained. In it, they describe a (general) quantum measurement on a system with state space $\mathcal{H}$ as an indexed collection $\{M_m\}_{m \in \mathcal{I}}$ of operators (called *measurement operators*) in $\mathcal{L}(\mathcal{H})$ satisfying $\sum_{m \in \mathcal{I}} M_m^* M_m = I$. (Thus the M_m satisfy the same completeness condition as the Kraus operators did previously.) I use $\mathcal{I}$ here again to describe the set of possible outcomes. According to the postulate, when a system in state ρ is measured using $\{M_m\}$, the probability of seeing an outcome $m \in \mathcal{I}$ is given by $\langle M_m^* M_m, \rho \rangle$, and the post-measurement state assuming outcome m occurred is $M_m \rho M_m^* / \langle M_m^* M_m, \rho \rangle$.

All of this is fine except that it is not general enough. There are legitimate physical measurements that do not take this form. The measurements described by the book are all guaranteed to produce pure states after the measurement, assuming that a pure state was measured. There are,

however, more "imprecise" measurements that may yield mixed states after the measurement, even if the pre-measurement state was pure.

As before with quantum channels, there are two equivalent views of a general measurement: the coupled-systems view and the operator-sum view. The textbook gives the operator-sum view. I'll describe both views, pointing out how the true operator-sum view differs from the text, but I'll omit the proof of equivalence, which is very similar to what I did earlier with quantum channels. If you want a chance to practice "index gymnastics" yourself, I'll leave the details of the proof to you as an exercise.

We'll only consider finitary measurements here, *i.e.*, measurements with only a finite set of possible outcomes. One can generalize our analysis to infinitary measurements as well.

In the coupled-systems view, a general measurement on a system S with state space $\mathcal{H}_S$ proceeds as follows, assuming ρ is the pre-measurement state of S:

1. Prepare another system T with (finite dimensional) state space $\mathcal{H}_T$ in some initial pure state σ.

2. Couple T with S, and let the combined system evolve unitarily according to some unitary $U \in \mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$, producing the state $U(\sigma \otimes \rho)U^*$.

3. Perform a projective measurement on the system TS, using some complete set $\{P^{(m)} : m \in \mathcal{I}\}$ of orthogonal projectors in $\mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S)$. $\mathcal{I}$ is the (finite) set of possible outcomes. By the usual rules, the probability of seeing any outcome $m \in \mathcal{I}$ is $\Pr[m] = \left\langle P^{(m)}, U(\sigma \otimes \rho)U^* \right\rangle$, and, assuming m is the outcome, the post-measurement state of TS is

$$\rho_m^{TS} := \frac{P^{(m)}U(\sigma \otimes \rho)U^*(P^{(m)})^*}{\Pr[m]} = \frac{P^{(m)}U(\sigma \otimes \rho)(P^{(m)}U)^*}{\mathrm{tr}(P^{(m)}U(\sigma \otimes \rho)(P^{(m)}U)^*)}.$$

4. Trace out the system T of the post-measurement state to obtain the post-measurement state of S:

$$\rho_m^S := \frac{\mathrm{tr}_T(P^{(m)}U(\sigma \otimes \rho)(P^{(m)}U)^*)}{\mathrm{tr}(P^{(m)}U(\sigma \otimes \rho)(P^{(m)}U)^*)}.$$

Remember that the projectors satisfy

$$\sum_{m \in \mathcal{I}} P^{(m)} = \sum_{m \in \mathcal{I}} (P^{(m)})^* P^{(m)} = I \in \mathcal{L}(\mathcal{H}_T \otimes \mathcal{H}_S).$$

This means that we have

$$\sum_{m \in \mathcal{I}} (P^{(m)}U)^* P^{(m)}U = U^* \left(\sum_{m \in \mathcal{I}} (P^{(m)})^* P^{(m)} \right) U = U^*U = I$$

as well.

In the operator-sum view, a general measurement $\mathcal{M}$ of system S is described by a (finite) set $\mathcal{I}$ of possible outcomes, and for each outcome $m \in \mathcal{I}$ a finite list $M_1^{(m)}, \ldots, M_N^{(m)}$ of operators in

$\mathcal{L}(\mathcal{H}_S)$,[23] all satisfying

$$\sum_{m\in\mathcal{I}}\sum_{j=1}^{N}(M_j^{(m)})^*M_j^{(m)} = I.$$

When system S in state ρ is measured according to $\mathcal{M}$, the probability of seeing an outcome m is given by

$$\Pr[m] := \operatorname{tr}\left(\sum_{j=1}^{N} M_j^{(m)}\rho(M_j^{(m)})^*\right) = \left\langle M^{(m)}, \rho\right\rangle, \tag{95}$$

where we set $M^{(m)} := \sum_{j=1}^{N}(M_j^{(m)})^*M_j^{(m)}$. If m is observed, then the post-measurement state of S is

$$\rho_m := \frac{\mathcal{E}_m(\rho)}{\Pr[m]} := \frac{\sum_{j=1}^{N} M_j^{(m)}\rho(M_j^{(m)})^*}{\Pr[m]}. \tag{96}$$

Here, we define $\mathcal{E}_m(\rho)$ to be the numerator of the right-hand side. Note that $\operatorname{tr}\mathcal{E}_m(\rho) = \Pr[m] \leqslant 1$. Any operator $\tau \geqslant 0$ such that $\operatorname{tr}\tau \leqslant 1$ will be called a *partial state* or an *incomplete state*. The incompleteness of $\mathcal{E}_m(\rho)$ is a reflection of the fact that it might not occur with certainty—its occurrence is conditioned on the outcome of the measurement being m. Similarly, we'll call $\mathcal{E}_m$ an *incomplete quantum channel*, since it might not be applied with certainty. $\mathcal{E}_m$ is clearly linear and maps positive operators to positive operators, but it is not (necessarily) trace-preserving. The map $\rho \mapsto \sum_{m\in\mathcal{I}}\mathcal{E}_m(\rho)$ is, however, a trace-preserving (*i.e.*, complete) quantum channel.

I'll finish with two remarks about Equations (95) and (96): First, it's easy to see that the operators $\{M^{(m)}\}_{m\in\mathcal{I}}$ of (95) form a POVM, and the converse is also true—any POVM arises from some general measurement where the post-measurement state is neglected. To see this, let $\{M^{(m)}\}_{m\in\mathcal{I}}$ be any (finitary) POVM. If we define measurement elements $K^{(m)} := \sqrt{M^{(m)}}$ for each $m \in \mathcal{I}$, then these elements form the operator-sum view of a generalized measurement, one operator per outcome, and the resulting outcome probabilities are the same as with the given POVM. Second, Postulate 3 only allows one operator per outcome, and so it is the special case of (96) where $N = 1$.

Completely Positive Maps. This is another optional topic. We've seen that every (complete) quantum channel $\mathcal{E}$ maps states to states; equivalently, it has two properties:

1. $\mathcal{E}$ preserves positivity, *i.e.*, if $A \geqslant 0$ then $\mathcal{E}(A) \geqslant 0$, and
2. $\mathcal{E}$ is trace-preserving, *i.e.*, $\operatorname{tr}\mathcal{E}(A) = \operatorname{tr}A$.

We'll see shortly that the converse does *not* hold. That is, there are linear maps satisfying (1) and (2) above that are not legitimate quantum channels according to Definition 24.6. To get a characterization, we need to strengthen (1) a bit. We say that $\mathcal{E}$ is *positive* if (1) holds, *i.e.*, if $\mathcal{E}$ maps positive operators to positive operators. The stronger condition we need is that $\mathcal{E}$ be *completely positive*—a condition that we now explain.

[23] Actually, the lists could contain different numbers of operators, but we can assume they are all the same length by padding shorter lists with copies of the zero operator.

Quantum channels are linear maps, and we can form tensor products of these linear maps just as we can with any linear maps. So, given two superoperators $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ and $\mathcal{F} \in \mathcal{T}(\mathcal{K}, \mathcal{M})$ ($\mathcal{H}$, $\mathcal{J}$, $\mathcal{K}$, and $\mathcal{M}$ are Hilbert spaces), we define $\mathcal{E} \otimes \mathcal{F}$ as usual to be the unique superoperator in $\mathcal{T}(\mathcal{H} \otimes \mathcal{K}, \mathcal{J} \otimes \mathcal{M})$ that takes $A \otimes B$ to $\mathcal{E}(A) \otimes \mathcal{F}(B)$ for every $A \in \mathcal{L}(\mathcal{H})$ and $B \in \mathcal{L}(\mathcal{K})$.

For every Hilbert space $\mathcal{H}$ we have the identity superoperator $\mathcal{I} \in \mathcal{T}(\mathcal{H})$ defined by $\mathcal{I}(A) = A$ for all $A \in \mathcal{L}(\mathcal{H})$. $\mathcal{I}$ is certainly a quantum channel, given by the single Kraus operator $I \in \mathcal{L}(\mathcal{H})$. The next definition gives the strengthening of property (1) that we need:

Definition 24.7 Let $\mathcal{H}$ and $\mathcal{J}$ be Hilbert spaces. A superoperator $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ is *completely positive* if for every Hilbert space $\mathcal{K}$, the map $\mathcal{I} \otimes \mathcal{E} \in \mathcal{T}(\mathcal{K} \otimes \mathcal{H}, \mathcal{K} \otimes \mathcal{J})$ is positive, where $\mathcal{I} \in \mathcal{T}(\mathcal{K})$ is the identity map on $\mathcal{L}(\mathcal{K})$.

If $\mathcal{E}$ is completely positive as in Definition 24.7, then $\mathcal{E}$ is also positive: If X is any operator in either $\mathcal{L}(\mathcal{H})$ or $\mathcal{L}(\mathcal{J})$, then it is easy to check that $X \geqslant 0$ if and only if $I \otimes X \geqslant 0$, where $I \in \mathcal{L}(\mathcal{K})$ is the identity operator. Then if $X \geqslant 0$, then $I \otimes X \geqslant 0$, and so by assumption,

$$(\mathcal{I} \otimes \mathcal{E})(I \otimes X) = \mathcal{I}(I) \otimes \mathcal{E}(X) = I \otimes \mathcal{E}(X) \geqslant 0,$$

and thus $\mathcal{E}(X) \geqslant 0$. This means that $\mathcal{E}$ is positive. Therefore, complete positivity is at least as strong a condition as positivity.

It may be counterintuitive, but there are maps $\mathcal{E}$ that are positive but not completely positive. Here's a great example. Fix some orthonormal basis for $\mathcal{H}$ so that we can identify operators on $\mathcal{H}$ with matrices. Now consider the transpose operator T that takes any square matrix to its transpose (not the adjoint, just the transpose), *i.e.*, $T(A) = A^T$ for any matrix A. With respect to the chosen basis, we can think of T as a map from operators in $\mathcal{L}(\mathcal{H})$ to operators in $\mathcal{L}(\mathcal{H})$, and it is clearly a linear map, and so $T \in \mathcal{T}(\mathcal{H})$. T is obviously trace-preserving, and it is also positive: For any square matrix A, it is easily checked that if $A \geqslant 0$ then $A^T \geqslant 0$ as well (if A is normal, then so is A^T, and both matrices have the same spectrum). T is not completely positive, however, provided $\dim(\mathcal{H}) \geqslant 2$. Suppose $\mathcal{H} = \mathcal{K}$ is the state space of a single qubit, and we fix the standard computational basis $\{|0\rangle, |1\rangle\}$ for $\mathcal{H}$. Consider the matrix

$$A = \left|\Phi^+\right\rangle\left\langle\Phi^+\right| = \frac{1}{2}\left[\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right] \geqslant 0.$$

Applying $\mathcal{I} \otimes T$ (sometimes called the *partial transpose*) to A means taking the transpose of each 2×2 block (the T part), but not rearranging the blocks at all (the $\mathcal{I}$ part). Thus,

$$(\mathcal{I} \otimes T)(A) = \frac{1}{2}\left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] = \frac{1}{2}\text{SWAP},$$

where we recall that the two-qubit SWAP operator swaps the qubits, *i.e.*, $\text{SWAP}|a\rangle|b\rangle = |b\rangle|a\rangle$ for any $a, b \in \{0, 1\}$. The eigenvalues of SWAP are $1, 1, 1, -1$ (see Exercise 12.2), and so SWAP is not a positive operator. This shows that $\mathcal{I} \otimes T$ is not a positive map, and so T is not completely positive.

Theorem 24.8 *Let $\mathcal{H}$ and $\mathcal{J}$ be Hilbert spaces, and let $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ be superoperator. $\mathcal{E}$ is a quantum channel (Definition 24.6) if and only if $\mathcal{E}$ is trace-preserving and completely positive.*

Proof. First the forward direction. Let $\mathcal{E}$ be a quantum channel given in the operator-sum representation by Kraus operators $K_1, \ldots, K_N$ (linear maps from $\mathcal{H}$ to $\mathcal{J}$) such that $\sum_{j=1}^{N} K_j^* K_j = I_{\mathcal{H}}$, where $I_{\mathcal{H}}$ is the identity operator on $\mathcal{H}$, and $\mathcal{E}(X) = \sum_j K_j X K_j^*$ for any $X \in \mathcal{L}(\mathcal{H})$. (Recall that each K_j^* is a linear map from $\mathcal{J}$ to $\mathcal{H}$. See page 19.) We have

$$\operatorname{tr}(\mathcal{E}(X)) = \sum_{j=1}^{N} \operatorname{tr}(K_j X K_j^*) = \operatorname{tr}\left[\left(\sum_j K_j^* K_j\right) X\right] = \operatorname{tr} X$$

for any $X \in \mathcal{L}(\mathcal{H})$, so $\mathcal{E}$ is trace-preserving.

We show that $\mathcal{E}$ is completely positive in two easy steps: (1) we show that any quantum channel is a positive map, and (2) we show that if $\mathcal{K}$ is any Hilbert space and $\mathcal{I}$ is the identity on $\mathcal{L}(\mathcal{K})$, then $\mathcal{I} \otimes \mathcal{E}$ is also a quantum channel, hence $\mathcal{I} \otimes \mathcal{E}$ is positive, hence $\mathcal{E}$ is completely positive. Step 1 was done in the case where $\mathcal{H} = \mathcal{J}$ in Exercise 24.2. The general case is similar: If $X \in \mathcal{L}(\mathcal{H})$ is any positive operator and $v \in \mathcal{J}$ is any vector, then letting $Y := \mathcal{E}(X)$, we want to show that $v^* Y v \geqslant 0$, which shows that Y is a positive operator in $\mathcal{L}(\mathcal{J})$, hence $\mathcal{E}$ is a positive map. For $1 \leqslant j \leqslant N$, define $u_j := K_j^* v \in \mathcal{H}$. We now have

$$v^* Y v = \sum_{j=1}^{N} v^* K_j X K_j^* v = \sum_j (K_j^* v)^* X K_j^* v = \sum_j u_j^* X u_j \geqslant 0$$

as desired, since $X \geqslant 0$. Because $\mathcal{E}$ is an arbitrary quantum channel, this shows that every quantum channel is a positive map.

For Step 2, let $\mathcal{K}$ be any Hilbert space, let $I_{\mathcal{K}} \in \mathcal{L}(\mathcal{K})$ be the identity operator on $\mathcal{K}$, and let $\mathcal{I} \in \mathcal{T}(\mathcal{K})$ be the identity map on $\mathcal{L}(\mathcal{K})$. We want to show that $\mathcal{I} \otimes \mathcal{E} \in \mathcal{T}(\mathcal{K} \otimes \mathcal{H}, \mathcal{K} \otimes \mathcal{J})$ is a quantum channel, so we must come up with Kraus operators for $\mathcal{I} \otimes \mathcal{E}$: For $1 \leqslant j \leqslant N$, define $L_j = I_{\mathcal{K}} \otimes K_j$. Each L_j is a linear map from $\mathcal{K} \otimes \mathcal{H}$ to $\mathcal{K} \otimes \mathcal{J}$, *i.e.*, $L_j \in \mathcal{L}(\mathcal{K} \otimes \mathcal{H}, \mathcal{K} \otimes \mathcal{J})$, and for completeness, we have

$$\sum_{j=1}^{N} L_j^* L_j = \sum_j (I_{\mathcal{K}} \otimes K_j)^* (I_{\mathcal{K}} \otimes K_j) = I_{\mathcal{K}} \otimes \left(\sum_j K_j^* K_j\right) = I_{\mathcal{K}} \otimes I_{\mathcal{H}},$$

which is the identity map on $\mathcal{K} \otimes \mathcal{H}$. Finally, if $A \in \mathcal{L}(\mathcal{K})$ and $B \in \mathcal{L}(\mathcal{H})$ are arbitrary operators and we set $C := A \otimes B$, we have

$$(\mathcal{I} \otimes \mathcal{E})(C) = (\mathcal{I} \otimes \mathcal{E})(A \otimes B) = A \otimes \mathcal{E}(B) = \sum_{j=1}^{N} (I_{\mathcal{K}} \otimes K_j)(A \otimes B)(I_{\mathcal{K}} \otimes K_j^*) = \sum_j L_j C L_j^*. \qquad (97)$$

Both sides of Equation (97) are linear in C, and so (97) extends to arbitrary $C \in \mathcal{L}(\mathcal{K} \otimes \mathcal{H})$. This shows that $\mathcal{I} \otimes \mathcal{E}$ is a quantum channel and hence positive, making $\mathcal{E}$ completely positive.

Now the reverse direction. Suppose $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ is trace-preserving and completely positive. We need to come up with Kraus operators for $\mathcal{E}$. Let $n := \dim(\mathcal{H})$, and let $\mathcal{I}$ be the identity map on

$\mathcal{L}(\mathcal{H})$. Fix an orthonormal basis $\{e_1, \ldots, e_n\}$ for $\mathcal{H}$. Taking the product of this basis with itself, we get a basis $\{e_{i,j} : 1 \leqslant i, j \leqslant n\}$ for $\mathcal{H} \otimes \mathcal{H}$, where for convenience we define $e_{i,j} := e_i \otimes e_j$. Define the vector

$$v := \sum_{i=1}^{n} e_{i,i} = \sum_i e_i \otimes e_i \in \mathcal{H} \otimes \mathcal{H}.$$

The operator $vv^* \in \mathcal{L}(\mathcal{H} \otimes \mathcal{H})$ is clearly positive, and so by assumption, the operator

$$J := J(\mathcal{E}) := (\mathcal{I} \otimes \mathcal{E})(vv^*) \in \mathcal{L}(\mathcal{H} \otimes \mathcal{J})$$

is also positive. (We are letting $\mathcal{K} = \mathcal{H}$.) We have

$$vv^* = \sum_{i,j=1}^{n} (e_i \otimes e_i)(e_j^* \otimes e_j^*) = \sum_{i,j=1}^{n} e_i e_j^* \otimes e_i e_j^* = \sum_{i,j} E_{ij} \otimes E_{ij} \ ,$$

where we let $E_{ij} := e_i e_j^*$ as usual. Thus

$$J = (\mathcal{I} \otimes \mathcal{E}) \left(\sum_{i,j} E_{ij} \otimes E_{ij} \right) = \sum_{i,j=1}^{n} E_{ij} \otimes \mathcal{E}(E_{ij}) \ .^{24}$$

Because $J \geqslant 0$, we can choose some eigenbasis $\{g_1, \ldots, g_N\}$ for J, where $N := n^2 = \dim(\mathcal{H} \otimes \mathcal{H})$. This allows us to write

$$J = \sum_{k=1}^{N} \lambda_k g_k g_k^*,$$

where $\lambda_1, \ldots, \lambda_N \geqslant 0$ are the eigenvalues of J. For $1 \leqslant k \leqslant N$, we can now define the Kraus operator $K_k \in \mathcal{L}(\mathcal{H})$ by its matrix with respect to the $\{e_{i,j}\}$ basis: for all $1 \leqslant i, j \leqslant n$, define

$$[K_k]_{ij} := \sqrt{\lambda_k} \langle e_{j,i}, g_k \rangle = \sqrt{\lambda_k} e_{j,i}^* g_k \ .$$

We need to check that $\sum_{k=1}^{N} K_k^* K_k = I$ (completeness) and that $\mathcal{E}(X) = \sum_{k=1}^{N} K_k X K_k^*$ for all $X \in \mathcal{L}(\mathcal{H})$.

For completeness, fix some $a, b \in \{1, \ldots, n\}$, and using the fact that $\mathcal{E}$ is trace-preserving, compute

$$\begin{aligned}
\left[\sum_{k=1}^{N} K_k^* K_k \right]_{ab} &= \sum_k \sum_{c=1}^{n} [K_k^*]_{ac} [K_k]_{cb} = \sum_k \sum_c [K_k]_{ca}^* [K_k]_{cb} = \sum_k \lambda_k \sum_c \langle e_{a,c}, g_k \rangle^* \langle e_{b,c}, g_k \rangle \\
&= \sum_k \lambda_k \sum_c \langle e_{b,c}, g_k \rangle \langle g_k, e_{a,c} \rangle = \sum_k \lambda_k \sum_c e_{b,c}^* g_k g_k^* e_{a,c} \\
&= \sum_c e_{b,c}^* \left(\sum_k \lambda_k g_k g_k^* \right) e_{a,c} = \sum_c e_{b,c}^* J e_{a,c} = \sum_c \sum_{i,j=1}^{n} e_{b,c}^* (E_{ij} \otimes \mathcal{E}(E_{ij})) e_{a,c} \\
&= \sum_{c,i,j} e_b^* E_{ij} e_a \otimes e_c^* (\mathcal{E}(E_{ij})) e_c = \sum_c e_c^* (\mathcal{E}(E_{ba})) e_c = \operatorname{tr}[\mathcal{E}(E_{ba})] = \operatorname{tr} E_{ba} = \delta_{ab}.
\end{aligned}$$

[24] It is not needed for the proof, but it is interesting to note that the matrix J contains complete information about $\mathcal{E}$. It includes all the n^2 matrices $\mathcal{E}(E_{ij})$ laid out as $n \times n$ blocks in an $n^2 \times n^2$ matrix. Since the E_{ij} form a basis for $\mathcal{L}(\mathcal{H})$, $\mathcal{E}$ is completely determined by the matrices $\mathcal{E}(E_{ij})$. $J = J(\mathcal{E})$ is called the *Choi representation* of $\mathcal{E}$. The condition $J \geqslant 0$ is actually equivalent to $\mathcal{E}$ being completely positive, for any superoperator $\mathcal{E}$.

From this we get that $\sum_{k=1}^{N} K_k^* K_k$ is the identity matrix. This shows completeness.

Now let $X \in \mathcal{L}(\mathcal{H})$ be arbitrary. Again, we compare matrix elements with respect to the $\{e_i\}$ basis. For any $1 \leqslant a, b \leqslant n$, we have

$$\begin{aligned}
\left[\sum_{k=1}^{N} K_k X K_k^*\right]_{ab} &= \sum_k \sum_{c,d=1}^{n} [K_k]_{ac}[X]_{cd}[K_k^*]_{db} = \sum_k \sum_{c,d=1}^{n} [X]_{cd}[K_k]_{ac}[K_k]_{bd}^* \\
&= \sum_k \lambda_k \sum_{c,d} [X]_{cd}\langle e_{c,a}, g_k\rangle\langle g_k, e_{d,b}\rangle = \sum_k \lambda_k \sum_{c,d} [X]_{cd} e_{c,a}^* g_k g_k^* e_{d,b} \\
&= \sum_{c,d} [X]_{cd} e_{c,a}^* J e_{d,b} \qquad \text{(just as before)} \\
&= \sum_{c,d} [X]_{cd} \sum_{i,j=1}^{n} e_{c,a}^*(E_{ij} \otimes \mathcal{E}(E_{ij}))e_{d,b} \\
&= \sum_{c,d} [X]_{cd} \sum_{i,j} e_c^* E_{ij} e_d \otimes e_a^*(\mathcal{E}(E_{ij}))e_b = \sum_{c,d} [X]_{cd} e_a^*(\mathcal{E}(E_{cd}))e_b \\
&= e_a^* \left(\sum_{c,d} [X]_{cd}\mathcal{E}(E_{cd})\right) e_b = e_a^* \left(\mathcal{E}\left(\sum_{c,d} [X]_{cd} E_{cd}\right)\right) e_b \\
&= e_a^*(\mathcal{E}(X))e_b = [\mathcal{E}(X)]_{ab}.
\end{aligned}$$

Thus $\mathcal{E}(X) = \sum_{k=1}^{N} K_k X K_k^*$ as we wanted. □

Exercise 24.9 (Optional) Show that the composition of two quantum channels is a quantum channel. That is, let $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ and $\mathcal{F} \in \mathcal{T}(\mathcal{J}, \mathcal{K})$ be quantum channels. Show that $\mathcal{F} \circ \mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{K})$ is a quantum channel, where $\mathcal{F} \circ \mathcal{E}$ is defined as $(\mathcal{F} \circ \mathcal{E})(X) := \mathcal{F}(\mathcal{E}(X))$ for all $X \in \mathcal{L}(\mathcal{H})$.

Exercise 24.10 (Challenging, Optional) Show that the tensor product of two quantum channels is a quantum channel. That is, let $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ and $\mathcal{F} \in \mathcal{T}(\mathcal{K}, \mathcal{M})$ be quantum channels. Show that $\mathcal{E} \otimes \mathcal{F} \in \mathcal{T}(\mathcal{H} \otimes \mathcal{K}, \mathcal{J} \otimes \mathcal{M})$ is a quantum channel.

Definition 24.6 defines what are sometimes called *complete quantum channels*, and a *general quantum channel* (not necessarily complete) is defined the same way, except that we replace the completeness condition $\sum_{j=1}^{N} K_j^* K_j = I_{\mathcal{H}}$ with the looser condition $\sum_{j=1}^{N} K_j^* K_j \leqslant I_{\mathcal{H}}$. Incomplete quantum channels are used to describe physical processes that may not happen with certainty, *e.g.*, a general measurement that results in some outcome m.

Exercise 24.11 (Challenging, Optional) Show that a superoperator $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ is a general quantum channel, as described above, if and only if (1) $\mathcal{E}$ is completely positive, and (2) for every state (positive operator with unit trace) $\rho \in \mathcal{L}(\mathcal{H})$, we have $0 \leqslant \mathrm{tr}(\mathcal{E}(\rho)) \leqslant 1$. The quantity $\mathrm{tr}(\mathcal{E}(\rho))$ is interpreted as the probability that $\mathcal{E}$ actually occurs. [Hint: Set $L := I_{\mathcal{H}} - \sum_{j=1}^{N} K_j^* K_j$, where the K_j are the Kraus operators corresponding to $\mathcal{E}$ as above. Since $L \geqslant 0$, you can define $K_{N+1} := \sqrt{L}$, and then define $\mathcal{E}'(X) := \sum_{j=1}^{N+1} K_j X K_j^*$ for any $X \in \mathcal{L}(\mathcal{H})$. Notice that $\mathcal{E}'$ is a *complete* quantum channel and that $\mathcal{E}(X) = \mathcal{E}'(X) - \sqrt{L}X\sqrt{L}$. Also note that $L \leqslant I_{\mathcal{H}}$. Apply Theorem 24.8 to $\mathcal{E}'$, and use it to prove facts about $\mathcal{E}$.]

Exercise 24.12 (Challenging, Optional) Show that the partial trace map is always a (complete) quantum channel. [Hint: Let $\mathrm{tr}_{\mathcal{H}} : \mathcal{L}(\mathcal{H} \otimes \mathcal{J}) \to \mathcal{L}(\mathcal{J})$ be a partial trace map. Note that by linearity, $\mathrm{tr}_{\mathcal{H}} \in \mathcal{T}(\mathcal{H} \otimes \mathcal{J}, \mathcal{J})$ is a superoperator. Fix orthonormal bases $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_m\}$ for $\mathcal{H}$ and $\mathcal{J}$, respectively, and for each j with $1 \leqslant j \leqslant n$, define the Kraus operator $K_j \in \mathcal{L}(\mathcal{H} \otimes \mathcal{J}, \mathcal{J})$ by

$$K_j := e_j^* \otimes I_{\mathcal{J}} = e_j^* \otimes \sum_{k=1}^{m} f_k f_k^* = \sum_{k=1}^{m} f_k (e_j^* \otimes f_k^*),$$

where $I_{\mathcal{J}}$ is the identity map on $\mathcal{J}$. In other words, for every vector in $\mathcal{H} \otimes \mathcal{J}$ of the form $u \otimes v$, we have $K_j(u \otimes v) = \langle e_j, u \rangle v$. Show that the K_j satisfy the completeness condition and then characterize $\mathrm{tr}_{\mathcal{H}}$ accordingly.]

25 Week 12: Distance and fidelity

Distance Measures. First, some basic definitions from probability theory.

Recall that we have been talking about a *probability distribution* as a finite list of values $p = (p_1, p_2, \ldots, p_k)$ such that $p_j \geqslant 0$ for all $1 \leqslant j \leqslant k$ and $\sum_{j=1}^{k} p_j = 1$. Here, the set $\{1, \ldots, k\}$ is called the *sample space*. More generally, any finite or countable set Ω can be used as a sample space, in which case, a *probability distribution on* Ω is a map $p : \Omega \to \mathbb{R}$ such that $p(a) \geqslant 0$ for all $a \in \Omega$, and $\sum_{a \in \Omega} p(a) = 1$. Subsets of Ω are called *events*, and elements of Ω, which we identify with singleton subsets of Ω, are called *elementary events*. If $S \subseteq \Omega$ is some event, then the *probability of* S (with respect to the probability distribution p, above) is defined as

$$\Pr_p[S] := \sum_{a \in S} p(a).$$

We might drop the subscript p if it is clear what probability distribution we are using.

If p and q are two probability distributions over the same sample space, we are interested in measures of the similarity or difference between p and q. We'll discuss two here: the *trace distance* and the *fidelity*.

Definition 25.1 Let p and q be two probability distributions on the same sample space Ω. The *trace distance* (also called the L_1 *distance* or the *Kolmogorov distance*) between p and q is defined as

$$D(p, q) := \frac{1}{2} \sum_{a \in \Omega} |p(a) - q(a)|.$$

It is easy to check that D satisfies the axioms for a metric on the set of all probability distributions on Ω. These are:

1. $D(p, q) \geqslant 0$,
2. $D(p, q) = 0$ iff $p = q$,
3. $D(p, q) = D(q, p)$, and

4. $D(p,r) \leqslant D(p,q) + D(q,r)$,

for any probability distributions p, q, r on Ω. Here's another way of characterizing the trace distance: for any probability distributions p and q on Ω,

$$D(p,q) = \max_{S \subseteq \Omega} |\Pr_p[S] - \Pr_q[S]| = \max_{S \subseteq \Omega} (\Pr_p[S] - \Pr_q[S]). \tag{98}$$

Exercise 25.2 Prove Equation (98).

The trace distance guages the difference between two distributions p and q. The fidelity, on the other hand, is a measure of their similarity; it is *maximized* when $p = q$.

Definition 25.3 Let p and q be two probability distributions on the same sample space Ω. The *fidelity* of p and q is defined as

$$F(p,q) := \sum_{a \in \Omega} \sqrt{p(a)q(a)}.$$

$F(p,q)$ can be seen as the dot product of two real unit vectors—the vector whose a'th entry is $\sqrt{p(a)}$ and the vector whose a'th entry is $\sqrt{q(a)}$. Since these two vectors clearly have unit norm, the fidelity is then the cosine of the angle between them. Thus we immediately get $0 \leqslant F(p,q) \leqslant 1$, with $F(p,q) = 1$ iff $p = q$.

Trace Distance and Fidelity of Operators. We'd like to extend these definitions to quantum states, *i.e.*, operators. A reasonable sanity check on the way we should define such an extension would be to say that if ρ and σ are mixtures of the same set of pairwise orthogonal pure states with (eigenvalue) probability distributions r and s, respectively, then $D(\rho,\sigma)$ should be equal to $D(r,s)$, and $F(\rho,\sigma)$ should be equal to $F(r,s)$. Let's see this in more detail. Suppose $\rho = \sum_{j=1}^{k} r_j \rho_j$ and $\sigma = \sum_{j=1}^{k} s_j \rho_j$, where the pure states ρ_j project onto mutually orthogonal subspaces (equivalently, $\rho_i \rho_j = \delta_{ij} \rho_i$ for any i and j). Now consider the operator $|\rho - \sigma|$. We have

$$\begin{aligned}
|\rho - \sigma| &= \sqrt{(\rho-\sigma)^*(\rho-\sigma)} \\
&= \sqrt{(\rho-\sigma)^2} \\
&= \left[\left(\sum_{j=1}^{k} (r_j - s_j)\rho_j \right)^2 \right]^{1/2} \\
&= \left(\sum_{j=1}^{k} (r_j - s_j)^2 \rho_j \right)^{1/2},
\end{aligned}$$

because the cross-terms ($\rho_i \rho_j$ for $i \neq j$) all vanish when we expand the expression inside the square brackets. Since the ρ_j project onto mutually orthogonal subspaces, we can choose an orthonormal basis in which all the ρ_j are diagonal matrices simultaneously. Permuting the basis vectors if need

be, we can assume that each ρ_j (which is a one-dimensional projector) is given by the matrix E_{jj}. Thus $\sum_{j=1}^{k}(r_j - s_j)^2\rho_j = \sum_j (r_j - s_j)^2 E_{jj} = \mathrm{diag}[(r_1 - s_1)^2, (r_2 - s_2)^2, \ldots, (r_k - s_k)^2, 0, \ldots, 0]$. To take the square root of this matrix, we just take the square root of each diagonal entry, which gives the matrix $\mathrm{diag}[|r_1 - s_1|, |r_2 - s_2|, \ldots, |r_k - s_k|, 0, \ldots, 0]$, and so this is $|\rho - \sigma|$ in matrix form. Taking one half of the trace of this gives

$$\frac{1}{2}\,\mathrm{tr}\,|\rho - \sigma| = \frac{1}{2}\sum_{j=1}^{k}|r_j - s_j| = D(r, s).$$

This suggests that we can now define the trace distance $D(\rho, \sigma)$ for *arbitrary* operators A and B as

$$D(A, B) := \frac{1}{2}\,\mathrm{tr}\,|A - B| = \frac{1}{2}\|A - B\|_1.$$

We can do something similar to define the fidelity of two arbitrary positive operators. I won't do the details here, but a reasonable definition is

$$F(A, B) := \mathrm{tr}\,\sqrt{A^{1/2}BA^{1/2}} = \left\| \sqrt{B}\,\sqrt{A} \right\|_1 \tag{99}$$

for arbitrary operators $A, B \geqslant 0$. It can be shown that $F(A, B) = F(B, A)$, and if ρ and σ are states, then $0 \leqslant F(\rho, \sigma) \leqslant 1$ with $F(\rho, \sigma) = 1$ iff $\rho = \sigma$.

We do the same sanity check for F as we did for D, above. If ρ and σ are commuting states as before, *i.e.*, $\rho = \mathrm{diag}(r_1, \ldots, r_k, 0, \ldots, 0)$ and $\sigma = \mathrm{diag}(s_1, \ldots, s_k, 0, \ldots, 0)$ with respect to the same orthonormal basis, then we have

$$F(\rho, \sigma) = \mathrm{tr}\,\sqrt{\mathrm{diag}(r_1 s_1, \ldots, r_k s_k, 0, \ldots, 0)} = \sum_{j=1}^{k}\sqrt{r_j s_j} = F(r, s).$$

Exercise 25.4 Show that $\|AB\|_1 = \|BA\|_1$ for any Hermitean operators A and B. Thus the fidelity function F of (99) is symmetric. [Hint: Use Property 10 of the norm, which says that $\|C\|_1 = \|C^*\|_1$ for any operator C.]

Properties of the Trace Distance. The trace distance of operators has an alternate characterization analogous to Equation (98). If A and B are operators, we say that $A \leqslant B$ if $B - A \geqslant 0$. We'll show that for any states ρ and σ,

$$D(\rho, \sigma) = \max_{\text{projectors } P} \mathrm{tr}(P(\rho - \sigma)) = \max_{P \geqslant 0\ \&\ \|P\| = 1} \mathrm{tr}(P(\rho - \sigma)) = \max_{0 \leqslant P \leqslant I} \mathrm{tr}(P(\rho - \sigma)), \tag{100}$$

where the three maxima are taken over all projectors P, all positive operators P of unit operator norm (L_∞ norm), and all operators P such that $0 \leqslant P \leqslant I$, respectively. Equation (100) has many uses. We won't bother to do it here, but it is straightforward to check—as a consequence of Equation (100)—that $D(\rho, \sigma)$ is the maximum probability difference of any outcome of a POVM applied to ρ and to σ. The function D is also a metric on the set of all quantum states of a given system, that is, it can be shown to satisfy the axioms for a metric on page 114, and (100) helps with showing the triangle inequality for D.

Actually, we'll show a result slightly more general than Equation (100):

Proposition 25.5 *Suppose that A is a traceless Hermitean operator, i.e., $\operatorname{tr} A = 0$ and $A = A^*$. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ be the eigenvalues of A (A acts on an n-dimensional space). The following quantities are all equal:*

1. $(1/2)\|A\|_1$,
2. $(1/2)\operatorname{tr}|A|$,
3. $\sum_{i:\lambda_i>0} \lambda_i$,
4. $\max_{\text{projectors } P} \operatorname{tr}(PA)$,
5. $\max_{0\leqslant P \;\&\; \|P\|=1} \operatorname{tr}(PA)$,
6. $\max_{0\leqslant P\leqslant I} \operatorname{tr}(PA)$.

Proof. We'll do these in increasing order of difficulty.

(1) = (2) follows directly from the definition of $\|\cdot\|_1$ (Equation (83)). For (2) = (3), let $p := \sum_{i:\lambda_i>0} \lambda_i$ and let $q := \sum_{i:\lambda_i<0} \lambda_i$. Note that $p + q = \sum_{i=1}^n \lambda_i = \operatorname{tr} A = 0$, and so $q = -p$. Also note that the eigenvalues of $|A|$ are $|\lambda_1|, \ldots, |\lambda_n|$. This is easiest to see by taking an eigenbasis for A (A is normal because it is Hermitean) and looking at the matrices for A and $|A|$. So we have

$$\operatorname{tr}|A| = \sum_{i=1}^n |\lambda_i| = \sum_{i:\lambda_i>0} \lambda_i - \sum_{i:\lambda_i<0} \lambda_i = p - q = 2p.$$

The inequalities (3) $\leqslant$ (4) $\leqslant$ (5) $\leqslant$ (6) are pretty straightforward and we leave these as exercises.

It remains to show that (6) $\leqslant$ (3). Consider the expression $\max_{0\leqslant P\leqslant I} \operatorname{tr}(PA)$ of (6). The key insight is to show first that the maximum is achieved by some P that *commutes* with A (*i.e.*, $PA = AP$). Once that fact is established, the rest is easy: we can pick a common eigenbasis for P and A and look at diagonal matrices.

Suppose that $0 \leqslant P \leqslant I$ and that P does not commute with A. We will find an operator P' such that $0 \leqslant P' \leqslant I$ and $\operatorname{tr}(P'A) > \operatorname{tr}(PA)$, and so the maximum is not achieved by P.[25] Set $C := i(AP - PA)$. Note that C is Hermitean, because both P and A are, and $C \neq 0$ by assumption. (The quantity $AP - PA$, for any operators A and P, is called the *commutator* or the *Lie bracket* (pronounced, "LEE") of A and P, and is denoted by $[A, P]$.) For any $\varepsilon > 0$, define

$$U_\varepsilon := e^{-i\varepsilon C} = I - i\varepsilon C + O(\varepsilon^2).$$

Then U_ε is unitary by Item 4 of Exercise 9.3. The "$O(\varepsilon^2)$" here denotes an operator (depending on ε) whose norm (it doesn't matter which norm) is bounded by some positive constant times ε^2. We now define

$$P' := U_\varepsilon P U_\varepsilon^*$$

[25] We are tacitly assuming that the maximum is achieved by *some* P such that $0 \leqslant P \leqslant I$. This is in fact true, and it follows from concepts in topology that we won't go into here, namely, continuity and compactness.

for some $\varepsilon > 0$ that we will choose later. It is easy to check that $0 \leqslant P' \leqslant I$. Now we have

$$\begin{aligned}
\operatorname{tr}(P'A) &= \operatorname{tr}(U_\varepsilon P U_\varepsilon^* A) \\
&= \operatorname{tr}\left[(I - i\varepsilon C + O(\varepsilon^2))P(I + i\varepsilon C + O(\varepsilon^2))A\right] \\
&= \operatorname{tr}\left[PA + i\varepsilon PCA - i\varepsilon CPA + O(\varepsilon^2)\right] \\
&= \operatorname{tr}(PA) + i\varepsilon\left[\operatorname{tr}(PCA) - \operatorname{tr}(CPA)\right] + O(\varepsilon^2) \\
&= \operatorname{tr}(PA) + i\varepsilon\left[\operatorname{tr}(CAP) - \operatorname{tr}(CPA)\right] + O(\varepsilon^2) \\
&= \operatorname{tr}(PA) + i\varepsilon \operatorname{tr}\left[C(AP - PA)\right] + O(\varepsilon^2) \\
&= \operatorname{tr}(PA) + \varepsilon \operatorname{tr}(C^2) + O(\varepsilon^2).
\end{aligned}$$

Now $C^2 = C^*C \geqslant 0$, and since $C \neq 0$, we must then have $\operatorname{tr}(C^2) > 0$, either by Exercise 9.28 or by observing that $\operatorname{tr}(C^*C) = \langle C, C\rangle > 0$ (Hilbert-Schmidt inner product). Now we can choose ε small enough so that $\varepsilon \operatorname{tr}(C^2)$ strictly dominates the $O(\varepsilon^2)$ error term, yielding $\operatorname{tr}(P'A) > \operatorname{tr}(PA)$. This shows that the maximum value of $\operatorname{tr}(PA)$ is achieved only when P commutes with A, *i.e.*,

$$\max_{0 \leqslant P \leqslant I} \operatorname{tr}(PA) = \max_{0 \leqslant P \leqslant I \;\&\; PA = AP} \operatorname{tr}(PA).$$

Finally suppose that $0 \leqslant P \leqslant I$ and that P commutes with A. Pick a common eigenbasis for P and A so that, with respect to this basis, $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $P = \operatorname{diag}(\mu_1, \ldots, \mu_n)$. Since $0 \leqslant P \leqslant I$, we must have $0 \leqslant \mu_1, \ldots, \mu_n \leqslant 1$, but otherwise, we are free to choose the μ_j arbitrarily (see the hint to Exercise 25.7, below). We now have

$$\operatorname{tr}(PA) = \sum_{j=1}^{n} \mu_j \lambda_j,$$

and this sum is the largest possible when we define

$$\mu_j := \begin{cases} 1 & \text{if } \lambda_j > 0, \\ 0 & \text{otherwise.} \end{cases}$$

For this choice of the μ_j, we get $\operatorname{tr}(PA) = \sum_{j:\lambda_j > 0} \lambda_j$, and we've shown that this is the largest possible value for $\operatorname{tr}(PA)$ with $0 \leqslant P \leqslant I$. (Note that the optimal P is a projector. That's a direct way to see that (4) $\leqslant$ (3).) □

Exercise 25.6 Prove (3) $\leqslant$ (4) in Proposition 25.5, above. [Hint: Find a projector P such that $\operatorname{tr}(PA) = \sum_{i:\lambda_i > 0} \lambda_i$. This shows that $\sum_{i:\lambda_i > 0} \lambda_i \leqslant \max_{\text{projectors } P} \operatorname{tr}(PA)$. To find P, consider the subspace spanned by all the eigenvectors of A with positive eigenvalues.]

Exercise 25.7 Prove (4) $\leqslant$ (5) $\leqslant$ (6) in Proposition 25.5, above. [Hint: The following easy facts are useful for any operator P:

- $0 \leqslant P \leqslant I$ if and only if P is normal and all its eigenvalues are in the closed interval $[0, 1] \subseteq \mathbb{R}$ (consider an eigenbasis for P).
- Recall that $\|P\|$ is the maximum eigenvalue of $|P|$.

- Recall (Exercise 9.35) that $0 \leqslant P$ iff $P = |P|$, and thus if $0 \leqslant P$ then $\|P\|$ is the largest eigenvalue of P itself.

For (4) $\leqslant$ (5), note that every nonzero projector P satisfies $0 \leqslant P$ and $\|P\| = 1$. You need to treat the case where $P = 0$ separately. (5) $\leqslant$ (6) is straightforward.]

Exercise 25.8 Use Proposition 25.5 to prove Equation (100).

Exercise 25.9 Let $\rho_1 = (I + \vec{r} \cdot \vec{\sigma})/2 = (I + r_x X + r_y Y + r_z Z)/2$ and $\tau = (I + \vec{t} \cdot \vec{\sigma})/2 = (I + t_x X + t_y Y + t_z Z)/2$ be single-qubit states, where X, Y, Z are the usual Pauli spin matrices and $\vec{r} = (r_x, r_y, r_z) \in \mathbb{R}^3$ and $\vec{t} = (t_x, t_y, t_z) \in \mathbb{R}^3$ are either on or inside the Bloch sphere. Show that

$$D(\rho, \tau) = \frac{\|\vec{r} - \vec{t}\|}{2} = \frac{1}{2}\sqrt{(r_x - t_x)^2 + (r_y - t_y)^2 + (r_z - t_z)^2}\,,$$

i.e., half the Euclidean distance between $\vec{r}$ and $\vec{t}$.

In Proposition 25.13, below, I'll mention one more interesting property of the trace distance: it can never increase via a quantum channel. This says that all (complete) quantum channels are *contractive* with respect to the metric D. So if no classical information is coming out of an open quantum system, its dynamics tends to cause states to become less distinguishable, not more. This is not necessarily the case with incomplete quantum channels, where some classical information is obtained.

Lemma 25.10 (*Jordan-Hahn decomposition*) *For any Hermitean operator A, there exist unique positive operators Q and S such that $QS = SQ = 0$ and $A = Q - S$.*

Proof. To prove existence, define

$$\begin{aligned} Q &:= \frac{|A| + A}{2}\,, \\ S &:= \frac{|A| - A}{2}\,. \end{aligned}$$

Evidently, $A = Q - S$; moreover,

$$QS = \frac{1}{4}(|A|^2 - |A|A + A|A| - A^2) = |A|^2 - A^2\,,$$

because A commutes with $|A|$. Since A is Hermitean, we have $A^2 = A^*A = |A|^2$, and hence, $QS = 0$. A similar argument gives $SQ = 0$. It remains to show that Q and S are both positive. Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of A. The λ_i are all real, since A is Hermitean. By Corollary 9.17, we have a unique decomposition

$$A = \lambda_1 P_1 + \cdots + \lambda_k P_k\,,$$

where the P_j form a complete set of orthogonal projectors. Then

$$|A| = |\lambda_1| P_1 + \cdots + |\lambda_k| P_k\,,$$

which immediately implies

$$\begin{aligned} Q &= \frac{1}{2}\left((|\lambda_1|+\lambda_1)P_1+\cdots+(|\lambda_k|+\lambda_k)P_k\right) = \sum_{j:\lambda_j>0}\lambda_j P_j\ , \\ S &= \frac{1}{2}\left((|\lambda_1|-\lambda_1)P_1+\cdots+(|\lambda_k|-\lambda_k)P_k\right) = \sum_{j:\lambda_j<0}(-\lambda_j)P_j\ . \end{aligned}$$

All the coefficients above are nonnegative real numbers, and since all the P_i are positive, Q and S must both be positive.

To prove uniqueness, suppose some positive operators Q and S satisfy the conditions of the lemma. Since $QS = 0 = SQ$, Q and S commute with each other, and thus Q and S both commute with $Q - S = A$. Since A, Q, and S are all normal operators, they share a common eigenbasis $\mathcal{B}$ by Theorem 9.41. With respect to $\mathcal{B}$, these three operators are represented by diagonal matrices:

$$\begin{aligned} A &= \operatorname{diag}(a_1,\ldots,a_n)\ , \\ Q &= \operatorname{diag}(q_1,\ldots,q_n)\ , \\ S &= \operatorname{diag}(s_1,\ldots,s_n)\ , \end{aligned}$$

for some $a_1,\ldots,a_n \in \mathbb{R}$ (because A is Hermitean) and $q_1,\ldots,q_n,s_1,\ldots,s_n \geqslant 0$ (because $Q,S \geqslant 0$). Let $1 \leqslant i \leqslant n$ be arbitrary. The conditions $A = Q - S$ and $QS = 0$ imply $a_i = q_i - s_i$ and $q_i s_i = 0$, respectively. The latter equation implies at least one of q_i and s_i is zero. Since $q_i, s_i \geqslant 0$, we must have $(q_i, s_i) = (a_i, 0)$ if $a_i \geqslant 0$ and $(q_i, s_i) = (0, -a_i)$ if $a_i \leqslant 0$. (If $a_i = 0$, then $q_i = s_i = 0$.). Thus q_i and s_i are uniquely determined, given a_i. Since i was arbitrary, Q and S are uniquely determined given A. □

The condition that $QS = SQ = 0$ is often referred to as Q and S having "*orthogonal support*," and it is equivalent to $\langle Q, S\rangle = 0$ (for any positive operators Q and S.).

Exercise 25.11 Show that if A, Q, and S are as in Lemma 25.10, then $|A| = |Q - S| = |Q + S| = \sqrt{Q^2 + S^2}$.

Exercise 25.12 Show that if A, Q, and S are as in Lemma 25.10, then $\operatorname{tr}|A| = \operatorname{tr} Q + \operatorname{tr} S$. [Hint: Either pick a common eigenbasis for Q and S or use the decomposition in the proof of Lemma 25.10.]

Proposition 25.13 *Let $\mathcal{E} \in \mathcal{T}(\mathcal{H},\mathcal{J})$ be a (complete, i.e., trace-preserving) quantum channel, and let ρ and σ be states in $\mathcal{L}(\mathcal{H})$. Then $D(\mathcal{E}(\rho),\mathcal{E}(\sigma)) \leqslant D(\rho,\sigma)$.*

Proof. The operator $\rho - \sigma$ satisfies Lemma 25.10, so uniquely write $\rho - \sigma = Q - S$, where $Q, S \geqslant 0$ and $QS = SQ = 0$. Now we have

$$\begin{aligned} D(\rho,\sigma) &= \frac{1}{2}\operatorname{tr}|\rho-\sigma| = \frac{1}{2}(\operatorname{tr} Q + \operatorname{tr} S) && \text{(Exercise 25.12)} \\ &= \operatorname{tr} Q && \text{(because } \operatorname{tr} Q - \operatorname{tr} S = \operatorname{tr}(Q-S) = \operatorname{tr}(\rho-\sigma) = 1-1=0\text{)} \\ &= \operatorname{tr}(\mathcal{E}(Q))\ . && (\mathcal{E} \text{ is trace-preserving}) \end{aligned}$$

Noticing that $\mathcal{E}(\rho) - \mathcal{E}(\sigma) = \mathcal{E}(\rho - \sigma)$ is a traceless Hermitean operator, we can choose a projector P that maximizes $\mathrm{tr}(P(\mathcal{E}(\rho) - \mathcal{E}(\sigma)))$. Continuing on, we get

$$\begin{aligned}
\mathrm{tr}(\mathcal{E}(Q)) &\geqslant \mathrm{tr}(P\mathcal{E}(Q)) \qquad \text{(by Corollary 9.40 and } P \geqslant 0\text{)} \\
&\geqslant \mathrm{tr}(P(\mathcal{E}(Q) - \mathcal{E}(S))) \qquad \text{(because } \mathrm{tr}(P(\mathcal{E}(S))) = \langle P, \mathcal{E}(S)\rangle \geqslant 0 \text{ by Theorem 9.31)} \\
&= \mathrm{tr}(P(\mathcal{E}(Q - S))) = \mathrm{tr}(P(\mathcal{E}(\rho - \sigma))) \\
&= \mathrm{tr}(P(\mathcal{E}(\rho) - \mathcal{E}(\sigma))) \\
&= D(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \qquad \text{(choice of } P \text{ and Equation (100))}
\end{aligned}$$

Thus $D(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \leqslant D(\rho, \sigma)$, which proves the theorem. □

A particular example of Proposition 25.13 is when $\mathcal{E}$ is a partial trace operator (see Exercise 24.12). The interpretation is that if we ignore part of a system, we lose distinguishability between its states.

Properties of the Fidelity. An important special case of Equation (99) is when $\rho = uu^*$ is a pure state (u is a unit vector). We may prepare a pure state ρ, then send it through a noisy quantum channel (quantum channel $\mathcal{E}$), producing a state σ at the other end. The fidelity $F(\rho, \sigma)$ is a good measure of how much the state was garbled in the transmission—the higher the fidelity, the less garbling. For $\rho = uu^*$ and any state σ, we have

$$F(uu^*, \sigma) = \mathrm{tr}\ \sqrt{u(u^*\sigma u)u^*} = \sqrt{u^*\sigma u}\ \mathrm{tr}\ \sqrt{uu^*} = \sqrt{u^*\sigma u}\,, \tag{101}$$

noting that $\sqrt{uu^*} = uu^*$, which has unit trace. In Dirac notation, letting $|\psi\rangle := u$, this becomes

$$F(|\psi\rangle\langle\psi|, \sigma) = \mathrm{tr}\ \sqrt{|\psi\rangle\langle\psi|\sigma|psi\rangle\langle psi|} = \sqrt{\langle\psi|\sigma|\psi\rangle}\ \mathrm{tr}\ \sqrt{|psi\rangle\langle psi|} = \sqrt{\langle\psi|\sigma|\psi\rangle}\,. \tag{102}$$

There is a fact about the fidelity analogous with Proposition 25.13 regarding the trace distance. We'll state it without proof.

Proposition 25.14 *Suppose $\mathcal{E} \in \mathcal{T}(\mathcal{H}, \mathcal{J})$ is a complete quantum channel. For any two states $\rho, \sigma \in \mathcal{L}(\mathcal{H})$,*

$$F(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \geqslant F(\rho, \sigma)\,.$$

Comparing Trace Distance and Fidelity. The trace distance and fidelity are roughly interchangeable as measures of distinctness/similarity. For pure states ρ and σ it can be shown that $D(\rho, \sigma) = \sqrt{1 - F(\rho, \sigma)^2}$. For arbitrary states ρ and σ, it can be shown that

$$1 - F(\rho, \sigma) \leqslant D(\rho, \sigma) \leqslant \sqrt{1 - F(\rho, \sigma)^2}\,,$$

or equivalently,

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

These inequalities are known as the *Fuchs-van de Graaf inequalities*. So in most situations, it doesn't really matter which measure is used. The book uses the fidelity measure almost exclusively. The inequalities above are illustrated in Figure 10.

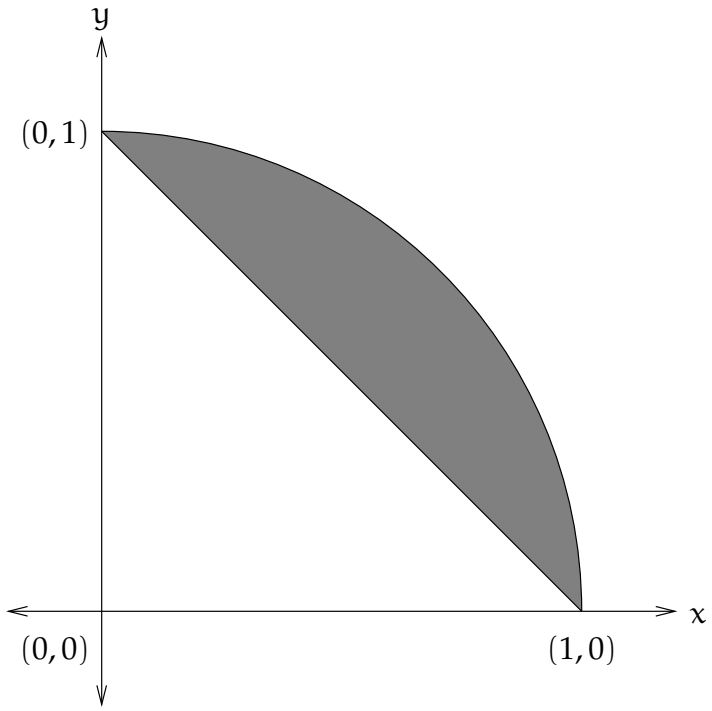


Figure 10: For any two states ρ and σ, let $x := D(\rho, \sigma)$ and $y := F(\rho, \sigma)$. The Fuchs-van de Graaf inequalities say that the point (x, y) must lie in the shaded region bounded by the line $x + y = 1$ and the arc of the unit circle in the first quadrant. This region is symmetric with respect to reflection through the line $x = y$.

26 Week 13: Quantum error correction

Quantum Error Correction. In this topic, we'll see ways to reduce the effects of noise in a quantum channel, thereby increasing the fidelity between the input state to the channel and the output state.

First, we'll see a typical scenario where this is done classically. Suppose Alice sends individual bits to Bob across a channel that is noisy in the following sense: any bit b is flipped to the opposite bit $1 - b$ with probability p, independent of the other bits. Such a channel is called the *binary symmetric channel* or *bit-flip channel*, and is an often-used model of classical noise. We can assume that $p \leqslant 1/2$, because if $p > 1/2$, then Bob would do well to flip each bit he receives, making the effective error probability $1 - p < 1/2$ per bit sent. If $p = 1/2$, then all hope is lost; no information at all can be carried by the bits; Bob receives independently random bits that are completely uncorrelated with those that Alice sent. So we'll assume that $p < 1/2$ from now on.

To reduce the chances of error per bit, Alice and Bob agree on an *binary error-correcting code*, which is some mapping

$$\begin{aligned} 0 &\mapsto c_0, \\ 1 &\mapsto c_1, \end{aligned}$$

where c_0 and c_1 are strings over the binary alphabet $\{0, 1\}$ (*binary strings*) of equal length, called *codewords*. Instead of sending each bit b, Alice sends c_b instead, and Bob decodes what he receives to (hopefully) recover b. An obvious error-correcting code is

$$\begin{aligned} 0 &\mapsto 000, \\ 1 &\mapsto 111, \end{aligned}$$

which we'll call the *majority-of-3 code*[26]. Alice wants to send a bit b (the *plaintext* or *cleartext*) to Bob, so she sends bbb across the channel. When Bob receives the possibly garbled string xyz of three bits from Alice, he decodes xyz to get the bit c as follows:

$$c := \text{majority}(x, y, z).$$

The bit b was decoded successfully iff $c = b$. What is the failure probability, *i.e.*, the probability that $c \neq b$ due to unrecoverable errors? There will be a failure iff at least two of the three bits were flipped in transit. Since each is flipped with probability p independent of the others, we have

$$\Pr[\text{failure}] = 3(1-p)p^2 + p^3 = 3p^2 - 2p^3 \,.$$

The first term in the middle is the probability that exactly two of the three bits were flipped, and the second term in the middle is the probability that all three bits were flipped. It is easy to see that $\Pr[\text{failure}] < p$ if $p < 1/2$, and so this code reduces the probability of error per plaintext bit from no encoding at all. Finally, note that $\Pr[\text{failure}] = O(p^2)$ as p tends to 0, and so for tiny $p \ll 1$, the failure probability is reduced by a considerable factor.

[26] This is an example of a *repetition code.*

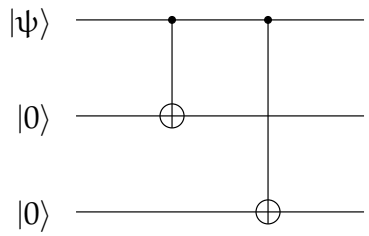


Figure 11: The three-qubit quantum majority-of-3 code. An arbitrary one-qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is encoded as $|\psi_L\rangle = \alpha|0_L\rangle + \beta|1_L\rangle = \alpha|000\rangle + \beta|111\rangle$.

The Quantum Bit-Flip Channel. Now we can "quantize" the scheme above. Suppose Alice sends qubits one at a time to Bob across a noisy quantum channel that we will call the *quantum bit-flip channel*. In the quantum bit-flip channel, a Pauli X operator is applied to each transmitted qubit with probability $p < 1/2$, independendently for each qubit. The corresponding quantum channel is thus

$$\mathcal{E}(\rho) := (1-p)\rho + pX\rho X, \tag{103}$$

whose set of Kraus operators is $\{\sqrt{1-p}\, I, \sqrt{p}\, X\}$.[27] Suppose that Alice sends some unencoded one-qubit pure state $|\psi\rangle\langle\psi|$ through the bit-flip channel $\mathcal{E}$. Ideally, Bob wants to receive $|\psi\rangle\langle\psi|$, but in reality, Bob receives $\rho' := \mathcal{E}(|\psi\rangle\langle\psi|) = (1-p)|\psi\rangle\langle\psi| + pX|\psi\rangle\langle\psi|X$. The fidelity between Alice's sent state and Bob's received state is, by Equation (102),

$$F(|\psi\rangle\langle\psi|, \rho') = \sqrt{\langle\psi|\rho'|\psi\rangle} = \sqrt{(1-p) + p\langle\psi|X|\psi\rangle^2} \geqslant \sqrt{1-p}\,,$$

with equality holding if $|\psi\rangle = |0\rangle$ or $|\psi\rangle = |1\rangle$. So the fidelity without encoding can be as low as $\sqrt{1-p}$.

Now suppose that Alice and Bob employ a quantum version of the majority-of-3 code. Alice encodes each plaintext qubit she sends to Bob as a three-qubit code state using the map

$$\begin{aligned} |0\rangle &\mapsto |0_L\rangle := |000\rangle, \\ |1\rangle &\mapsto |1_L\rangle := |111\rangle \end{aligned}$$

extended to all one-qubit pure states by linearity. Here the subscript "L" stands for "logical"—three *physical* qubits are being used to encode one *logical* (uncoded) qubit. Figure 11 shows how Alice encodes a single qubit in state $|\psi\rangle := \alpha|0\rangle + \beta|1\rangle$ as a three-qubit state $|\psi_L\rangle := \alpha|0_L\rangle + \beta|1_L\rangle$. $|\psi_L\rangle$ lies in the *code space*, *i.e.*, the two-dimensional subspace of the eight-dimensional Hilbert space of three qubits spanned by $|0_L\rangle$ and $|1_L\rangle$. The three qubits in state $|\psi_L\rangle$ are sent through the channel, and (we assume) each qubit is subjected to the $\mathcal{E}$ of Equation (103) independently of the other two. Thus the channel yields the output state

$$\sigma := (\mathcal{E} \otimes \mathcal{E} \otimes \mathcal{E})(|\psi_L\rangle\langle\psi_L|) = \mathcal{E}^{\otimes 3}(|\psi_L\rangle\langle\psi_L|)$$

[27] $\mathcal{E}$ is an example of a *mixed unitary channel*. Generally, a mixed unitary channel maps $A \in \mathcal{L}(\mathcal{H})$ to $\sum_{i=1}^{k} p_i U_i A U_i^*$, for some k, probability distribution $(p_1, \ldots, p_k)$, and unitary operators $U_1, \ldots, U_k \in \mathcal{L}(\mathcal{H})$. The Kraus operators are then $\sqrt{p_i}U_i$ for $1 \leqslant i \leqslant k$.

Bob receives σ and wants to decode it to (hopefully) recover $|\psi\rangle$. Now some issues arise that aren't a problem in the classical case. Most importantly, Bob cannot just measure the physical qubits he receives, since this will destroy the superposition making up $|\psi\rangle$. In fact, Bob's error correction operation cannot give him any classical information about $|\psi\rangle$; any such information would disrupt $|\psi\rangle$. Instead, Bob can measure what *kind* of error occurred (if any) and correct the error directly without disturbing $|\psi\rangle$. The type of error is called the *error syndrome*. Bob's decoding is a two-step process: First, Bob will measure the error syndrome, *i.e.*, which bit (if any) was flipped, without gaining any knowledge of what the values of the bit were before and after. Second, knowing which qubit was flipped, Bob applies an X gate to that qubit, and this will allow him to recover $|\psi\rangle$ with high probability.

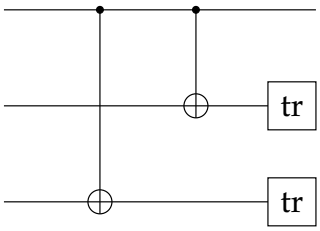
To measure the error syndrome, Bob makes a four-outcome projective measurement on his three received qubits using the four projectors

$$\begin{aligned} P_0 &= |000\rangle\langle 000| + |111\rangle\langle 111|, \\ P_1 &= |100\rangle\langle 100| + |011\rangle\langle 011|, \\ P_2 &= |010\rangle\langle 010| + |101\rangle\langle 101|, \\ P_3 &= |001\rangle\langle 001| + |110\rangle\langle 110|. \end{aligned}$$

$P_0, \ldots, P_3$ form a complete set of projectors, and each P_j is a two-dimensional projector. P_0 projects onto the code space and corresponds to the outcome of either no qubits flipped or all three qubits flipped. P_1 corresponds to the outcome of either the first qubit flipped and the other two left alone, or the other two flipped and the first left alone. P_2 and P_3 are similar for the second and third qubits, respectively. Note that, whatever the state was before the syndrome measurement, the post-measurement state is in one of the four subspaces projected onto by $P_0, \ldots, P_3$, respectively.

Let $j \in \{0, 1, 2, 3\}$ be the outcome of Bob's syndrome measurement, above. After the measurement, Bob tries to recover $|\psi_L\rangle$ as follows: if $j = 0$, then Bob assumes that no qubits were flipped (which is way more likely than all three being flipped), and so he does nothing; if $1 \leqslant j \leqslant 3$, then Bob assumes that the jth qubit was flipped (which is somewhat more likely than the other two being flipped), and so he flips the jth qubit back by applying an X gate to it. No matter what qubits were flipped in the channel, Bob has a state in the code space after the correction. If at most one qubit was flipped, then Bob has $|\psi_L\rangle$, and the recovery is successful. If more than one qubit was flipped, then Bob has the state $X_L|\psi_L\rangle$, where X_L is some three-qubit operator that swaps $|0_L\rangle$ with $|1_L\rangle$, and the recovery failed. (Bob doesn't know at this point whether he succeeded or failed.)

In a moment, we'll see in detail how Bob can perform these steps, but once he has $|\psi_L\rangle$—and if he really wants to—he can convert $|\psi_L\rangle$ back into $|\psi\rangle$ by applying the circuit



which is the inverse of Alice's circuit of Figure 11. The two gates labeled "tr" are what I call *trace gates*. A trace gate just signifies that a qubit is no longer useful and is to be ignored, *i.e.*, traced

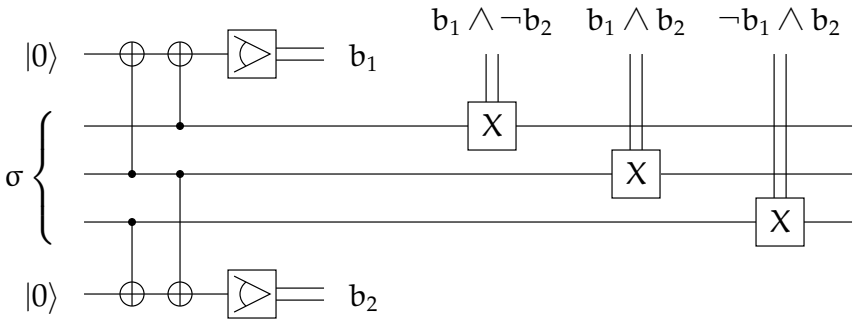


Figure 12: Bob's error-recovery circuit for the quantum bit-flip channel. The middle three qubits are what he receives from Alice, and the two outer qubits are ancillæ used in the syndrome measurement.

out. So mathematically, a trace gate corresponds to a partial trace. Assuming the input state to the circuit is $|\psi_L\rangle$, the first qubit of the output will be in state $|\psi\rangle$. The traced-out qubits will both hold $|0\rangle$ if the input state is in the code space. Bob usually does *not* want to recover $|\psi\rangle$ by decoding if the encoded state will be used for further computation, because those computations are more fault-tolerant using the encoded state.

A circuit for Bob's syndrome measurement and subsequent correction is shown in Figure 12. Bob received the three-qubit state σ in the middle three qubits. His syndrome measurement is split into two binary measurements: He first measures whether the first two of the three qubits from Alice are different. The value b_1 measured on the upper ancilla will be 1 iff they are, and 0 otherwise. In other words, b_1 gives the parity (sum modulo 2) of the values of the first two qubits. Similarly, the lower ancilla measurement is 1 iff the second two qubits are different. To correct the state, Bob combines these two Boolean values to determine which qubit value, if any, is different from the other two, and applies a classically controlled X gate to this qubit.

Exercise 26.1 Show mathematically that the syndrome measurement portion of Figure 12 is the same as the projective measurement $\{P_0, P_1, P_2, P_3\}$ described earlier. What values of $b_1 b_2$ correspond to which P_j?

Bob's entire recovery process in Figure 12 can be described as a quantum channel $\mathcal{R}$ that maps three-qubit states to three-qubit states: For input state σ, we have

$$\mathcal{R}(\sigma) = P_0 \sigma P_0 + \sum_{j=1}^{3} X_j P_j \sigma P_j X_j, \tag{104}$$

where X_j is the Pauli X gate applied to the jth qubit. That is,

$$\begin{aligned} X_1 &= X \otimes I \otimes I, \\ X_2 &= I \otimes X \otimes I, \\ X_3 &= I \otimes I \otimes X. \end{aligned}$$

Thus the state after Bob's recovery is

$$\tau := \mathcal{R}(\sigma) = \mathcal{R}(\mathcal{E}^{\otimes 3}(|\psi_L\rangle\langle\psi_L|)).$$

To get a handle on what τ is, first notice that $|\psi_L\rangle\langle\psi_L|$ is a linear combination of operators of the form $|a_L\rangle\langle b_L| = |aaa\rangle\langle bbb|$ for $a, b \in \{0, 1\}$. By Equation (103) we have $\mathcal{E}(|a\rangle\langle b|) = (1-p)|a\rangle\langle b| + p|\bar{a}\rangle\langle\bar{b}|$, where we let $\bar{a} := 1 - a$ and $\bar{b} := 1 - b$. Then,

$$\begin{aligned}
\mathcal{E}^{\otimes 3}(|aaa\rangle\langle bbb|) &= \mathcal{E}^{\otimes 3}\left[(|a\rangle\langle b|)^{\otimes 3}\right] \\
&= [\mathcal{E}(|a\rangle\langle b|)]^{\otimes 3} \\
&= [(1-p)|a\rangle\langle b| + p|\bar{a}\rangle\langle\bar{b}|]^{\otimes 3} \\
&= (1-p)^3|aaa\rangle\langle bbb| \\
&\quad + (1-p)^2 p\,(|\bar{a}aa\rangle\langle\bar{b}bb| + |a\bar{a}a\rangle\langle b\bar{b}b| + |aa\bar{a}\rangle\langle bb\bar{b}|) \\
&\quad + (1-p)p^2\,(|a\bar{a}\bar{a}\rangle\langle b\bar{b}\bar{b}| + |\bar{a}a\bar{a}\rangle\langle\bar{b}b\bar{b}| + |\bar{a}\bar{a}a\rangle\langle\bar{b}\bar{b}b|) \\
&\quad + p^3|\bar{a}\bar{a}\bar{a}\rangle\langle\bar{b}\bar{b}\bar{b}|.
\end{aligned}$$

(Alternatively, we can expand $\mathcal{E}^{\otimes 3}(\rho)$ for any three-qubit operator ρ to get an operator-sum expression for $\mathcal{E}^{\otimes 3}$:

$$\begin{aligned}
\mathcal{E}^{\otimes 3}(\rho) &= (1-p)^3\rho \\
&\quad + (1-p)^2 p(X_1\rho X_1 + X_2\rho X_2 + X_3\rho X_3) \\
&\quad + (1-p)p^2(X_2X_3\rho X_2X_3 + X_1X_3\rho X_1X_3 + X_1X_2\rho X_1X_2) \\
&\quad + p^3 X_1X_2X_3\rho X_1X_2X_3,
\end{aligned}$$

then plug in $|aaa\rangle\langle bbb|$ for ρ to get the same expression for $\mathcal{E}^{\otimes 3}(|aaa\rangle\langle bbb|)$.) Applying the $\mathcal{R}$ of Equation (104) to $\mathcal{E}^{\otimes 3}(|aaa\rangle\langle bbb|)$ above, we get, after much simplification,

$$\begin{aligned}
\mathcal{R}(\mathcal{E}^{\otimes 3}(|a_L\rangle\langle b_L|)) &= (1 - 3p^2 + 2p^3)|aaa\rangle\langle bbb| + (3p^2 - 2p^3)|\bar{a}\bar{a}\bar{a}\rangle\langle\bar{b}\bar{b}\bar{b}| \\
&= (1 - 3p^2 + 2p^3)|a_L\rangle\langle b_L| + (3p^2 - 2p^3)X_1X_2X_3|a_L\rangle\langle b_L|X_1X_2X_3.
\end{aligned}$$

Exercise 26.2 Verify this last equation. This may be tedious, but it is good practice.

Since this equation holds for all four bowtie operators $|a_L\rangle\langle b_L|$, by linearity, we get

$$\tau = (1 - 3p^2 + 2p^3)|\psi_L\rangle\langle\psi_L| + (3p^2 - 2p^3)X_1X_2X_3|\psi_L\rangle\langle\psi_L|X_1X_2X_3.$$

The first term represents Bob's successful recovery of $|\psi_L\rangle$, and this occurs with probability $1 - 3p^2 + 2p^3$, which is greater than $1 - p$ if $p < 1/2$. In fact, it is $1 - O(p^2)$, which is significant if p is small. For the fidelity, we get

$$F(|\psi_L\rangle\langle\psi_L|, \tau) = \sqrt{\langle\psi_L|\tau|\psi_L\rangle} \geqslant \sqrt{1 - 3p^2 + 2p^3} > \sqrt{1-p},$$

and so the minimum fidelity of τ with $|\psi_L\rangle\langle\psi_L|$ is strictly greater than the minimum fidelity of σ with $|\psi_L\rangle\langle\psi_L|$. So, recovery improves the worst-case fidelity.

The Quantum Phase-Flip Channel. Bit flips are not the only possible errors in a quantum channel. Consider the one-qubit *phase-flip channel* given by

$$\mathcal{F}(\rho) := (1-p)\rho + pZ\rho Z,$$

which applies a Pauli Z operator to the qubit (thus flipping the relative phase between $|0\rangle$ and $|1\rangle$ by a factor of -1) with probability $p < 1/2$.

This kind of channel has no classical analogue, but in a very real sense it is closely analogous to the quantum bit-flip channel—the two channels are "unitarily conjugate" to each other via the Hadamard H operator. Here's what I mean by that: Since $HX = ZH$ and $XH = HZ$, we have

$$H(\mathcal{F}(\rho))H = (1-p)H\rho H + pHZ\rho ZH = (1-p)H\rho H + pXH\rho HX = \mathcal{E}(H\rho H)$$

for every one-qubit operator ρ. Similarly, $H(\mathcal{E}(\rho))H = \mathcal{F}(H\rho H)$.[28] So by conjugating everything by H on each qubit, we can reduce the problem of error recovery in the phase-flip channel to that of error recovery in the bit-flip channel.

Compare the following with the previous discussion about the quantum bit-flip channel: If Alice sends a one-qubit pure state $|\psi\rangle\langle\psi|$ unencoded across the channel $\mathcal{F}$ to Bob, then Bob receives some $\rho' = \mathcal{F}(|\psi\rangle\langle\psi|) = (1-p)|\psi\rangle\langle\psi| + pZ|\psi\rangle\langle\psi|Z$. The fidelity between $|\psi\rangle\langle\psi|$ and ρ' is

$$F(|\psi\rangle\langle\psi|, \rho') = \sqrt{\langle\psi|\rho'|\psi\rangle} = \sqrt{(1-p) + p\langle\psi|Z|\psi\rangle^2} \geqslant \sqrt{1-p},$$

with equality holding if $|\psi\rangle = H|0\rangle$ or $|\psi\rangle = H|1\rangle$. So the worst-case fidelity is the same as with the bit-flip channel.

To get an error-correcting code for the phase-flip channel, we take our majority-of-3 construction for the bit-flip channel and insert Hadamard gates in the right places. Recall that we've defined $|+\rangle := H|0\rangle$ and $|-\rangle := H|1\rangle$. Alice now encodes her one-qubit pure state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ as

$$\left|\psi'_L\right\rangle := H^{\otimes 3}(\alpha|000\rangle + \beta|111\rangle) = \alpha|{+}{+}{+}\rangle + \beta|{-}{-}{-}\rangle = \alpha\left|0'_L\right\rangle + \beta\left|1'_L\right\rangle,$$

where $\left|0'_L\right\rangle := |{+}{+}{+}\rangle$, $\left|1'_L\right\rangle := |{-}{-}{-}\rangle$, and $H^{\otimes 3} = H_1H_2H_3$, defined analogously with X_1, X_2, and X_3 previously. Figure 13 shows the circuit Alice uses to do this.

Note that $Z|+\rangle = |-\rangle$ and $Z|-\rangle = |+\rangle$. In other words, Z is represented in the $\{|+\rangle, |-\rangle\}$ basis by the same matrix as X is in the computational basis. So a phase flip in the channel $\mathcal{F}$ will flip a $+$ to a $-$ and vice versa. This means that we can do the same analysis of the channel $\mathcal{F}$ as we did with $\mathcal{E}$ by substituting the labels $+$ for 0 and $-$ for 1. Bob receives the state $\sigma' := \mathcal{F}^{\otimes 3}(\left|\psi'_L\right\rangle\left\langle\psi'_L\right|)$ from Alice, measures the error syndrome with projectors Q_0, Q_1, Q_2, Q_3, where each $Q_j := H^{\otimes 3}P_jH^{\otimes 3}$. If Bob sees that the relative phase of one of the qubits is different from that of the other two, then Bob assumes that the qubit's phase was flipped and applies a Z gate to that qubit. The circuit for doing all this is shown in Figure 14.

The quantum channel corresponding to Bob's whole procedure is given by

$$\mathcal{S}(\sigma) \quad := \quad Q_0\sigma Q_0 + \sum_{j=1}^{3} Z_jQ_j\sigma Q_jZ_j = H^{\otimes 3}P_0H^{\otimes 3}\sigma H^{\otimes 3}P_0H^{\otimes 3} + \sum_{j=1}^{3} Z_jH^{\otimes 3}P_jH^{\otimes 3}\sigma H^{\otimes 3}P_jH^{\otimes 3}Z_j$$

[28] Put more succinctly, $\mathcal{U} \circ \mathcal{F} = \mathcal{E} \circ \mathcal{U}$ and $\mathcal{U} \circ \mathcal{E} = \mathcal{F} \circ \mathcal{U}$, where $\mathcal{U}$ is the one-qubit unitary quantum channel that maps $\rho \mapsto H\rho H$.

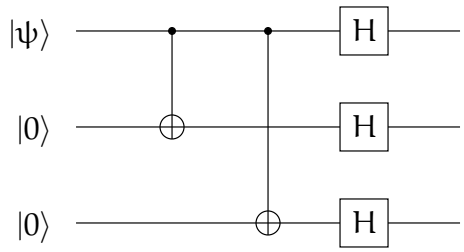


Figure 13: The three-qubit code for the phase-flip channel. An arbitrary one-qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is encoded as $|\psi'_L\rangle = \alpha|0'_L\rangle + \beta|1'_L\rangle = \alpha|{+}{+}{+}\rangle + \beta|{-}{-}{-}\rangle$.

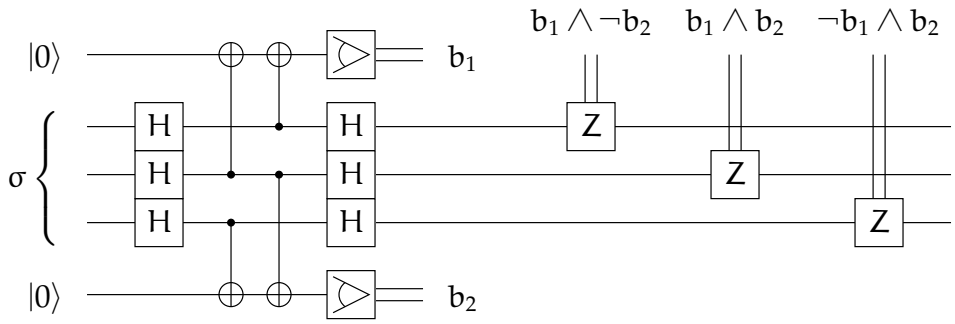


Figure 14: Bob's error recovery procedure for the phase-flip channel.

$$
\begin{aligned}
&= H^{\otimes 3}P_0H^{\otimes 3}\sigma H^{\otimes 3}P_0H^{\otimes 3} + \sum_{j=1}^{3} H^{\otimes 3}X_jP_jH^{\otimes 3}\sigma H^{\otimes 3}P_jX_jH^{\otimes 3} \\
&= H^{\otimes 3}\left(P_0(H^{\otimes 3}\sigma H^{\otimes 3})P_0 + \sum_{j=1}^{3} X_jP_j(H^{\otimes 3}\sigma H^{\otimes 3})P_jX_j\right)H^{\otimes 3} = H^{\otimes 3}\left(\mathcal{R}(H^{\otimes 3}\sigma H^{\otimes 3})\right)H^{\otimes 3}
\end{aligned}
$$

and thus

$$H^{\otimes 3}(\mathcal{S}(\sigma))H^{\otimes 3} = \mathcal{R}(H^{\otimes 3}\sigma H^{\otimes 3}) \tag{105}$$

for any three-qubit operator σ, where $\mathcal{R}$ is the bit-flip recovery channel of Equation (104). That is, $\mathcal{S}$ is unitarily conjugate to $\mathcal{R}$ via $H^{\otimes 3}$. In a similar fashion, we can get that

$$H^{\otimes 3}(\mathcal{F}^{\otimes 3}(\rho))H^{\otimes 3} = \mathcal{E}^{\otimes 3}(H^{\otimes 3}\rho H^{\otimes 3}) \tag{106}$$

for any three-qubit operator ρ.

Letting $\tau' := \mathcal{S}(\mathcal{F}^{\otimes 3}(|\psi'_L\rangle\langle\psi'_L|))$ and stringing these channels together, (105) and (106) give us

$$
\begin{aligned}
H^{\otimes 3}\tau' H^{\otimes 3} &= H^{\otimes 3}(\mathcal{S}(\mathcal{F}^{\otimes 3}(|\psi'_L\rangle\langle\psi'_L|)))H^{\otimes 3} \\
&= \mathcal{R}(H^{\otimes 3}(\mathcal{F}^{\otimes 3}(|\psi'_L\rangle\langle\psi'_L|))H^{\otimes 3}) \\
&= \mathcal{R}(\mathcal{E}^{\otimes 3}(H^{\otimes 3}|\psi'_L\rangle\langle\psi'_L|H^{\otimes 3})) \\
&= \mathcal{R}(\mathcal{E}^{\otimes 3}(|\psi_L\rangle\langle\psi_L|)) \\
&= \tau\,,
\end{aligned}
$$

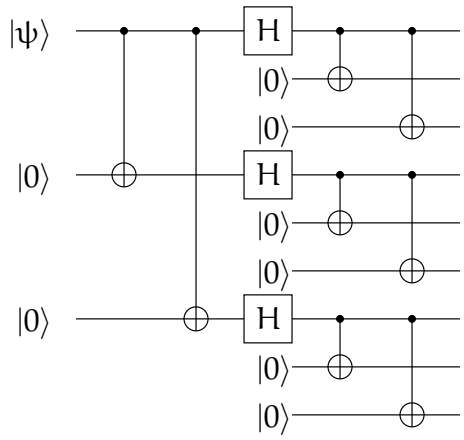


Figure 15: The nine-qubit Shor code. This concatenates the phase-flip and bit-flip codes.

or equivalently,

$$\tau' = H^{\otimes 3}\tau H^{\otimes 3} = (1 - 3p^2 + 2p^3)\left|\psi'_L\right\rangle\left\langle\psi'_L\right| + (3p^2 - 2p^3)Z_1Z_2Z_3\left|\psi'_L\right\rangle\left\langle\psi'_L\right|Z_1Z_2Z_3.$$

Thus we get the same success probability here as with the bit-flip channel, and the fidelity is at worst the same as it was then:

$$F(\left|\psi'_L\right\rangle\left\langle\psi'_L\right|, \tau') = \sqrt{\left\langle\psi'_L|\tau'|\psi'_L\right\rangle} \geqslant \sqrt{1 - 3p^2 + 2p^3}.$$

The Shor Code. We can combine the bit-flip and phase-flip error correcting codes above to correct against both kinds of errors, even on the same qubit. As a bonus, we'll show that the resulting code corrects against *arbitrary* errors on a single qubit. A typical one-qubit channel that has all three kinds of errors (bit flip, phase flip, and combined bit and phase flip) is called the *depolarizing channel*, and it maps

$$\rho \mapsto \mathcal{D}(\rho) := (1-p)\rho + \frac{p}{3}(X\rho X + Z\rho Z + ZX\rho XZ) = (1-p)\rho + \frac{p}{3}(X\rho X + Y\rho Y + Z\rho Z). \quad (107)$$

This channel leaves the qubit alone with probability $1 - p > 1/2$ and produces each of the three possible errors with the same probability $p/3$.

To help correct against all three types of errors, Alice first encodes a single qubit using the three-qubit phase-flip code of Figure 13, then she encodes *each* of the three qubits using the majority-of-3 code for the bit-flip channel, shown in Figure 11. The resulting encoding circuit, shown in Figure 15, produces the nine-qubit *Shor code*, named after its inventor, Peter Shor. Such a code is called a *concatenated code*, in that it combines (concatenates) two or more simpler codes into a single code. Using the Shor code, Alice encodes a single qubit in state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ as the nine-qubit state $|\psi_S\rangle := \alpha|0_S\rangle + \beta|1_S\rangle$, where

$$|0_S\rangle := \frac{1}{2\sqrt{2}}(|000\rangle + |111\rangle)^{\otimes 3} = |+_L\rangle^{\otimes 3}, \quad (108)$$

$$|1_S\rangle := \frac{1}{2\sqrt{2}}(|000\rangle - |111\rangle)^{\otimes 3} = |-_L\rangle^{\otimes 3}, \quad (109)$$

where we define the three-qubit states $|+_L\rangle := (|000\rangle + |111\rangle)/\sqrt{2}$ and $|-_L\rangle := (|000\rangle - |111\rangle)/\sqrt{2}$. The nine qubits are naturally divided into three subblocks of three qubits each, which I'll call *3-blocks*. Alice sends Bob $|\psi_S\rangle\langle\psi_S|$ through a channel (*e.g.*, the depolarizing channel) that may cause one of the three errors on each of the nine qubits with some probability independently of the others. If more than one qubit is affected, then the recovery may not work, and so we hope that the probability of this happening is low.

Bob receives the nine-qubit state σ sent from Alice, and we'll assume (with high probability) that at most one of the nine qubits endured either a bit flip, phase flip, or both. For example, suppose Alice sends $|0_S\rangle = |+_L\rangle^{\otimes 3}$ to Bob. If the first qubit is bit-flipped en route, then Bob receives $(1/\sqrt{2})(|100\rangle + |011\rangle) \otimes |+_L\rangle^{\otimes 2}$. If the first qubit is phase-flipped, then Bob gets the state $(1/\sqrt{2})(|000\rangle - |111\rangle) \otimes |+_L\rangle^{\otimes 2} = |-_L\rangle \otimes |+_L\rangle^{\otimes 2}$. (Note that a phase flip in a qubit contributes an overall phase flip in its 3-block; phase flips on two different qubits in the same block would cancel each other.) Finally, if the first qubit is bit-flipped and then phase-flipped, then Bob gets $(1/\sqrt{2})(-|100\rangle + |011\rangle) \otimes |+_L\rangle^{\otimes 2}$.

To recover, Bob first applies the bit-flip error recovery operation $\mathcal{R}$ of Figure 12 and Equation (104) to each of the three 3-blocks independently. This will correct up to a single bit-flip error within each 3-block. Crucially, this intrablock bit-flip recovery works regardless of whether there was also a phase-flip in the 3-block. After Bob corrects bit flips within each 3-block, he must then correct phase flips. He does this by comparing phase differences between adjacent 3-blocks, either finding which 3-block's phase doesn't match the other two and flipping that 3-block's phase back, or else determining that the phases of the 3-blocks are all equal and nothing needs to be done. A circuit that accomplishes this phase-flip recovery portion of the overall recovery is shown in Figure 16. In essence, Bob's procedure first applies a bank of H gates to turn phase flips into bit flips, then it measures the bit flips (bit parity), then converts the state back into phase flips (more H gates), which it can then correct with Z gates based on what bit flips were measured. To see that this all works, define

$$\begin{aligned} |\text{even}\rangle &:= \frac{1}{2}(|000\rangle + |011\rangle + |101\rangle + |110\rangle), \\ |\text{odd}\rangle &:= \frac{1}{2}(|100\rangle + |010\rangle + |001\rangle + |111\rangle). \end{aligned}$$

$|\text{even}\rangle$ is a superposition of all computational basis states of three qubits with an even number of 1's; $|\text{odd}\rangle$ is a superposition of the other computational basis states. One can readily check that $|\text{even}\rangle = H^{\otimes 3}|+_L\rangle$ and that $|\text{odd}\rangle = H^{\otimes 3}|-_L\rangle$. Suppose for example that Alice sends $|1_S\rangle = |-_L\rangle|-_L\rangle|-_L\rangle$ to Bob (we are surpressing the $\otimes$ operator symbol for now), and there is a phase flip on some qubit in the first 3-block (it doesn't matter which). So after correcting bit flips, Bob's state is now $|+_L\rangle|-_L\rangle|-_L\rangle$, which feeds into the circuit of Figure 16. Applying the Hadamard gates yields the state $|\text{even}\rangle|\text{odd}\rangle|\text{odd}\rangle$. As a result of the CNOTs, the upper ancilla's bit value will flip an even + odd = odd number of times, and so $b_1 = 1$. The lower ancilla's bit value will flip an odd + odd = even number of times, and so $b_2 = 0$. The next layer of Hadamards converts the state back to $|+_L\rangle|-_L\rangle|-_L\rangle$, and then Bob recovers by applying a Z gate to the first qubit, yielding $|-_L\rangle|-_L\rangle|-_L\rangle = |1_S\rangle$.

Let's look briefly at the quantum channels involved. Let $\mathcal{T}$ be the quantum channel corresponding to Bob's entire recovery procedure for the Shor code.

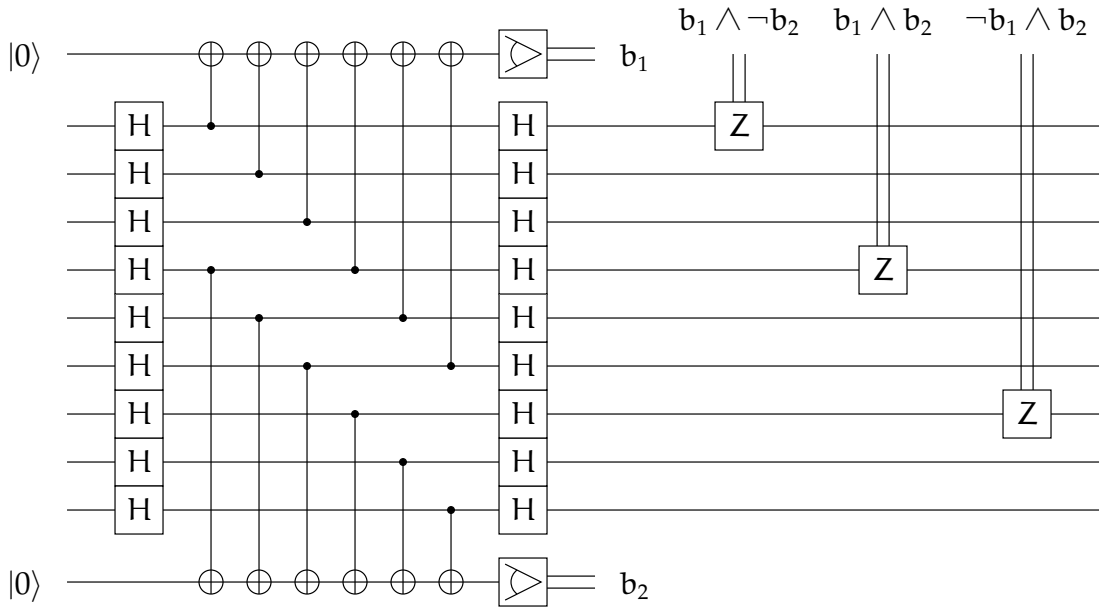


Figure 16: Recovering from a phase flip with the Shor code. When this circuit starts, Bob assumes that he has already corrected any bit flip in a 3-block if there was one, and so the incoming state is a linear combination of the eight states $|\pm_L\rangle|\pm_L\rangle|\pm_L\rangle$.

Exercise 26.3 (Challenging) Give an expression for $\mathcal{T}$ applied to an arbitrary nine-qubit state ρ. Make your expression as succinct as possible but still mathematically precise. You may use the following notations for operators without having to expand them:

- Let R_0, R_1, R_2, R_3 represent the projectors for the measurement performed in Figure 16, corresponding to the outcomes 00, 10, 11, and 01, respectively, for $b_1 b_2$.
- For $j \in \{0, 1, 2\}$, let $P_0^{(j)}, P_1^{(j)}, P_2^{(j)}, P_3^{(j)}$ be the projectors used for the bit-flip syndrome measurement in the $(j+1)$st 3-block, as described in the bit-flip channel discussion.
- For any single-qubit operator A and $k \in \{1, \ldots, 9\}$, let A_k be the nine-qubit operator that applies A to the kth qubit and leaves the other qubits alone.

Suppose the channel between Alice and Bob is the one-qubit depolarizing channel $\mathcal{D}$ of Equation (107). If Alice and Bob use the Shor code, then nine qubits will be transferred per single plaintext qubit. For an arbitrary nine-qubit state ρ, the effect of $\mathcal{D}$ on ρ is then

$$\mathcal{D}^{\otimes 9}(\rho) = (1-p)^9\rho + (1-p)^8\frac{p}{3}\sum_{j=1}^{9}(X_j\rho X_j + Y_j\rho Y_j + Z_j\rho Z_j) + O(p^2).$$

Where the terms hidden in the "$O(p^2)$" represent errors on two or more qubits, from which Bob might not recover. Bob, however, *can* recover from any of the single-qubit errors showing in the expression above, provided ρ is in the code space of the Shor code. That is, if $\rho = |\psi_S\rangle\langle\psi_S|$ for some

one-qubit state $|\psi\rangle$, then we've shown that $\mathcal{T}(X_j \rho X_j) = \mathcal{T}(Y_j \rho Y_j) = \mathcal{T}(Z_j \rho Z_j) = \rho$ for all $1 \leqslant j \leqslant 9$, and thus the final error-corrected state is

$$\upsilon := \mathcal{T}(\mathcal{D}^{\otimes 9}(\rho)) = ((1-p)^9 + 9p(1-p)^8)\rho + O(p^2) = (1-p)^8(1+8p)\rho + O(p^2).$$

The hidden terms are all of the form $K\rho K^* = K|\psi_S\rangle\langle\psi_S|K^*$ for some Kraus operators K, and thus the fidelity is

$$\begin{aligned} F(|\psi_S\rangle\langle\psi_S|, \upsilon) &= \sqrt{\langle\psi_S|\upsilon|\psi_S\rangle} \\ &= \sqrt{(1-p)^8(1+8p) + (\text{nonnegative terms})} \\ &\geqslant \sqrt{(1-p)^8(1+8p)} \\ &= (1-p)^4\sqrt{1+8p} \\ &= 1 - O(p^2). \end{aligned}$$

The value $(1-p)^4\sqrt{1+8p}$ is actually an underestimate for the minimum fidelity, because using $\mathcal{T}$ Bob can correct some errors involving more than one qubit—for example, bit-flips of qubits in different 3-blocks, or even three phase flips and one bit flip within a single 3-block. The only errors he potentially cannot recover from are either two or more bit flips within the same 3-block, or net phase flips in two or more different 3-blocks, and even then, Bob can recover from some of these errors.

Exercise 26.4 Show that using the Shor code, Bob recovers from $X_2X_4X_9Z_1Z_2Z_3Z_4Z_5Z_7Z_9$.

Exercise 26.5 (Challenging) What is the worst-case fidelity of sending an unencoded one-qubit pure state $|\psi\rangle\langle\psi|$ through the depolarizing channel $\mathcal{D}$? About how small does p have to be so that the worst-case estimate of the fidelity for the Shor code, above, is better than this? A numerical approximation will suffice.

27 Week 13: Error correction (cont.)

Quantum Error Correction: The General Theory. Here we want to determine, in the most general terms that we can, when it is possible to recover from a noisy quantum channel through the use of an error correcting code. Let $\mathcal{H}$ be the Hilbert space of states that are to be sent through some noisy channel. We'll assume that information sent through the channel is encoded into states in some linear subspace $C \subseteq \mathcal{H}$, *i.e.*, the *code space*. Let P be the projector that projects orthogonally onto C. We'll assume that the noisy channel is modeled by some (possibly incomplete) quantum *error channel* $\mathcal{E} \in \mathcal{T}(\mathcal{H})$. For example, in our discussion of the Shor code, $\mathcal{H}$ is the 2^9-dimensional space of nine qubits, and C is the 2-dimensional subspace spanned by the vectors $|0_S\rangle$ and $|1_S\rangle$; the noisy channel $\mathcal{E}$ of interest may be the portion of the depolarizing channel $\mathcal{D}^{\otimes 9}$ in which at most one qubit is affected. This channel sends a state $\rho \in \mathcal{L}(\mathcal{H})$ to

$$\mathcal{E}(\rho) := (1-p)^9\rho + (1-p)^8 p/3 \sum_{j=1}^{9} (X_j \rho X_j + Y_j \rho Y_j + Z_j \rho Z_j), \tag{110}$$

and represents the portion of $\mathcal{D}^{\otimes 9}$ from which we know Bob can recover. Note that $\mathcal{E}$ is an incomplete (non-trace-preserving) channel, because we are omitting the terms of $\mathcal{D}^{\otimes 9}$ where more than one qubit is subjected to an error, and from which Bob may not be able to recover. The incompleteness reflects the fact that this unrecoverability happens with nonzero probability.

Back to the general case. We'll say that a quantum state ρ is in the code space C iff it is a convex sum of pure states in C, *i.e.*, $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$ where each $|\psi_i\rangle \in C$. Equivalently, ρ is in the code space iff $\rho = P\rho P$ (equivalently, $\rho = P\rho$, or equivalently, $\rho = \rho P$, by Exercise 27.1, below).

Exercise 27.1 Prove that the following are equivalent for any projector P and Hermitean operator A: (1) $A = PAP$; (2) $A = AP$; (3) $A = PA$. [Hint: No decompositions are needed for any of these—just simple substitutions and taking adjoints.]

The error channel $\mathcal{E}$ can be given in operator-sum form by some Kraus operators $E_1, \ldots, E_N \in \mathcal{L}(\mathcal{H})$ such that $\sum_{j=1}^{N} E_j^* E_j \leqslant I$ (inequality because $\mathcal{E}$ is not necessarily complete), and

$$\mathcal{E}(\rho) = \sum_{j=1}^{N} E_j \rho E_j^*$$

for any $\rho \in \mathcal{L}(\mathcal{H})$. Suppose that $\mathcal{R} \in \mathcal{T}(\mathcal{H})$ is some (not necessarily complete) quantum channel representing a recovery procedure. We will say that $\mathcal{R}$ *successfully recovers from* $\mathcal{E}$ if $(\mathcal{R} \circ \mathcal{E})(\rho) = c\rho$ for any ρ in C, where c is a real constant depending on ρ, $\mathcal{E}$, and $\mathcal{R}$ and satisfying $0 \leqslant c \leqslant 1$. (If $\mathcal{E}$ and $\mathcal{R}$ are both complete (hence trace-preserving), then we must have $c = 1$.) We will say that $\mathcal{E}$ is *recoverable* if there exists an $\mathcal{R}$ that successfully recovers from $\mathcal{E}$. The next theorem gives a quantitative criterion for when an error channel is recoverable.

Theorem 27.2 *Let $\mathcal{E}$ be a nonzero, possibly incomplete error channel on $\mathcal{L}(\mathcal{H})$ given by Kraus operators $E_1, \ldots, E_N \in \mathcal{L}(\mathcal{H})$. Fix a code space $C \subseteq \mathcal{H}$ and let P be the projector projecting onto C. $\mathcal{E}$ is recoverable (with respect to C) if and only if there exists an $N \times N$ matrix M such that, for all $1 \leqslant i, j \leqslant N$,*

$$PE_i^* E_j P = [M]_{ij} P. \tag{111}$$

Further, if such an M exists, then $M \geqslant 0$, $\operatorname{tr} M$ is the probability that $\mathcal{E}$ occurs given any state in C, and a (complete) recovery channel $\mathcal{R}$ exists such that $(\mathcal{R} \circ \mathcal{E})(\rho) = (\operatorname{tr} M)\rho$ for any ρ in C.

I call Equation (111) the "peep" condition, because of the left-hand side. We will only be interested in the backwards implication, giving sufficient conditions for $\mathcal{E}$ to be recoverable by actually constructing a recovery procedure. So we won't prove the forward implication (the Nielsen & Chuang textbook does it). Here is some intuition about the forward implication: Suppose $\mathcal{E}$ is recoverable. Let $|\psi\rangle$ and $|\varphi\rangle$ be any two states in the code space C. Then $P|\psi\rangle = |\psi\rangle$ and $P|\varphi\rangle = |\varphi\rangle$, and so for all i, j, we have $\langle\psi|E_i^* E_j|\varphi\rangle = \langle\psi|PE_i^* E_j P|\varphi\rangle$. The left-hand side is the inner product of the vectors $E_i|\psi\rangle$ and $E_j|\varphi\rangle$. Equation (111) is equivalent to saying that this value is proportional to $\langle\psi|P|\varphi\rangle = \langle\psi|\varphi\rangle$, where the constant of proportionality ($[M_{ij}]$) depends only on i and j and not on $|\psi\rangle$ or $|\varphi\rangle$. One might expect that this is needed, so that the error operators do not "distort" the code space, i.e., they preserve inner products up to a constant, because to recover, we must apply a unitary operation to restore the code space undistorted, so that superpositions of states in the code space are preserved up to a constant.

Proof. [backward implication] Suppose that M exists satisfying (111) for all i, j. We can assume that $P \neq 0$; otherwise, the theorem is trivial. Taking the adjoint of each side of (111), we have, for all $1 \leqslant i, j \leqslant N$,

$$[M]_{ij}^* P = P E_j^* E_i P = [M]_{ji} P,$$

and so $[M]_{ij}^* = [M]_{ji}$ since $P \neq 0$, which means that M is Hermitean. The next thing to do is to simplify (111) by diagonalizing M. Since M is normal, there is an $N \times N$ unitary matrix U and scalars $d_1, \ldots, d_N \in \mathbb{R}$ (the eigenvalues of M) such that $U^* M U = \operatorname{diag}(d_1, \ldots, d_N)$. For $1 \leqslant k \leqslant N$, define

$$F_k := \sum_{j=1}^{N} [U]_{jk} E_j.$$

Then

$$\sum_{k=1}^{N} F_k^* F_k = \sum_{i,j=1}^{N} \left(\sum_k [U]_{jk} [U]_{ik}^* \right) E_i^* E_j = \sum_{i,j} \delta_{ji} E_i^* E_j = \sum_j E_j^* E_j \leqslant I,$$

and for any $\rho \in \mathcal{L}(\mathcal{H})$,

$$\sum_{k=1}^{N} F_k \rho F_k^* = \sum_{i,j=1}^{N} \left(\sum_k [U]_{ik} [U]_{jk}^* \right) E_i \rho E_j^* = \sum_{i,j} \delta_{ij} E_i \rho E_j^* = \sum_j E_j \rho E_j^* = \mathcal{E}(\rho).$$

Thus $F_1, \ldots, F_N$ are also a set of Kraus operators for $\mathcal{E}$. Now Equation (111) becomes, for all $1 \leqslant k, \ell \leqslant N$,

$$P F_k^* F_\ell P = \sum_{i,j=1}^{N} [U]_{ik}^* [U]_{j\ell} P E_i^* E_j P = \sum_{i,j} [U^*]_{ki} [M]_{ij} [U]_{j\ell} P = \sum_{i,j} [U^* M U]_{k\ell} P = d_k \delta_{k\ell} P. \tag{112}$$

Taking the trace of both sides of (112) with $k = \ell$, we get

$$d_k = \frac{\operatorname{tr}(P F_k^* F_k P)}{\operatorname{tr} P} = \frac{\langle F_k P, F_k P \rangle}{\operatorname{tr} P} \geqslant 0 .$$

This implies $M \geqslant 0$. Also, for any state ρ in C, the probability that $\mathcal{E}$ actually occurs is given by

$$\operatorname{tr}(\mathcal{E}(\rho)) = \operatorname{tr}(\mathcal{E}(P\rho P)) = \sum_{k=1}^{N} \operatorname{tr}(F_k P \rho P F_k^*) =$$
$$\sum_k \operatorname{tr}(P F_k^* F_k P \rho) = \sum_k d_k \operatorname{tr}(P\rho) = \sum_k d_k \operatorname{tr}\rho = \sum_k d_k = \operatorname{tr} M\,.$$

Note that if $d_k = 0$ for some k, then $\langle F_k P, F_k P\rangle = 0$, and so $F_k P = 0$. This implies that if ρ is any state in C, then $F_k \rho F_k^* = F_k P \rho P F_k^* = 0$, and so this term is dropped from the operator-sum expression for $\mathcal{E}(\rho)$. Since we only care about the behavior of $\mathcal{E}$ on states in C, we can effectively ignore the cases where $d_k = 0$ and assume instead that all the d_k are positive.

By the Polar Decomposition (Theorem B.8 in Section B.3), for each $1 \leqslant k \leqslant N$ there is a unitary $U_k \in \mathcal{L}(\mathcal{H})$ such that

$$F_k P = U_k |F_k P| = U_k \sqrt{P F_k^* F_k P} = \sqrt{d_k} U_k P. \tag{113}$$

U_k rotates C to the subspace C_k that is the image of the projector P_k defined as

$$P_k := U_k P U_k^* = \frac{F_k P U_k^*}{\sqrt{d_k}}. \tag{114}$$

The crucial fact that makes $\mathcal{E}$ recoverable is that these C_k subspaces are mutually orthogonal:

$$P_k P_\ell = P_k^* P_\ell = \frac{U_k P F_k^* F_\ell P U_\ell^*}{\sqrt{d_k d_\ell}} = \frac{U_k (d_k \delta_{k\ell} P) U_\ell^*}{\sqrt{d_k d_\ell}} = 0 \tag{115}$$

if $k \neq \ell$. To help see what's going on, it's worth seeing what happens when $\mathcal{E}$ is applied to some pure state $|\psi\rangle\langle\psi|$ with $|\psi\rangle \in C$. We have $|\psi\rangle = P|\psi\rangle$, and so by Equation (113) we have

$$\mathcal{E}(|\psi\rangle\langle\psi|) = \sum_{k=1}^{N} F_k |\psi\rangle\langle\psi| F_k^* = \sum_k F_k P |\psi\rangle\langle\psi| P F_k^* = \sum_k d_k U_k |\psi\rangle\langle\psi| U_k^* = \sum_k d_k |\psi_k\rangle\langle\psi_k|,$$

where $|\psi_k\rangle := U_k|\psi\rangle \in C_k$ for each k. So $\mathcal{E}(|\psi\rangle\langle\psi|)$ is a mixture of pure states $|\psi_k\rangle$ in the various subspaces C_k. We can thus interpret $\mathcal{E}$ as mapping $|\psi\rangle$ to $|\psi_k\rangle \in C_k$ with probability d_k. Since the C_k are mutually orthogonal, the $|\psi_k\rangle$ are pairwise orthogonal. To recover, we can first measure to which C_k the state $\mathcal{E}(|\psi\rangle\langle\psi|)$ belongs. This projective measurement projects to one of the states $|\psi_k\rangle = U_k|\psi\rangle$, where k is the outcome of the measurement. Then to correct the error, we simply apply U_k^* to get $U_k^*|\psi_k\rangle = |\psi\rangle$.

Now we describe $\mathcal{R}$ formally. $\mathcal{R}$ consists of two stages: (1) measure the error syndrome (*i.e.*, "which C_k?"), and (2) apply the appropriate (unitary) correction U_k^*. By Equation (115), the projectors $P_1, \ldots, P_N$ form a set of orthogonal projectors. If this is not a complete set, *i.e.*, if $\sum_{k=1}^N P_k \neq I$, then we add one more projector $P_{N+1} := I - \sum_{k=1}^N P_k$ to the set to make it complete. Otherwise, we set $P_{N+1} := 0$. The syndrome measurement is then a projective measurement with the P_k. (If the outcome is $N+1$, which signifies "none of the above," then we really don't know what to do, so we'll give up and define $U_{N+1} := I$ for completeness. If the state being measured is the result of applying $\mathcal{E}$ to some state in the code space C, however, then outcome $N+1$ will never actually occur.)

So we define, for any $\sigma \in \mathcal{L}(\mathcal{H})$,

$$\mathcal{R}(\sigma) := \sum_{k=1}^{N+1} U_k^* P_k \sigma P_k U_k.$$

Thus $\mathcal{R}$ has Kraus operators $U_k^* P_k$ for $1 \leqslant k \leqslant N+1$. We first check that $\mathcal{R}$ is compete:

$$\sum_{k=1}^{N+1} P_k U_k U_k^* P_k = \sum_{k=1}^{N+1} P_k = I\,.$$

It remains to check that $\mathcal{R}$ successfully recovers from $\mathcal{E}$ for arbitrary states in C—not just pure states. The following equation will make things easier: for all $1 \leqslant k, \ell \leqslant N$,

$$U_k^* P_k F_\ell P = U_k^* P_k^* F_\ell P = \frac{U_k^* U_k P F_k^* F_\ell P}{\sqrt{d_k}} = \frac{P F_k^* F_\ell P}{\sqrt{d_k}} = \sqrt{d_k}\delta_{k\ell} P\,, \tag{116}$$

using Equations (112) and (114). Also, for $1 \leqslant \ell \leqslant N$, we have $P_{N+1} P_\ell = 0$ by orthogonality, and thus, using Equations (113) and (114),

$$U_{N+1}^* P_{N+1} F_\ell P = P_{N+1} F_\ell P = \sqrt{d_\ell} P_{N+1} U_\ell P = \sqrt{d_\ell} P_{N+1} P_\ell U_\ell = 0\,, \tag{117}$$

and so Equation (116) holds for $k = N+1$ as well.

So finally, if ρ is in C, we have, by Equations (116) and (117),

$$\begin{aligned}
\mathcal{R}(\mathcal{E}(\rho)) = \mathcal{R}(\mathcal{E}(P\rho P)) &= \sum_{k=1}^{N+1}\sum_{\ell=1}^{N} U_k^* P_k F_\ell P \rho P F_\ell^* P_k U_k \\
&= \sum_k \sum_\ell (U_k^* P_k F_\ell P)\, \rho\, (U_k^* P_k F_\ell P)^* \\
&= \sum_k \sum_\ell \left(\sqrt{d_k}\delta_{k\ell} P\right) \rho \left(\sqrt{d_k}\delta_{k\ell} P\right) \\
&= \left(\sum_k \sum_\ell d_k \delta_{k\ell}\right) P\rho P \\
&= (\operatorname{tr} M)\rho\,.
\end{aligned}$$

□

Exercise 27.3 (Challenging) Recall the quantum bit-flip channel for a single qubit:

$$\mathcal{E}(\rho) := (1-p)\rho + pX\rho X.$$

Also recall the recoverable portion of the three-qubit bit-flip channel:

$$\mathcal{E}'(\rho) = (1-p)^3\rho + (1-p)^2 p \sum_{j=1}^{3} X_j \rho X_j.$$

Show directly that $\mathcal{E}'$, with Kraus operators $(1-p)^{3/2} I, (1-p)\sqrt{p}X_1, (1-p)\sqrt{p}X_2, (1-p)\sqrt{p}X_3$, satisfies the peep condition (111) of Theorem 27.2, where C is the usual majority-of-3 code space given by the projector $P = |000\rangle\langle 000| + |111\rangle\langle 111|$. What is the matrix M? What are the P_k and U_k? Is the $\mathcal{R}$ constructed by the Theorem the same as it was before?

Discretization of Errors. The great thing about the Shor code is that it can recover from an *arbitrary* single-qubit error. There are many possible single-qubit errors, as there are a continuum of possible one-qubit Kraus operators. Yet they are all corrected by the Shor code, with no additional work. This happy fact follows from the following two general theorems:

Theorem 27.4 *Suppose $C \subseteq \mathcal{H}$ is the code space for a quantum code, P is the projector projecting orthogonally onto C, $\mathcal{E} \in \mathcal{T}(\mathcal{H})$ is a not necessarily complete quantum error channel with Kraus operators $F_1, \ldots, F_N$, and $\mathcal{R} \in \mathcal{T}(\mathcal{H})$ is a quantum channel with Kraus operators $R_1, \ldots, R_M$ such that, for any $1 \leqslant j \leqslant N$ there exist scalars $d_j \geqslant 0$ such that*

$$R_k F_j P = \sqrt{d_j}\delta_{kj} P \tag{118}$$

for any $1 \leqslant k \leqslant M$. Suppose also that $\mathcal{G}$ is an error channel whose Kraus operators $G_1, \ldots, G_K$ are all linear combinations of $F_1, \ldots, F_N$. Then $\mathcal{R}$ successfully recovers from $\mathcal{G}$.

Proof. For all $1 \leqslant \ell \leqslant K$ we have $G_\ell = \sum_{j=1}^N m_{j\ell} F_j$, for some scalars $m_{j\ell}$. Using (118), we get

$$R_k G_\ell P = \sum_{j=1}^N m_{j\ell} R_k F_j P = \sum_j m_{j\ell} \sqrt{d_j}\delta_{kj} P = m_{k\ell} \sqrt{d_k} P \,,$$

where we set $d_k := 0$ if $N < k \leqslant M$. Then for every state ρ in C, we have

$$\mathcal{R}(\mathcal{G}(\rho)) = \mathcal{R}(\mathcal{G}(P\rho P)) = \sum_{k=1}^M \sum_{\ell=1}^K (R_k G_\ell P)\rho(R_k G_\ell P)^* = \sum_k \sum_\ell |m_{k\ell}|^2 d_k P\rho P = c\rho,$$

where $c := \sum_{k=1}^M \sum_{\ell=1}^K |m_{k\ell}|^2 d_k$. Thus $\mathcal{R}$ successfully recovers from $\mathcal{G}$ given code space C. □

Theorem 27.5 *Suppose $C \subseteq \mathcal{H}$ is the code space for a quantum code, P is the projector projecting orthogonally onto C, $\mathcal{E} \in \mathcal{T}(\mathcal{H})$ is a not necessarily complete quantum error channel with Kraus operators $E_1, \ldots, E_N$ that satisfy the peep condition (111), i.e.,*

$$P E_i^* E_j P = [M]_{ij} P$$

for all $1 \leqslant i, j \leqslant N$, for some matrix M. Suppose also that $\mathcal{G}$ is an error channel whose Kraus operators $G_1, \ldots, G_K$ are all linear combinations of $E_1, \ldots, E_N$. Then the channel $\mathcal{R}$ constructed in the proof of Theorem 27.2 to recover from $\mathcal{E}$ also successfully recovers from $\mathcal{G}$, given code space C.

Proof. In the proof of Theorem 27.2 above, we chose new Kraus operators $F_1, \ldots, F_N$ for $\mathcal{E}$ where $F_k := \sum_{j=1}^N [U]_{jk} E_j$ for all $1 \leqslant k \leqslant N$, where U is an $N \times N$ unitary matrix that diagonalizes M so that there are real numbers $d_1, \ldots, d_N \geqslant 0$ such that $P F_k^* F_\ell P = d_k \delta_{k\ell} P$ for all $1 \leqslant k, \ell \leqslant N$. The F_k are clearly linear combinations of the E_j, but the E_j are also linear combinations of the F_k; indeed, it is easily checked that $E_j = \sum_{k=1}^N [U]_{jk}^* F_k$, using the unitarity of U. Thus the G_ℓ, being linear combinations of the E_j, are linear combinations of the F_k as well.

The $\mathcal{R}$ we constructed in the proof of Theorem 27.2 has Kraus operators

$$U_1^* P_1, \ldots, U_N^* P_N, U_{N+1} P_{N+1}.$$

Setting $R_k := U_k^* P_k$ for all $1 \leqslant k \leqslant N+1$, Equations (116) and (117) say that

$$R_k F_j P = \sqrt{d_j}\delta_{kj} P$$

for all $1 \leqslant k \leqslant N+1$ and all $1 \leqslant j \leqslant N$. This is exactly the discretization condition of Equation (118) (with $M = N+1$). Therefore, $\mathcal{G}$, the R_k, and the F_j together satisfy the hypotheses of Theorem 27.4, and so $\mathcal{R}$ successfully recovers from $\mathcal{G}$ by that theorem. □

We can apply either Theorem 27.4 or Theorem 27.5 to the Shor code to show that Bob's recovery procedure can correct any single-qubit error. The key point is that the four Pauli operators I, X, Y, Z form a basis for the space of all single-qubit operators, and so any single-qubit error channel has Kraus operators that are linear combinations of the Pauli operators. Since Bob can recover from any error of the form X_j, Y_j, or Z_j, for $1 \leqslant j \leqslant 9$ in a way that satisfies Theorem 27.4, he can recover from any linear combination of these—in particular, any error on any one of the nine qubits.

Exercise 27.6 (Challenging) Show that Bob's recovery channel for the Shor code can recover from any error on any one of the nine qubits. [Hint: By the preceding discussion, it only remains to show that Bob's recovery procedure satisfies the discretization condition of Equation (118) for the recoverable portion of the depolarizing channel given by Equation (110).]

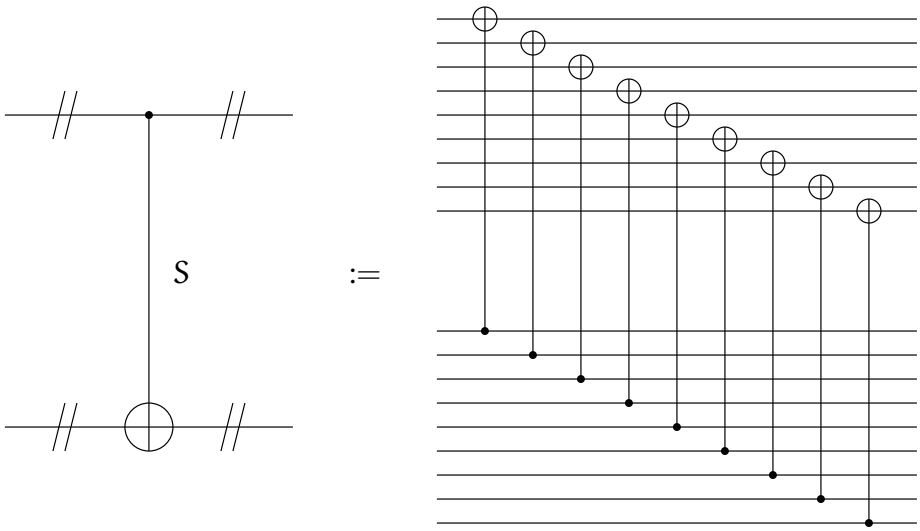


Figure 17: Implementing the C-NOT gate fault-tolerantly using the Shor code. The double slashes on the left indicate that each line represents a multi-qubit register (nine qubits in this case). The circuit maps $|a_S\rangle|b_S\rangle$ to $|a_S\rangle|(a \oplus b)_S\rangle$ for all $a, b \in \{0, 1\}$.

28 Week 14: Fault tolerance

Fault-Tolerant Quantum Computation. If a qubit is in an encoded state, such as with the Shor code, then we can repeatedly apply an error-recovery operation to "restore the logic," *i.e.*, the state of the logical qubit, assuming isolated errors in the physical qubits. Depending on the implementation and frequency of the restore operations, we can maintain a logical qubit state indefinitely with high probability. There is more to a quantum computation, however, than simply maintaining qubits. We must apply quantum gates to them. A not-so-good way to apply a quantum gate is to decode each qubit involved in the gate, then apply the gate on the unencoded qubits, then re-encode the qubits. This is bad because qubits spend time unencoded and subject to unrecoverable errors, defeating the whole purpose of error correction. A better way is to keep all qubits in an encoded state always, never decoding them, so that we prepare, work with, save, and measure qubits in their encoded states only. This practice is called *fault-tolerant quantum computation*, and it works by replacing each gate of a standard, non-fault-tolerant quantum circuit with a quantum mini-circuit that affects the state of the logical qubits in the same way the original gate affects the state of its unencoded qubits.

With the Shor code as well as other quantum error-correcting codes, we can implement several types of quantum gates fault-tolerantly. It can be shown that these codes can implement a family of gates big enough to provide a basis for any feasible quantum computation (a so-called, "universal" family of gates). We will not do an exhaustive treatment here, but will at least show how to implement the C-NOT and Pauli gates explicitly using the Shor code.

Figure 17 shows how to implement the C-NOT gate fault-tolerantly using the Shor code. Each logical qubit is implemented by nine physical qubits.

Exercise 28.1 Verify that the circuit in Figure 17 really implements the C-NOT gate with respect to the Shor code. That is, show that the circuit maps $|a_S\rangle|b_S\rangle$ to $|a_S\rangle|(a \oplus b)_S\rangle$ for all $a, b \in \{0, 1\}$.

29 Week 15: Stabilizers, Entanglement, and Bell inequalities

29.1 Stabilizers

The stabilizer formalism gives us two things: (1) a convenient way of describing some quantum states and some of the gates that act on them; and (2) a large family of quantum error-correcting codes that are efficient and allow easy recovery from errors. The Shor code is an example of a stabilizer code. We will see others.

The Pauli Group. For ease of reference, here we recall the four 1-qubit Pauli operators (matrices are with respect to the computational basis):

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

These operators satisfy $X^2 = Y^2 = Z^2 = I$ and $XY = -YX = iZ$ and $YZ = -ZY = iX$ and $ZX = -XZ = iY$. Also $\operatorname{tr} I = 2$ and $\operatorname{tr} X = \operatorname{tr} Y = \operatorname{tr} Z = 0$. All four Pauli operators are both Hermitean and unitary.

First, some basic definitions. Fix $n \geqslant 1$. The *Pauli group* Π^n on n qubits is the set of all n-qubit unitary operators of the form

$$g = \alpha(\sigma_1 \otimes \sigma_2 \otimes \cdots \otimes \sigma_n),$$

where each σ_i is a Pauli operator (acting on the ith qubit), and the scalar α is in the set $\{1, -1, i, -i\}$. We call α the *coefficient* of g, writing $\operatorname{coeff}(g)$ for α, and we call $\sigma_1 \otimes \cdots \otimes \sigma_n$ the *principal part* of g, writing $\operatorname{princ}(g)$ for $\sigma_1 \otimes \cdots \otimes \sigma_n$. Clearly, both α and all the σ_i are uniquely determined by g. It follows that Π^n has exactly 4^{n+1} many elements.

We call Π^n a *group* because it has an identity element $I^{\otimes n}$ (which we will hereafter abuse notation and simply denote as 1) with respect to multiplication (i.e., composition) and it is closed under multiplication and inverses: if g and h are elements of Π^n, then writing $g = \alpha(\sigma_1 \otimes \cdots \otimes \sigma_n)$ and $h = \beta(\tau_1 \otimes \cdots \otimes \tau_n)$, we have

$$gh = \alpha\beta(\sigma_1\tau_1 \otimes \cdots \otimes \sigma_n\tau_n)$$

and

$$g^{-1} = g^* = \alpha^*(\sigma_1 \otimes \cdots \otimes \sigma_n).$$

Notice that, because of the commutation properties of the Pauli operators, g and h either commute or anticommute, that is, either $gh = hg$ or $gh = -hg$. The latter occurs just when there are an odd number of positions with anticommuting Pauli components in g and h.

A *subgroup* of Π^n is any subset of Π^n which is itself a group. If $g_1, \ldots, g_k$ are elements of Π^n, then the subgroup of Π_n *generated* by $S = \{g_1, \ldots, g_k\}$ is the smallest subgroup of Π_n that includes S. It is the closure of $S \cup \{1\}$ under multiplication and inverses, and we denote it $\langle S \rangle$ or $\langle g_1, \ldots, g_k \rangle$.[29] We say that S is a *minimal* set of generators if no proper subset of S generates the same subgroup.

[29] The angle bracket notation $\langle \cdots \rangle$ we use here has nothing to do with the Hermitean inner product on a Hilbert space. We have no occasion to use the latter meaning anywhere in this section.

Here is an example. Let $n := 3$ and let $g := -i(X \otimes Z \otimes Y) = -iX_1Z_2Y_3$ and $h := -(Y \otimes I \otimes Z) = -Y_1Z_3$. Then

$$\begin{aligned}
\text{coeff}(g) &= -i\,, \\
\text{princ}(g) &= X \otimes Z \otimes Y = X_1Z_2Y_3\,, \\
\text{coeff}(h) &= -1\,, \\
\text{princ}(h) &= Y \otimes I \otimes Z = Y_1Z_3\,, \\
gh &= -i(Z \otimes Z \otimes X) = -iZ_1Z_2X_3\,, \\
hg &= -i(Z \otimes Z \otimes X) = -iZ_1Z_2X_3 = gh\,, \\
g^2 &= -1\,, \\
h^2 &= 1\,.
\end{aligned}$$

Exercise 29.1 Redo the example above, this time where $n = 4$ and $g = I \otimes Z \otimes Z \otimes X$ and $h = i(Y \otimes Y \otimes X \otimes Y)$.

Stabilizing Subgroups. There are subgroups of Π^n that are of particular interest to us.

Definition 29.2 Let G be a subgroup of Π^n. We will say that G is a *stabilizing subgroup* iff $-1 \notin G$.

This simple condition has strong consequences.

Lemma 29.3 *Let G be a stabilizing subgroup of Π^n. Then*

1. $\text{coeff}(g) = \pm 1$ *(and hence g is Hermitean) for all* $g \in G$,
2. $g^2 = 1$ *for all* $g \in G$,
3. *no two distinct elements of* G *share the same principal part, and*
4. $gh = hg$ *for all* $g, h \in G$.[30]

Exercise 29.4 Prove Lemma 29.3. [Hint: The last item follows easily from the second item and the fact that any two elements of Π^n either commute or anticommute.]

Exercise 29.5 List the elements of $\langle g_1, g_2 \rangle$, where $g_1 = X \otimes Z$ and $g_2 = Z \otimes X$. Is $\langle g_1, g_2 \rangle$ a stabilizing subgroup of Π^2? Explain.

Exercise 29.6 Let G be a stabilizing subgroup of Π^n with a k-element minimal generating set $S = \{g_1, \ldots, g_k\}$. Show that G has exactly 2^k many elements, each one obtained by multiplying together the elements of a different subset of S. That is,

$$G = \left\{ \prod_{g \in J} g \;:\; J \subseteq S \right\}\,, \tag{119}$$

and each choice of J yields a distinct product. [Hint: Use items 2 and 4 of Lemma 29.3.]

[30] A group with this property is called *commutative* or *abelian*.

The previous exercise shows that any minimal generating set for G has size k. We will call k the *dimension* of G.

Stabilizing Subgroups Acting on $\mathcal{H}$. Fix $n \geqslant 1$ as before, and let $\mathcal{H} = \mathbb{C}^{2^n}$ be the n-qubit Hilbert space with the usual computational basis. Notice that each element of Π^n is an operator in $\mathcal{L}(\mathcal{H})$. Let $E \subseteq \mathcal{H}$ be any nontrivial (i.e., positive-dimensional) subspace of $\mathcal{H}$. The *stabilizer* of E in Π^n, written $\mathrm{Stab}(E)$, is the set of all $g \in \Pi^n$ that fix E pointwise, that is, the set of all $g \in \Pi^n$ such that $gv = v$ for all $v \in E$. $\mathrm{Stab}(E)$ is a subgroup of Π^n, and in fact it must be a stabilizing subgroup, because $-1 \in \Pi^n$ maps any v to $-v$, and so does not fix anything except the zero vector.

Conversely, given a stabilizing subgroup G of Π^n, we let E_G be the subspace of $\mathcal{H}$ stabilized by G, that is,

$$E_G := \{v \in \mathcal{H} : (\forall g \in G)[\, gv = v \,]\} .$$

(If G is not stabilizing, then E_G is evidently the trivial space $\{0\}$.) Clearly, $G \subseteq \mathrm{Stab}(E_G)$. We will see shortly (Proposition 29.14, below) that the two groups are the same.

We can recast all this in terms of eigenvectors, eigenvalues, and projectors. If $g \in \Pi^n$ and $\mathrm{coeff}(g) \in \{1, -1\}$, then $g^2 = 1$. This means that the only two eigenvalues of g are $+1$ and -1. Obviously, the identity element $1 \in \Pi^n$ only has eigenvalues $+1$. If $g \neq \pm 1$, however, then g has both eigenvalues, and in fact, $\dim \mathcal{E}_{+1}(g) = \dim \mathcal{E}_{-1}(g) = 2^{n-1}$. This is because first of all, $\mathrm{princ}(g)$ must have a Pauli operator $\sigma \neq I$ in at least one position, and since $\mathrm{tr}\,\sigma = 0$ we then have $\mathrm{tr}\, g = 0$ (see the last item of Proposition 10.1). Secondly, $\mathrm{tr}\, g$ is the sum of g's eigenvalues with multiplicity. Thus $+1$ and -1 occur with the same multiplicity, which is then the common dimension of the two eigenspaces of g. The projectors onto these two eigenspaces are seen to be

$$P_g^{\pm} := \frac{1 \pm g}{2} \tag{120}$$

(here as well, $1 \in \Pi^n$ is the n-qubit identity operator).

Notice, by the way, that these projectors are sums over $\langle g \rangle = \{1, g\}$ and $\langle -g \rangle = \{1, -g\}$, which are both two-element (stabilizing) subgroups of Π^n. More generally:

Lemma 29.7 *Let G be a k-dimensional stabilizing subgroup of Π^n, and let $\{g_1, \ldots, g_k\}$ be a minimal generating set for G. Then $\dim(E_G) = 2^{n-k}$, and the operator*

$$P_G := 2^{-k} \sum_{g \in G} g \tag{121}$$

orthogonally projects onto E_G.

Proof. We set $P := P_G$ for this proof. Note that for any $g \in G$,

$$gP = g\left(2^{-k} \sum_{h \in G} h\right) = 2^{-k} \sum_{h \in G} gh = 2^{-k} \sum_{h \in G} h = P\,, \tag{122}$$

the third equality following from the fact that for fixed $g \in G$, gh runs through the elements of G as h runs through the elements of G.

P is Hermitean because all elements of G are Hermitean. Next, we check that P projects onto E_G by fixing every vector in E_G and mapping every vector in $\mathcal{H}$ into E_G (whence it follows that $P^2 = P$). For any $v \in E_G$, we have

$$Pv = 2^{-k} \sum_{g \in G} gv = 2^{-k} \sum_{g \in G} v = v \, .$$

Now let $u \in \mathcal{H}$ be arbitrary. For any $g \in G$, we have $gPu = Pu$, and so $Pu \in E_G$.

We have established that P projects onto E_G. What is $\dim(E_G)$ then? The dimension of E_G is equal to the trace of P:

$$\dim(E_G) = \operatorname{tr} P = 2^{-k} \sum_{g \in G} \operatorname{tr} g = 2^{-k} \operatorname{tr} 1 \, ,$$

because all the other elements of G besides 1 have zero trace. We have $\operatorname{tr} 1 = \operatorname{tr}(I^{\otimes n}) = 2^n$, and so finally,

$$\dim(E_G) = \operatorname{tr} P = 2^{n-k} \, .$$

□

The next exercise shows that we can also characterize E_G directly in terms of a generating set for G, using the P_g^+ projectors.

Exercise 29.8 Let G be a k-dimensional stabilizing subgroup of Π^n, and let $\{g_1, \ldots, g_k\}$ be a minimal generating set for G, as in the previous lemma. Show that $P_G = P_{g_1}^+ P_{g_2}^+ \cdots P_{g_k}^+$, where the $P_{g_i}^+$ are defined by Equation (120). [Hint: Expand the right-hand side and use Exercise 29.6.]

Let's look at some more examples. Suppose $n = 4$ and $G = \langle Z_1, Z_2, Z_3, Z_4 \rangle$. This is a minimal generating set for G with four elements, so $\dim(E_G) = 2^{4-4} = 1$. (Oh, and G is stabilizing.) What is a vector in E_G? We have $Z|0\rangle = |0\rangle$, so $|0000\rangle \in E_G$. Thus E_G is the 1-dimensional subspace spanned by $|0000\rangle$. Now suppose $G = \langle Z_1, Z_2, -Z_3, Z_4 \rangle$. Then you should check that E_G is spanned by $|0010\rangle$. Can you generalize these observations?

Exercise 29.9 For any $b_1, \ldots, b_n \in \{0, 1\}$, let $G := \langle (-1)^{b_1} Z_1, (-1)^{b_2} Z_2, \ldots, (-1)^{b_n} Z_n \rangle$. Show that E_G is the 1-dimensional subspace of $\mathcal{H}$ spanned by $|b_1 \cdots b_n\rangle$.

Now suppose $G = \langle X_1, X_2, X_3, X_4 \rangle$. Since $X|+\rangle = |+\rangle$, we get that E_G is the 1-dimensional space spanned by $|+\rangle^{\otimes 4}$.

Exercise 29.10 Suppose $G = \langle X_1, X_2, -X_3, X_4 \rangle$. Make a guess about E_G and verify that your guess is correct.

Exercise 29.11 In each case, find a 4-qubit vector that spans E_G.

- $G = \langle Z_1, Z_2, X_3, X_4 \rangle$.
- $G = \langle Y_1, Y_2, Y_3, Y_4 \rangle$.

- $G = \langle -Z_1, Z_1Z_2, -Z_1Z_2Z_3, Z_1Z_2Z_3Z_4 \rangle$.

Exercise 29.12 The last exercise suggests that the minimal generating set of a stabilizing G is not unique (although they all have the same size). Find an alternate minimum generating set for the group $\langle -Z_1, Z_1Z_2, -Z_1Z_2Z_3, Z_1Z_2Z_3Z_4 \rangle$ of the last exercise.

Corollary 29.13 *If G is a stabilizing subgroup of Π^n, then $\dim(E_G) > 0$ and G has dimension at most n.*

Now we have the following:

Proposition 29.14 *For any stabilizing subgroup G of Π^n,*

$$G = \mathrm{Stab}(E_G) .$$

Proof. We noticed before that $G \subseteq \mathrm{Stab}(E_G)$. For the reverse inclusion, let $H := \mathrm{Stab}(E_G)$. Then H is stabilizing and $E_G \subseteq E_H$ (since H fixes all of E_G at least). Thus $\dim(E_H) \geqslant \dim(E_G)$. Let k and ℓ be the dimensions of G and H, respectively. We have $\dim(E_G) = 2^{n-k}$ and $\dim(E_H) = 2^{n-\ell}$, and hence $\ell \leqslant k$. But then G has cardinality 2^k and H has cardinality $2^\ell \leqslant 2^k$, so we must have $k = \ell$ and $G = H$. □

Not every subspace of $\mathcal{H}$ is of the form E_G for some stabilizing G. There are infinitely many subspaces of $\mathcal{H}$ but only finitely many stabilizing subgroups of Π^n. For almost all subspaces E, we have $\mathrm{Stab}(E) = \{1\}$, but $E_{\{1\}} = \mathcal{H}$. The spaces of the form E_G are particularly nice. For one thing, we will use them as the code spaces for stabilizer error-correcting codes (see the topic, Stabilizer Codes, below).

Given a stabilizing subgroup $G \subseteq \Pi^n$ and some $h \in \Pi^n$ that anticommutes with at least one element of G, we end this topic with some results about how h "splits G in half."

Lemma 29.15 *Let G be a k-dimensional stabilizing subgroup of Π^n and let $h \in \Pi^n$ be arbitrary. Let $C := \{g \in G \mid gh = hg\}$. Then C is a subgroup of G, and if $C \neq G$, then C has exactly 2^{k-1} elements (that is, exactly half the elements of G).*

Proof. It is routine to check that C is a subgroup of G. Assuming $C \neq G$, there exists some $a \in G$ that anticommutes with h. Obviously, $a \neq 1$. Consider the map $a_L : G \to G$ that takes any $g \in G$ to ag. The map a_L is a bijection that satisfies $a_L(g) \neq g$ and $a_L(a_L(g)) = a^2 g = g$ for all $g \in G$. Thus we can partition the elements of G into disjoint pairs $\{g, a_L(g)\} = \{g, ag\}$ of distinct elements of G. Notice further, for any $g \in G$, that h commutes with g if and only if h anticommutes with ag. Thus exactly one element of each pair is in C. It follows that C has exactly 2^{k-1} elements. □

The next corollary will be useful when we discuss stabilizer codes for error correction.

Corollary 29.16 *Let G, k, and h be as in Lemma 29.15, and assume that h anticommutes with at least one element of G. Then $P_G h P_G = 0$.*

Proof. Let $C := \{g \in G \mid gh = hg\}$ as in the lemma, which implies that, since $C \neq G$ by assumption, C and $G - C$ have the same number of elements. Setting $P := P_G$ and using Equation (122), we compute

$$\begin{aligned} PhP &= 2^{-k} \sum_{g \in G} ghP = 2^{-k} \left(\sum_{g \in C} ghP + \sum_{g \in G-C} ghP \right) = 2^{-k} \left(\sum_{g \in C} hgP - \sum_{g \in G-C} hgP \right) \\ &= 2^{-k} h \left(\sum_{g \in C} gP - \sum_{g \in G-C} gP \right) = 2^{-k} h \left(\sum_{g \in C} P - \sum_{g \in G-C} P \right) = 0 \,. \end{aligned}$$

□

The next lemma will be used to prove the Gottesman-Knill theorem, below.

Lemma 29.17 *Let* G, k, h, *and* C *be as in Lemma 29.15, and suppose that* $\mathrm{coeff}(h) = \pm 1$ *and neither* h *nor* $-h$ *is in* G. *Let* $hC := \{hg \mid g \in C\}$. *Then* $C \cup hC$ *is a stabilizing subgroup of* Π^n *whose dimension is* $k+1$ *if* $C = G$ *and* k *otherwise.*

Proof. Let $H := C \cup hC$. It is routine to check that H is a subgroup of Π^n (using the fact that $h^2 = 1$) and that C and hC have the same number of elements. Furthermore, $C \cap hC = \emptyset$, for otherwise there are $g_1, g_2 \in C$ such that $g_1 = hg_2$, but then $h = g_1 g_2 \in G$; contradiction.[31] We conclude that H is twice as big as C. If $C = G$, then H has 2^{k+1} many elements. Otherwise, C has 2^{k-1} many elements by Lemma 29.15, whence H has exactly 2^k many elements.

Finally, we show that H is stabilizing, and thus has the appropriate dimension given its size. We already have $-1 \notin C$, since G is stabilizing, but we cannot have $hg = -1$ for any $g \in C$ either; otherwise, multiplying both sides by h gives $-h = g \in G$; contradiction. Therefore, $-1 \notin H$. □

Connection to Linear Algebra Over $\mathbb{Z}_2$. There is an illuminating way of describing the principal part of an element $g \in \Pi^n$ as a $2n$-dimensional row vector over the 2-element field $\mathbb{Z}_2$. We define two maps $\varphi_x, \varphi_z : \Pi_n \to \mathbb{Z}_2^n$ as follows: If $g = \alpha(\sigma_1 \otimes \cdots \sigma_n)$, where $\alpha \in \{\pm 1, \pm i\}$ and $\sigma_i \in \{I, X, Y, Z\}$ for $1 \leqslant i \leqslant n$, then define

$$\begin{aligned} \varphi_x(g) &:= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \in \mathbb{Z}_2^n \,, \\ \varphi_z(g) &:= \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} \in \mathbb{Z}_2^n \,, \end{aligned}$$

where for all $1 \leqslant i \leqslant n$,

$$\begin{aligned} x_i &:= \begin{cases} 1 & \text{if } \sigma_i \in \{X, Y\}, \\ 0 & \text{if } \sigma_i \in \{I, Z\}, \end{cases} \\ z_i &:= \begin{cases} 1 & \text{if } \sigma_i \in \{Z, Y\}, \\ 0 & \text{if } \sigma_i \in \{I, X\}. \end{cases} \end{aligned}$$

Observe that $\varphi_x(g)$ and $\varphi_z(g)$ together uniquely determine $\mathrm{princ}(g)$ but ignore $\mathrm{coeff}(g)$ completely. Most importantly, one should verify that for any $g, h \in \Pi^n$,

$$\begin{aligned} \varphi_x(gh) &= \varphi_x(g) + \varphi_x(h) \,, \\ \varphi_z(gh) &= \varphi_z(g) + \varphi_z(h) \,, \end{aligned}$$

[31] This is a basic result of group theory—cosets of a finite subgroup are all the same size and pairwise disjoint.

where the right-hand operations are both vector addition modulo 2.[32] Now define

$$\varphi(g) := \left[\begin{array}{c|c} \varphi_x(g) & \varphi_z(g) \end{array}\right],$$

the $2n$-dimensional row vector obtained by concatenating $\varphi_x(g)$ with $\varphi_z(g)$.[33] For any $g, h \in \Pi_n$, we have $\varphi(g) = \varphi(h)$ if and only if $\mathrm{princ}(g) = \mathrm{princ}(h)$, and

$$\varphi(gh) = \varphi(g) + \varphi(h)$$

as well.

If G is a stabilizing subgroup of Π^n, then φ is one-to-one when restricted to G (this follows from Lemma 29.3(3)). For any $g_1, \ldots, g_k \in G$, we can form the $k \times 2n$ matrix

$$M := \left[\begin{array}{c} \varphi(g_1) \\ \varphi(g_2) \\ \vdots \\ \varphi(g_k) \end{array}\right] = \left[\begin{array}{c|c} \varphi_x(g_1) & \varphi_z(g_1) \\ \varphi_x(g_2) & \varphi_z(g_2) \\ \vdots & \vdots \\ \varphi_x(g_k) & \varphi_z(g_k) \end{array}\right]$$

whose rows are the vectors $\varphi(g_i)$ for $1 \leqslant i \leqslant k$. Now assume $G = \langle g_1, \ldots, g_k \rangle$. Then the vectors $\varphi(g)$ for $g \in G$ are exactly the linear combinations (over $\mathbb{Z}_2$) of the rows of M. Furthermore, $\{g_1, \ldots, g_k\}$ is a minimal generating set if and only if the rows of M are linearly independent over $\mathbb{Z}_2$ (see Lemma 29.21 below). More generally, the dimension of G is equal to the rank of M.

The map φ can also easily tell us whether two given elements $g, h \in \Pi^n$ commute or anticommute. We define the following inner product[34] of the row vectors $\varphi(g)$ and $\varphi(h)$ as

$$\varphi(g) \cdot \varphi(h) := \varphi_x(g)(\varphi_z(h))^T + \varphi_z(g)(\varphi_x(h))^T \in \mathbb{Z}_2,$$

where the right-hand side operations are in $\mathbb{Z}_2$.

Exercise 29.18 For the g and h of Exercise 29.1, give $\varphi(g)$ and $\varphi(h)$ as well as $\varphi(g) \cdot \varphi(h)$. Do the same for the g and h of the example immediately preceding Exercise 29.1.

We have the following:

Lemma 29.19 *For every* $g, h \in \Pi^n$,

$$\varphi(g) \cdot \varphi(h) = \begin{cases} 1 & \text{if } gh = -hg, \\ 0 & \text{if } gh = hg. \end{cases}$$

Exercise 29.20 Prove this lemma.

The next lemma gives a linear algebraic way to determine if given elements of Π^n form a minimal generating set for a stabilizing subgroup. The linear algebra is over $\mathbb{Z}_2$.

[32] Any map that preserves operations in this way is known as a *group homomorphism*.

[33] This vector is sometimes written as $\varphi_x(g) \oplus \varphi_z(g)$, but we avoid that notation here as it may be confusing.

[34] This is an example of what is known as a *symplectic inner product*.

Lemma 29.21 *Let $g_1, \ldots, g_k \in \Pi^n$ be arbitrary. Then $\{g_1, \ldots, g_k\}$ is a minimal generating set for a stabilizing subgroup of Π^n if and only if*

1. $\text{coeff}(g_i) = \pm 1$ *for all* $1 \leqslant i \leqslant k$,
2. $\varphi(g_i) \cdot \varphi(g_j) = 0$ *(equivalently,* $g_i g_j = g_j g_i$ *by Lemma 29.19) for all* $1 \leqslant i, j \leqslant k$, *and*
3. *the vectors* $\varphi(g_1), \ldots, \varphi(g_k)$ *are linearly independent.*

Proof. The forward direction comes immediately from Lemma 29.3, except for the third item, which follows from Exercise 29.6: any linear dependence of the $\varphi(g_i)$ would correspond to a nonempty product of the g_i equalling 1, contradicting the minimality of the generating set.

For the reverse direction, assume all three conditions hold. Let $G := \langle g_1, \ldots, g_k \rangle$. To show that G is stabilizing, suppose that $-1 \in G$. By commutativity (the second condition) and the fact that $g_i^2 = 1$ for all $1 \leqslant i \leqslant k$ (the first condition), we must be able to write

$$-1 = g_1^{e_1} \cdots g_k^{e_k} ,$$

for some $e_1, \ldots, e_k \in \{0, 1\}$. But then,

$$0 = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix} = \varphi(-1) = \varphi(g_1^{e_1} \cdots g_k^{e_k}) = e_1 \varphi(g_1) + \cdots + e_k \varphi(g_k) .$$

So by the linear independence of $\{\varphi(g_1), \ldots, \varphi(g_k)\}$ (the third condition), we must have $e_1 = \cdots = e_k = 0$; but then, $g_1^0 \cdots g_k^0 = 1 \neq -1$. Contradiction. Thus G is stabilizing.

Finally, if $\{g_1, \ldots, g_k\}$ were not minimal, then we could write some g_j as a product of the other g_i, but then $\varphi(g_j)$ is a linear combination of the other $\varphi(g_i)$, contradicting linear independence. □

Stabilizer Circuits and the Gottesman-Knill Theorem. Our first application of stabilizers is to show the *Gottesman-Knill theorem*, which says that quantum circuits employing only H, S, and C-NOT gates can be simulated efficiently (i.e., in polynomial time) on a classical computer. We call these circuits *stabilizer circuits*. Initial states must be computational basis states, and all measurements are computational basis measurements.[35]

As before, we let $\mathcal{H} \cong \mathbb{C}^{2^n}$ be the n-qubit Hilbert space with the usual computational basis. The whole idea is to keep track of the quantum state $\rho = |\psi\rangle\langle\psi|$ inside an n-qubit circuit, not as a superposition of basis states as we have been doing, but rather as a set of generators of an n-dimensional stabilizing subgroup of Π^n that stabilizes $|\psi\rangle$. When some gate U is applied, mapping the state $|\psi\rangle$ to state $|\psi'\rangle = U|\psi\rangle$, we can easily update our generating set to that of a new subgroup that stabilizes $|\psi'\rangle$.

The gates H (Hadamard gate) and S (phase gate) applied to any qubit and the C-NOT gate applied to any pair of qubits have a special property that makes the above possible: they *normalize* Π^n. A unitary operator $U \in \mathcal{L}(\mathcal{H})$ is said to *normalize*[36] Π^n iff, for any unitary operator $g \in \mathcal{L}(\mathcal{H})$, g is in Π^n if and only if $UgU^* \in \Pi^n$. The unitary operators that normalize Π^n themselves form a group, called the n*-qubit Clifford group* $\mathcal{C}_n$. One can show that $\mathcal{C}_n$ is generated (up to an arbitrary global phase) by the three types of operators mentioned above: H, S, and C-NOT, which are sometimes called *Clifford gates*.

[35] Improvements and generalizations to this theorem were made in a subsequent paper by Aaronson and Gottesman.

[36] This term comes from group theory and has nothing to do with making a vector have unit length.

Exercise 29.22 The n-qubit Pauli group is clearly a subgroup of the n-qubit Clifford group, so we can allow Pauli gates in a stabilizer circuit "for free."

1. For all nine combinations of $g, U \in \{X, Y, Z\}$, give UgU^* ($= UgU$).
2. Show how the three Pauli operators X, Y, and Z (not necessarily in that order) can be written as products of H and S gates only. [Hint: It may help to picture how these gates rotate the Bloch sphere.]

We describe the classical simulation of a stabilizer circuit in three steps: (1) representing the initial quantum state before the circuit is applied; (2) showing how to update this representation as each Clifford gate of the circuit is applied; and (3) computing outcome probabilities and the post-measurement state of a 1-qubit measurement in the computational basis. We will take these in order, but first recall the projector $P = P_G$ of Equation (121) for an arbitrary stabilizing subgroup G. If G has dimension n, then P_G projects onto a subspace of dimension $2^{n-n} = 1$, in which case, P_G is a pure state that we can represent by a minimal generating set $\{g_1, \ldots, g_n\}$ of G. This is how we will represent states as the computation proceeds.

We assume the initial state being fed to the circuit is some computational basis state $|\varphi_0\rangle := |b_1 b_2 \cdots b_n\rangle$, where each $b_j \in \{0, 1\}$. In Exercise 29.9, you effectively showed that $|\varphi_0\rangle\langle\varphi_0| = P_G$ where $G = \langle (-1)^{b_1} Z_1, (-1)^{b_2} Z_2, \ldots, (-1)^{b_n} Z_n \rangle$. So this is our representation of the initial basis state of the circuit.

Now suppose that at some stage in the circuit's application the state is $\rho = P_G$ for some stabilizing $G = \langle g_1, \ldots, g_n \rangle$ just before some Clifford gate U is applied. We claim that immediately after U is applied, the new state $\rho' = U\rho U^*$ is equal to $P_{G'}$, where

$$G' := UGU^* = \langle Ug_1U^*, \ldots, Ug_nU^* \rangle .$$

To see the claim, first note that each Ug_jU^* is in Π^n for $1 \leqslant j \leqslant n$, because U is a Clifford gate. Next, notice that when multiplying terms of the form Ug_jU^* together, the inner U's cancel, e.g., $(Ug_1U^*)(Ug_2U^*) = Ug_1g_2U^*$. This fact helps one to see that $G' = \langle Ug_1U^*, \ldots, Ug_nU^* \rangle$. Next, we can see that G' must be a stabilizing subgroup of Π^n, for otherwise, $-1 \in G'$, and it follows that

$$-1 = -U^*U = U^*(-1)U \in U^*G'U = U^*(UGU^*)U = G ,$$

contradicting the fact that G is stabilizing. Now G' evidently has the same number of elements as G, which is 2^n, and so G' has dimension n. Finally, let $|\psi\rangle$ be any unit vector in E_G (and thus $\rho = |\psi\rangle\langle\psi|$) and let g be any element of G. Then

$$UgU^*(U|\psi\rangle) = Ug|\psi\rangle = U|\psi\rangle .$$

That is, $U|\psi\rangle$ is fixed by UgU^*. Since any element $g' \in G'$ can be written in this form, we get that $U|\psi\rangle \in E_{G'}$. Thus $\rho' = U|\psi\rangle\langle\psi|^*U$ projects onto $E_{G'}$ and so is equal to $P_{G'}$.

Having established the claim, it remains to compute Ug_jU^* given g_j, for $1 \leqslant j \leqslant n$. This is easy, given the limited choices for U. For example, if $U = H_1$ and $g_j = Z \otimes \cdots$, then

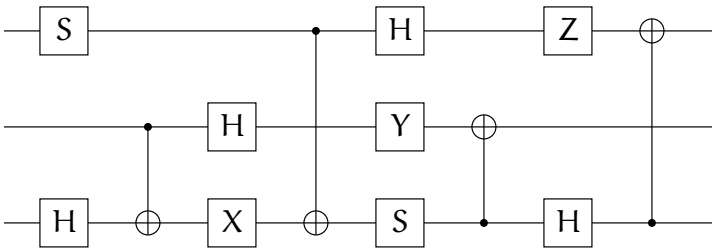
$$UgU^* = HZH \otimes \cdots = X \otimes \cdots ,$$

where the omitted Pauli operators remain unchanged. This example generalizes to H, S, and C-NOT acting on any of the qubits. The table below gives the results of the three Clifford gates U conjugating Pauli gates σ. Here, $1 \leqslant i, j \leqslant n$ and $i \neq j$, and we only include the cases where $U\sigma U^* \neq \sigma$. We could have omitted the Y-gates from the second column of the table because $Y = iXZ$, and so conjugating Y is the same as conjugating Z followed by conjugating X and inserting the global phase shift $i = e^{i\pi/2}$ (that is, $UYU^* = i(UXU^*)(UZU^*)$). Recall that C-NOT$_{i,j}$ has qubit i as the control and qubit j as the target.

U	σ	$U\sigma U^*$
H_i	X_i	Z_i
	Y_i	$-Y_i$
	Z_i	X_i
S_i	X_i	Y_i
	Y_i	$-X_i$
C-NOT$_{i,j}$	X_i	X_iX_j
	Y_i	Y_iX_j
	Y_j	Z_iY_j
	Z_j	Z_iZ_j

Exercise 29.23 Extend the table above to include entries for $U = X_i$, $U = Y_i$, and $U = Z_i$.

Exercise 29.24 Consider the somewhat randomly chosen stabilizer circuit below:



Assuming the initial state is $|110\rangle\langle 110|$, give a set of three generators for the group stabilizing the state after each C-NOT gate is applied. The initial state is stabilized by the generators

$$\begin{aligned} g_1 &= -Z \otimes I \otimes I = -Z_1 \\ g_2 &= -I \otimes Z \otimes I = -Z_2 \\ g_3 &= I \otimes I \otimes Z = Z_3 \end{aligned}$$

Give the other sets of generators in the same format, but you can omit all the $\otimes$'s.

Finally, we consider single-qubit computational basis measurements. Given a state $\rho = P_G$ where $G = \langle g_1, \ldots, g_n \rangle$ as before, we will assume for convenience that the first qubit is measured (measurements on other qubits are handled similarly). We have to determine the probabilities of the two possible outcomes 0 and 1, as well as the post-measurement states for each, represented as stabilizing subgroups G_0 and G_1 of Π^n, respectively. The projectors for the two outcomes are

$P^+_{Z_1} = (1+Z_1)/2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes I \otimes \cdots \otimes I$ for outcome 0 and $P^-_{Z_1} = (I-Z_1)/2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes I \otimes \cdots \otimes I$ for outcome 1 (see Equation (120)). The outcome probabilities are

$$\begin{aligned} \Pr[0] &= \left\langle P^+_{Z_1}, \rho \right\rangle = \mathrm{tr}\left(P^+_{Z_1}\rho\right) = \mathrm{tr}\left(P^+_{Z_1}\rho P^+_{Z_1}\right), \\ \Pr[1] &= \left\langle P^-_{Z_1}, \rho \right\rangle = \mathrm{tr}\left(P^-_{Z_1}\rho\right) = \mathrm{tr}\left(P^-_{Z_1}\rho P^-_{Z_1}\right), \end{aligned}$$

with respective post-measurement states

$$\begin{aligned} \rho_0 &:= \Pr[0]^{-1} P^+_{Z_1}\rho P^+_{Z_1}, \\ \rho_1 &:= \Pr[1]^{-1} P^-_{Z_1}\rho P^-_{Z_1}. \end{aligned}$$

To compute $P^\pm_{Z_1}\rho P^\pm_{Z_1}$, first note that for $g \in G$, if $gZ_1 = Z_1 g$, then $gP^\pm_{Z_1} = P^\pm_{Z_1} g$, but if $gZ_1 = -Z_1 g$, then $gP^\pm_{Z_1} = P^\mp_{Z_1} g$. Also note that $P^\pm_{Z_1} P^\mp_{Z_1} = 0$. From this you can perhaps see how this is going to go—we split $\rho = P_G$ into those terms that commute with Z_1 and those that anticommute with Z_1; the latter terms disappear:

$$\begin{aligned} P^\pm_{Z_1}\rho P^\pm_{Z_1} &= 2^{-n} \sum_{g\in G} P^\pm_{Z_1} g P^\pm_{Z_1} && (123) \\ &= 2^{-n} \left(\sum_{g\in G: gZ_1 = Z_1 g} P^\pm_{Z_1} g P^\pm_{Z_1} + \sum_{g\in G: gZ_1 = -Z_1 g} P^\pm_{Z_1} g P^\pm_{Z_1} \right) && (124) \\ &= 2^{-n} \left(\sum_{g\in G: gZ_1 = Z_1 g} P^\pm_{Z_1} P^\pm_{Z_1} g + \sum_{g\in G: gZ_1 = -Z_1 g} P^\pm_{Z_1} P^\mp_{Z_1} g \right) && (125) \\ &= 2^{-n} \sum_{g\in G: gZ_1 = Z_1 g} P^\pm_{Z_1} g && (126) \\ &= 2^{-n-1} \sum_{g\in G: gZ_1 = Z_1 g} (1 \pm Z_1) g = 2^{-n-1} \left(\sum_g g \pm \sum_g Z_1 g \right). && (127) \end{aligned}$$

Now we have to consider three cases: (1) Z_1 is in G; (2) $-Z_1$ is in G; and (3) neither Z_1 nor $-Z_1$ is in G. These cases are mutually exclusive, because G is stabilizing.

Case 1: $Z_1 \in G$**.** Then Z_1 commutes with all elements of G, making both sums $\sum_g g$ and $\sum_g Z_1 g$ sum over all elements of G and thus be equal. It follows that

$$\begin{aligned} P^+_{Z_1}\rho P^+_{Z_1} &= 2^{-n} \sum_{g\in G} g = \rho, \\ \Pr[0] &= \mathrm{tr}\,\rho = 1, \\ P^-_{Z_1}\rho P^-_{Z_1} &= 0, \\ \Pr[1] &= \mathrm{tr}\,0 = 0. \end{aligned}$$

To summarize, the outcome is 0 with certainty and the post-measurement state is unchanged: $\rho_0 = \rho$. (ρ_1 is undefined, because you cannot divide by 0.) We also have $G_0 = G$, so there is no need to change the stabilizing group representing the post-measurement state.

Case 2: $-Z_1 \in G$**.** This is similar to the previous case (particularly, Z_1 commutes with all of G) except that now, $\sum_{g\in G} g = \sum_{g\in G}(-Z_1)g = -\sum_{g\in G} Z_1 g$. This gives

$$\begin{aligned} P^+_{Z_1}\rho P^+_{Z_1} &= 0\,, \\ \Pr[0] &= \operatorname{tr} 0 = 0\,, \\ P^-_{Z_1}\rho P^-_{Z_1} &= 2^{-n}\sum_{g\in G} g = \rho\,, \\ \Pr[1] &= \operatorname{tr}\rho = 1\,. \end{aligned}$$

To summarize, the outcome is 1 with certainty and the post-measurement state is unchanged: $\rho_1 = \rho$. (ρ_0 is undefined.) Again, we have $G_1 = G$, so there is no need to change the stabilizing group for the post-measurement state.

Case 3: neither Z_1 **nor** $-Z_1$ **is in** G**.** This is the hardest of the three cases to analyze. The probability of outcome 0 is

$$\Pr[0] = \operatorname{tr}\left(P^+_{Z_1}\rho P^+_{Z_1}\right) = 2^{-n-1}\left(\sum_g \operatorname{tr} g + \sum_g \operatorname{tr}(Z_1 g)\right)\,,$$

where both sums are over those $g \in G$ that commute with Z_1. We now show that the second sum disappears entirely: for all $g \in G$ we must have $\operatorname{princ}(Z_1 g) \neq 1$ (for otherwise, $\operatorname{princ}(Z_1) = \operatorname{princ}(Z_1 gg) = \operatorname{princ}(Z_1 g)\operatorname{princ}(g) = \operatorname{princ}(g)$, and so $Z_1 = \pm g$, contradicting the fact that neither Z_1 nor $-Z_1$ is in G); thus $\operatorname{tr}(Z_1 g) = 0$ for all $g \in G$. The only term that survives in the first sum is $g = 1$; all others have zero trace. Thus

$$\Pr[0] = 2^{-n-1}\operatorname{tr} 1 = 2^{-n-1}\operatorname{tr} I^{\otimes n} = 2^{-n-1}2^n = \frac{1}{2} = \Pr[1]\,.$$

Thus outcomes 0 and 1 occur with equal odds. For the post-measurement states, we will see that G_0 and G_1 differ from each other and from G. In fact, $Z_1 \in G_0$ and $-Z_1 \in G_1$. A full analysis will use Lemma 29.25, below.

Lemma 29.25 *Let* G *be an* n*-dimensional stabilizing subgroup of* Π^n *and let* $h \in \Pi^n$ *be such that* $\operatorname{coeff}(h) = \pm 1$ *and neither* h *nor* $-h$ *is in* G*. Let* $C := \{g \in G \mid gh = hg\}$ *and* $hC := \{hg \mid g \in C\}$*. Then* $C \neq G$ *and* $C \cup hC$ *is an* n*-dimensional stabilizing subgroup of* Π^n*.*

Proof. This all follows immediately from Lemma 29.17 (with $k = n$) provided we can show that $C \neq G$. Let $H := C \cup hC$. H is a stabilizing subgroup of Π^n by Lemma 29.17 of dimension either n or $n+1$ (the latter if $C = G$). By Corollary 29.13, no stabilizing subgroup of Π^n can have more than 2^n elements, and therefore H has dimension n. This implies that $C \neq G$, since $|C| = |H|/2 = 2^{n-1} < 2^n = |G|$.[37] □

We apply Lemma 29.25 twice—once with $h = Z_1$ and again with $h = -Z_1$—to find the two alternative post-measurement states (actually the groups G_0 and G_1 that stabilize them) in the case where neither Z_1 nor $-Z_1$ is in G. Letting $C \subseteq G$ be the set of all elements of G that commute

[37] Here we use the vertical bars to indicate the cardinality of a set.

with Z_1, we define $G_0 := C \cup Z_1 C$ and $G_1 := C \cup (-Z_1)C$. By the lemma, both are n-dimensional stabilizing subgroups of Π^n. We now verify that for $b \in \{0,1\}$, P_{G_b} is the post-measurement state given outcome b. From Equations (123–127), the lemma, and the fact that $\Pr[0] = \Pr[1] = 1/2$, we have

$$\begin{aligned}
\rho_0 &= 2P^+_{Z_1}\rho P^+_{Z_1} = 2^{-n}\sum_{g\in C}(g + Z_1 g) = 2^{-n}\sum_{g\in G_0} g = P_{G_0}\,, \\
\rho_1 &= 2P^-_{Z_1}\rho P^-_{Z_1} = 2^{-n}\sum_{g\in C}(g - Z_1 g) = 2^{-n}\sum_{g\in C}(g + (-Z_1)g) = 2^{-n}\sum_{g\in G_1} g = P_{G_1}\,.
\end{aligned}$$

There are two things left to do to complete the proof of the Gottesman-Knill theorem: (1) show how to determine easily which of the three cases applies for a 1-qubit measurement; (2) in Case 3, determine *generators* for G_0 and G_1 given *generators* for G.

We are given a minimal set of generators $S := \{g_1, \ldots, g_n\}$ for $G = \operatorname{Stab}\rho$, the group stabilizing the pre-measurement state. We can easily distinguish Case 3 from the other two: by Lemma 29.25, one of $\pm Z_1$ is in G if and only if Z_1 commutes with all of G, if and only if Z_1 commutes with all of S. The latter can be easily checked: Z_1 commutes with g_j if and only if $\operatorname{princ}(g_j) = I \otimes \cdots$ or $\operatorname{princ}(g_j) = Z \otimes \cdots$.

If Z_1 commutes with all of S, then to determine which of Z_1 or $-Z_1$ is in G, we first find a subset $T \subseteq S$ whose elements multiply to $\pm Z_1$, then we actually multiply these together to see what we get—either Z_1 or $-Z_1$. Finding T is essentially a problem in linear algebra over $\mathbb{Z}_2$, via the correspondence given in the previous topic. Using standard techniques of linear algebra (Gaussian elimination, particularly), we can express $\varphi(Z_1)$ as a linear combination (actually a simple sum, since the field is $\mathbb{Z}_2$) of elements from the set $\{\varphi(g_1), \ldots, \varphi(g_n)\}$. Multiplying together those g_j such that $\varphi(g_j)$ appears in the sum will yield $\pm Z_1$.

Finally, here is how to get generators for G_b for $b \in \{0,1\}$ in the case where neither Z_1 nor $-Z_1$ is in G. Choose one of the generators that anticommutes with Z_1—suppose it is g_1 (it doesn't matter which one it is). Replace g_1 in the generating set with $(-1)^b Z_1$ (that is, Z_1 if $b = 0$ and $-Z_1$ if $b = 1$), then for any other generator g_j that anticommutes with Z_1, replace g_j by $g_1 g_j$. The result is a generating set S_b for G_b.

To see why, first note that all elements of S_b commute with Z_1, and so they are all in C except for $\pm Z_1$ itself. Thus $S_b \subseteq G_b$. It remains to show that all of G_b is generated by S_b. For this it suffices to show that all of C is generated by S_b, because $(-1)^b Z_1$ is also in S_b. Every element $g \in G$ is a unique product of distinct elements from S, say,

$$g = g_{i_1} g_{i_2} \cdots g_{i_k}$$

for some k and $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$. Then $g \in C$ (that is, g commutes with Z_1) if and only if Z_1 anticommutes with an *even* number of the factors g_{i_j}: starting with $Z_1 g = Z_1 g_{i_1} g_{i_2} \cdots g_{i_k}$, transpose Z_1 with g_{i_1} then g_{i_2}, etc., until Z_1 winds up on the far right. To keep all these expressions equal, every time Z_1 anticommutes with one of these factors, we must introduce a minus sign out front, so the minus signs cancel just when there are an even number of such anticommutations. Thus we can group the factors $g_{i_j} \notin C$ into adjacent pairs inside the product on the right-hand side, above. But now each pair is obtainable from S_b. For example, if g_3 and g_5 are paired, then $g_3 g_5 = g_1^2 g_3 g_5 = (g_1 g_3)(g_1 g_5)$, and both $g_1 g_3$ and $g_1 g_5$ are in S_b. The other unpaired factors are in

C and hence are already in S_b. This shows that every element of C is the product of factors from S_b, which finishes the proof of the Gottesman-Knill theorem.

Remark. The only difference between S_0 and S_1 above is that S_0 contains Z_1 and S_1 contains $-Z_1$ instead. One can easily simulate any number of 1-qubit measurements in the circuit by maintaining a single expression for S_b (involving b) after each successive measurement.

Exercise 29.26 Referring back to Exercise 29.24, suppose that after all the gates are applied, the three qubits are measured in order. Give the probability of the outcomes of each measurement and the corresponding post-measurement states. Subsequent measurements may depend on previous results. Describe all possible sequences of outcomes and their probabilities.

Stabilizer Codes.

29.2 Entanglement

Suppose we have two physical systems with Hilbert spaces $\mathcal{H}$ and $\mathcal{J}$. If the first system is prepared in some pure state $\rho \in \mathcal{L}(\mathcal{H})$ and the second is independently prepared in a pure state $\sigma \in \mathcal{L}(\mathcal{J})$—and the two systems do not interact in any way—then the (pure) state of the combined system is $\rho \otimes \sigma \in \mathcal{L}(\mathcal{H}) \otimes \mathcal{L}(\mathcal{J}) \cong \mathcal{L}(\mathcal{H} \otimes \mathcal{J})$.[38] We call such a pure state *separable* between the two systems, or a *tensor product state*. For pure states, being a tensor product state and being a separable state are equivalent, but for general (not necessarily pure) states, the notion of separability is looser. We will only consider pure states, and then only those of two combined systems—so-called "bipartite" pure states.

$\mathcal{L}(\mathcal{H} \otimes \mathcal{J})$ contains lots of pure states that are not of the form $\rho \otimes \sigma$ as above. Such pure states are said to be *entangled*. Roughly, two physical systems are in an entangled state when they are correlated in a non-classical (uniquely quantum) way. The four Bell states are entangled states of two single-qubit systems—maximally entangled, it will turn out. None of them can be written as the tensor product of two single-qubit states.

Our first task is to find a way to quantify mathematically the amount of entanglement of a pure state shared between two systems. For this we use the *Schmidt decomposition*.

Theorem 29.27 (Schmidt Decomposition) *Let $\mathcal{H}$ and $\mathcal{J}$ be Hilbert spaces, and let $u \in \mathcal{H} \otimes \mathcal{J}$ be any unit vector. There exists a unique integer $r > 0$ and unique positive values $s_1 \geqslant \cdots \geqslant s_r > 0$ such that there exist pairwise orthogonal unit vectors $x_1, \ldots, x_r \in \mathcal{H}$ and $y_1, \ldots, y_r \in \mathcal{J}$ such that $\sum_{k=1}^{r} s_k^2 = 1$ and*

$$u = \sum_{k=1}^{r} s_k (x_k \otimes y_k) \,. \tag{128}$$

In fact, $\{s_1^2, \ldots, s_r^2\}$ is the multiset of nonzero eigenvalues of $\mathrm{tr}_{\mathcal{J}}(uu^*)$ and of $\mathrm{tr}_{\mathcal{H}}(uu^*)$.

[38] The $\cong$ relation indicates that these two spaces are naturally isomorphic.

Proof. The Schmidt decomposition is really the singular value decomposition in disguise. Pick some standard orthonormal bases $e_1, \ldots, e_n$ for $\mathcal{H}$ and $f_1, \ldots, f_n$ for $\mathcal{J}$. (We will assume that $\dim(\mathcal{H}) = \dim(\mathcal{J}) = n$ for technical convenience, but this is not necessary.) We expand u with respect to the product basis $\{e_i \otimes f_j\}_{1 \leqslant i,j \leqslant n}$ as

$$u = \sum_{1 \leqslant i,j \leqslant n} \alpha_{i,j} (e_i \otimes f_j)$$

for some unique coefficients $\alpha_{i,j} \in \mathbb{C}$.

Let A be the $n \times n$ matrix whose entries are the $\alpha_{i,j}$, i.e., $[A]_{ij} = \alpha_{i,j}$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$. By the singular value decomposition (Theorem B.9 in Section B.3) there exist unique real values $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_n \geqslant 0$ such that there exist $n \times n$ unitary matrices V, W with $A = VDW$, where $D = \operatorname{diag}(s_1, \ldots, s_n)$. Let r be the largest natural number such that $s_r > 0$. We have $r \geqslant 1$ because $A \neq 0$. Now for all $1 \leqslant k \leqslant r$, let

$$\begin{aligned} x_k &:= \sum_{i=1}^{n} [V]_{ik} e_i \,, \\ y_k &:= \sum_{j=1}^{n} [W]_{kj} f_j \,. \end{aligned}$$

There are three things to check here. We first check that $\{x_k\}$ and $\{y_k\}$ are orthonormal sets of vectors. Using the fact that V is unitary, we have, for all $1 \leqslant k, \ell \leqslant r$, letting $v_{ik} := [V]_{ik}$,

$$\langle x_k, x_\ell \rangle = \left\langle \sum_{i=1}^{n} v_{ik} e_i, \sum_{j=1}^{n} v_{j\ell} e_j \right\rangle = \sum_{i,j} v_{ik}^* v_{j\ell} \langle e_i, e_j \rangle = \sum_{i=1}^{n} [V^*]_{ki} [V]_{i\ell} = [V^* V]_{k\ell} = [I]_{k\ell} = \delta_{k\ell} \,.$$

A similar calculation shows that $\langle y_k, y_\ell \rangle = \delta_{k\ell}$, using the unitarity of W. Thus both $\{x_k\}$ and $\{y_k\}$ are orthonormal sets.

Second, we have

$$\begin{aligned} u &= \sum_{1 \leqslant i,j \leqslant n} [A]_{ij} (e_i \otimes f_j) = \sum_{i,j} [VDW]_{ij} (e_i \otimes f_j) \\ &= \sum_{i,j} \sum_{1 \leqslant k,\ell \leqslant n} [V]_{ik} [D]_{k\ell} [W]_{\ell j} (e_i \otimes f_j) \\ &= \sum_{i,j,k,\ell} s_k [V]_{ik} \delta_{k\ell} [W]_{\ell j} (e_i \otimes f_j) \\ &= \sum_{i,j,k} s_k [V]_{ik} [W]_{kj} (e_i \otimes f_j) \\ &= \sum_{k=1}^{r} s_k \sum_{i,j} [V]_{ik} [W]_{kj} (e_i \otimes f_j) \\ &= \sum_{k=1}^{r} s_k \left(\sum_i [V]_{ik} e_i \right) \otimes \left(\sum_j [W]_{kj} f_j \right) \\ &= \sum_{k=1}^{r} s_k (x_k \otimes y_k) \,. \end{aligned}$$

Since u is a unit vector, we also have

$$1 = u^*u = \sum_{k=1}^{r}\sum_{\ell=1}^{r} s_k s_\ell x_k^* x_\ell y_k^* y_\ell = \sum_{k=1}^{r} s_k^2 \ ,$$

the last equation following from the fact that $x_k^* x_\ell = y_k^* y_\ell = \delta_{k\ell}$.

Finally, we have

$$\begin{aligned}
\mathrm{tr}_{\mathcal{J}}(uu^*) &= \mathrm{tr}_{\mathcal{J}}\left(\left(\sum_{k=1}^{r} s_k(x_k \otimes y_k)\right)\left(\sum_{\ell=1}^{r} s_\ell(x_\ell^* \otimes y_\ell^*)\right)\right) \\
&= \mathrm{tr}_{\mathcal{J}} \sum_{k,\ell} s_k s_\ell (x_k x_\ell^* \otimes y_k y_\ell^*) \\
&= \sum_{k,\ell} s_k s_\ell \,\mathrm{tr}(y_k y_\ell^*) x_k x_\ell^* \\
&= \sum_{k=1}^{r} s_k^2 x_k x_k^* \ ,
\end{aligned}$$

where for the last equation we have used the fact that $\mathrm{tr}(y_k y_\ell^*) = \mathrm{tr}(y_\ell^* y_k) = \langle y_\ell, y_k\rangle = \delta_{k\ell}$. Similarly, $\mathrm{tr}_{\mathcal{H}}(uu^*) = \sum_{k=1}^{r} s_k^2 y_k y_k^*$. Thus $\{s_k^2 \mid 1 \leqslant k \leqslant r\}$ is the multiset of the nonzero eigenvalues of $\mathrm{tr}_{\mathcal{J}}(uu^*)$ and of $\mathrm{tr}_{\mathcal{H}}(uu^*)$, with corresponding eigenvectors x_k and y_k, respectively. We then have that r is the common rank of $\mathrm{tr}_{\mathcal{J}}(uu^*)$ and of $\mathrm{tr}_{\mathcal{H}}(uu^*)$, and since $s_k > 0$ for all $1 \leqslant k \leqslant r$, it follows that r and $s_1, \ldots, s_r$ are uniquely determined by u. □

The number r is called the *Schmidt number*, or *Schmidt rank* of u; the λ_j are called the *Schmidt coefficients* of u. The (not necessarily unique) vectors $x_1, \ldots, x_r$ and $y_1, \ldots, y_r$ are known collectively as a *Schmidt basis* for u, although they may not span their respective spaces.

The Schmidt rank r of u tells us immediately whether u is entangled. The answer is yes if and only if $r > 1$. The Schmidt coefficients give more information about the degree of entanglement, which we now explore.

Shannon entropy and von Neumann entropy.

Definition 29.28 (Shannon entropy) Given a probability distribution $p = (p_1, \ldots, p_n)$,[39] we define the *Shannon entropy* of p to be the quantity

$$H(p) = -\sum_{i=1}^{n} p_i \lg p_i \ , \tag{129}$$

where $0 \lg 0 = 0$ by convention (alternately, we restrict the sum to those i for which $p_i > 0$).

$H(p)$ measures the amount of uncertainty inherent in a p-distributed random experiment. Equivalently, $H(p)$ gives the amount of information (in bits) about the outcome obtained by running

[39] That is, each $p_i \geqslant 0$, and $\sum_{i=1}^{n} p_i = 1$.

such an experiment, averaged over all the possible outcomes. For any probability distribution $p = (p_1, \ldots, p_n)$ on n outcomes,

$$0 \leqslant H(p) \leqslant \lg n \, .$$

$H(p) = 0$ exactly when p is *deterministic*, i.e., $p_i = 1$ for some i (and thus $p_j = 0$ for all $j \neq i$). $H(p) = \lg n$ if and only if p is the *uniform distribution*, i.e., $p_i = 1/n$ for all $1 \leqslant i \leqslant n$.

Shannon entropy has a quantum analogue. For any state ρ, we have $\rho \geqslant 0$ and $\operatorname{tr} \rho = 1$, which is equivalent to the spectrum of ρ being a probability distribution.

Definition 29.29 (Von Neumann entropy) Let $\mathcal{H}$ be a Hilbert space of dimension $n > 0$, let $\rho \in \mathcal{L}(\mathcal{H})$ be a state, and let $\lambda := (\lambda_1, \cdots, \lambda_n)$ be the spectrum (vector of eigenvalues) of ρ, where $\lambda_1 \geqslant \cdots \geqslant \lambda_n$. We define the *von Neumann entropy* of ρ as

$$H(\rho) := H(\lambda) \, , \tag{130}$$

where the right-hand side is the Shannon entropy of λ.

The von Neumann entropy of ρ can be written succinctly as

$$H(\rho) = -\operatorname{tr}(\rho \lg \rho) \, .$$

If $\rho > 0$, then this expression makes perfect sense using the rule for applying a scalar function to a normal operator that we discussed in Section 9. If one or more of ρ's eigenvalues is zero, however, then some care must be taken with this expression, just as we did when defining Shannon entropy in the case where one or more of the probabilities was zero. In this case we can confine ρ to a subspace on which it acts positive definitely. Let $V := \operatorname{im}(\rho) := \{\rho v \mid v \in \mathcal{H}\}$, and let σ be ρ restricted to V (V is a subspace of $\mathcal{H}$). Then $\sigma > 0$, and so we can define $H(\rho) := -\operatorname{tr}(\sigma \lg \sigma)$, and this coincides with Definition 29.29.

Analogous to Shannon entropy, von Neumann entropy quantifies the amount of uncertainty about a quantum state. We can view a pure state as one about which we have complete information. If $\rho_1, \ldots, \rho_k$ are pure states that are pairwise orthogonal (that is, $\langle \rho_i, \rho_j \rangle = 0$ for all $i \neq j$), then there is a projective measurement that can distinguish each ρ_i from the others with certainty. (The projectors of this measurement are the ρ_i themselves, each ρ_i corresponding to outcome i, possibly together with one additional projector $P := I - \sum_{i=1}^{k} \rho_i$ for the outcome "none of the above," assuming this projector is nonzero.) Keeping with this view, a mixed state $\rho \in \mathcal{L}(\mathcal{H})$ can be thought of as a state about which we have *incomplete* information, that is, our information about ρ is only statistical. We can regard ρ as a probabilistic mixture of pairwise orthogonal pure states, and mathematically, these component pure states can come from the spectral decomposition of ρ:

$$\rho = \sum_{i=1}^{n} p_i u_i u_i^* \, ,$$

where $n = \dim(\mathcal{H})$, $(p_1, \ldots, p_n)$ is a probability distribution, $\{u_i \mid 1 \leqslant i \leqslant n\}$ is an eigenbasis for ρ, and each pure state is $\rho_i := u_i u_i^*$ for $1 \leqslant i \leqslant n$. In this view, the von Neumann entropy of ρ is just the Shannon entropy of the classical probability distribution $(p_1, \ldots, p_n)$.

How does all this relate to entanglement? Given a pure state $\rho := uu^*$, where u is a unit vector in $\mathcal{H} \otimes \mathcal{J}$, if we decide to ignore (by tracing out) one or the other system, then the reduced state (i.e., the state of the remaining system) will be mixed if and only if ρ is entangled. This suggests a natural quantitative measure of the amount of entanglement in ρ—the amount of uncertainty we have about either of these reduced states, given by their von Neumann entropy. This quantity can be computed directly from the Schmidt coefficients $s_1, \ldots, s_r$ of ρ:

$$H(\mathrm{tr}_{\mathcal{H}}(\rho)) = H(\mathrm{tr}_{\mathcal{J}}(\rho)) = H(s_1^2, \ldots, s_r^2) \,,$$

noting that $(s_1^2, \ldots, s_r^2)$ is a probability distribution by Theorem 29.27.

29.3 Bell inequalities

In 1935, Albert Einstein, Boris Podolsky, and Nathan Rosen (EPR from now on) published a paper arguing that the laws of quantum mechanics, although correct to the best of anyone's knowledge, were not a complete description of nature. Their argument was based on two assumed principles, which are commonly called "locality" and "realism":

Locality. All physical influences act locally; put another way, there is no action at a distance. Object A cannot directly influence a distant object B without some intervening continuum of local influences connecting the two. For example, that two distant, oppositely charged particles attract is not due to any direct influence between them but rather due each responding (locally) to the other's electromagnetic field, which permeates all of space. For another example, according to general relativity, a massive object warps the spacetime around it so that nearby objects move along curved paths, even though locally they are still moving in straight lines.

Realism. A complete knowledge of the state of a physical system is, in principle, enough to predict the outcome of every possible measurement of that system. For example knowing the exact trajectory of an asteroid now (as well as the gravitational forces acting on it) allows us to predict where it will be a year from now, the accuracy of the prediction limited only by the precision of the initial measurements and of our calculations.

If we prepare an electron spin in state $|+\rangle$ (i.e., $|\rightarrow\rangle$, spin-right) and we then measure its spin in the vertical direction, we get an apparently uniformly random result: spin-up about half the time, spin-down the rest of the time. The realistic view says that this apparent randomness is not fundamental physics; it is instead an artifact of our incomplete understanding of the state of the electron. There are aspects of the electron's state that we don't know about and that our current theory is not accounting for—so called *hidden variables*—that predetermine its vertical spin before we measure it. If we think we are preparing each electron in the same state $|+\rangle$ but getting different results measuring the spin vertically, then the electrons really aren't in the same state to begin with, and our theory is not (as yet) adequate to account for that difference. A complete description of a physical state would determine all measurement outcomes; nature is not inherently random. (Einstein: "God does not play dice.") This is the realist view of physics.

In a modification[40] of EPR's argument, we consider a system of two spin-$\frac{1}{2}$ particles, created together in a lab then separated from each other by an arbitrary distance. The particles are created

[40] This modification is close to one due to David Bohm in 1951.

in a closed system with zero net angular momentum in any direction; conservation of angular momentum then requires that the two spins always be measured as opposite, resulting in a net spin of zero. Quantum mechanics dictates that the pair is in the entangled state $|\Psi^-\rangle = (|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)/\sqrt{2}$—called the *spin singlet state*—which has the property that if we measure the spin of each particle in the same direction, we will always get opposite results, regardless of the direction chosen. Now imagine the two particles being moved very far apart (even lightyears apart), Alice having one and Bob the other. According to quantum mechanics, if Alice or Bob measures their spin in the vertical direction they will see $\uparrow$ or $\downarrow$ uniformly at random, *but*, if, say, Alice measures her spin and sees $\uparrow$, then Bob *must* see $\downarrow$, and vice versa, although he interprets his own result as being uniformly random. According to the prevailing interpretation of quantum mechanics (the so-called Copenhagen interpretation), when Alice measures her spin and sees $\uparrow$, say, the state of the system "collapses" to $|\uparrow\downarrow\rangle$, ensuring that Bob will subsequently measure $\downarrow$. This interpretation appears to violate local realism: Alice's measurement result alters the system's state, thus magically influencing Bob's measurement, even though Bob is in another galaxy and (even worse), light may not have time to travel from one measurement event to the other (the two events are spacelike separated). Einstein criticized this interpretation as "spooky action at a distance."

EPR posited an alternative, local realist interpretation of this scenario. When the two particles were first created together in the lab, hidden variables were fixed locally between the two particles, predisposing Alice's particle to result in $\uparrow$ when measured, and Bob's particle to result in $\downarrow$. These hidden parameters perfectly correlated the two particles when they were in the same place (locality), then were carried with the particles as they were separated; the measurement results were not random, but were predetermined by these hidden variables (realism). Entanglement and the subsequent state collapse are fictions; Alice was always going to see $\uparrow$, say, and Bob $\downarrow$, the die being cast when the particles were created in the same place.

In the following decades, the EPR argument was taken less as physics than as philosophy, since there seemed to be no experiment that could confirm or refute it (quantum mechanics versus local realism). Then in 1964, John Bell showed how EPR's interpretation has direct physical consequences. He proposed a clever physics experiment that could test the EPR hypothesis. Bell showed that local realism implies that statistical correlations between measurements of spatially separated systems must satisfy certain constraints—now known as *Bell inequalities*—whereas quantum mechanics predicts that these constraints are violated. By taking a large enough sample of runs of the same experiment and gathering the statistics, one could either confirm or refute (based on statistical evidence, at least) the local realist interpretation.

A number of different Bell inequalities are now known, and we consider two in depth below. Several experiments have been performed to test these inequalities. Although doubts have been raised from time to time about statistical loopholes allowing for a local realist interpretation of some of the experimental results, the overwhelming evidence at this point is that nature violates the Bell inequalities. Hidden-variable theories are refuted, and there is every reason to think that quantum mechanics offers a complete description of reality. A good philosophical discussion of the EPR paradox can be found online in the *Stanford Encyclopedia of Philosophy* (`https://plato.stanford.edu/entries/qt-epr/`).

We give two examples of Bell inequalities in this section, showing how the laws of quantum mechanics violate each. Each is cast in terms of a *nonlocal game*, which we now describe. A nonlocal

game is a cooperative game played by two parties, Alice and Bob, and a referee. The referee first produces a pair (s, t) of values, probabilistically drawn from some finite set. The ref gives s to Alice and t to Bob. Alice then produces a value a and Bob a value b, which they send back to the referee. Then Alice and Bob *win* if the tuple (s, t, a, b) satisfies a certain finite condition (which depends on the type of game being played); otherwise, they lose. Before getting s and t from the ref, Alice and Bob can get together beforehand and share any information, randomness, strategies, etc. that they want, but they are not allowed to communicate with each other from the time they receive s and t until the time they send a and b to the ref.

We say that Alice and Bob employ a *classical* strategy if what they share beforehand is purely classical information, including shared randomness. They employ a *quantum* strategy if, in addition, they share an entangled quantum state beforehand. We define the *value* of a particular strategy to be the overall probability of Alice and Bob winning using that strategy.

In each of the two example games we discuss below, there is a quantum strategy whose value is strictly higher than any optimal classical strategy.[41] These then imply violations of the corresponding Bell inequalities: the local realist interpretation dictates that classical strategies are the only ones available to Alice and Bob.

Nonlocal games and their limitations are explored extensively in a paper by Cleve, Høyer, Toner, and Watrous (`https://arxiv.org/abs/quant-ph/0404076`).

The CHSH game. In this game, based on a Bell-type inequality discovered by Clauser, Horne, Shimony, and Holt, the referee chooses values $s, t \in \{0, 1\}$ uniformly at random and independently, so that each pair (s, t) occurs with probability $\frac{1}{4}$. Then Alice and Bob produce values a and b in $\{+1, -1\}$, respectively. Alice and Bob win if and only if $ab = (-1)^{st}$; more prosaically, if $s = t = 1$, then Alice and Bob win iff $a \neq b$, and otherwise, they win iff $a = b$. We will show that using any classical strategy, Alice and Bob can win with probability no greater than $\frac{3}{4} = 0.75$, but using a quantum strategy, they can win with probability $\cos^2(\pi/8) = (2 + \sqrt{2})/4 \approx 0.85$.

First, we consider the case where Alice and Bob employ a (classical) deterministic strategy, that is, Alice computes $a := A(s)$, where $A : \{0, 1\} \to \{+1, -1\}$ is a function she chooses beforehand, and in a similar fashion Bob computes $b := B(t)$ for some $B : \{0, 1\} \to \{+1, -1\}$ of his choosing. There are four such functions: the constant $+1$ function, the constant -1 function, the function mapping $x \mapsto (-1)^x$, and the function mapping $x \mapsto (-1)^{1+x}$. The following table gives, for each possible pair of functions A and B, the value of ab given each of the four possible choices of (s, t):

	$B(x) = +1$	$B(x) = -1$	$B(x) = (-1)^x$	$B(x) = (-1)^{1+x}$
$A(x) = +1$	$+ \quad +$ $+ \quad +$	$- \quad -$ $- \quad -$	$+ \quad -$ $+ \quad -$	$- \quad +$ $- \quad +$
$A(x) = -1$	$- \quad -$ $- \quad -$	$+ \quad +$ $+ \quad +$	$- \quad +$ $- \quad +$	$+ \quad -$ $+ \quad -$
$A(x) = (-1)^x$	$+ \quad +$ $- \quad -$	$- \quad -$ $+ \quad +$	$+ \quad -$ $- \quad +$	$- \quad +$ $+ \quad -$
$A(x) = (-1)^{1+x}$	$- \quad -$ $+ \quad +$	$+ \quad +$ $- \quad -$	$- \quad +$ $+ \quad -$	$+ \quad -$ $- \quad +$

[41] An optimal classical strategy always exists, but it may not be unique.

Each boxed entry of the table is a 2×2 matrix with entries ± 1. For brevity, we write "$+$" for $+1$ and "$-$" for -1. The rows of each matrix are indexed by the value of s (0 then 1) and the columns similarly by t. It follows that each matrix is of the form

$$M(A,B) := \begin{bmatrix} A(0) \\ A(1) \end{bmatrix} \begin{bmatrix} B(0) & B(1) \end{bmatrix} = \begin{bmatrix} A(0)B(0) & A(0)B(1) \\ A(1)B(0) & A(1)B(1) \end{bmatrix}$$

for the given choice of functions A and B. The matrix giving $(-1)^{st}$ (and hence the winning values for ab) is

$$W := \begin{bmatrix} + & + \\ + & - \end{bmatrix}.$$

If Alice and Bob choose functions A and B, respectively, then they win for each particular (s,t) if the corresponding entry in $M(A,B)$ matches that of W. By inspecting the table above, one observes that for each choice of A and B, the matrix $M(A,B)$ differs from W in at least one of the four entries. Since each combination (s,t) occurs with probability $\frac{1}{4}$, the probability that they win is at most $1 - \frac{1}{4} = \frac{3}{4}$, no matter which A and B they choose. (Alice and Bob can achieve this optimal probability by always outputting $a = b = +1$, for example. Other strategies are optimal as well.)

There are two useful things to note here that will also apply to the classical case in the next game we consider:

1. Each $M(A,B)$ (considered as a matrix over $\mathbb{R}$) is the product of a column vector with a row vector, and as such, has rank 1. On the other hand, we observe that W is nonsingular, of rank 2. This is an alternate, more succinct way of seeing that $M(A,B) \neq W$ for all A and B.

2. Deterministic strategies (also called *pure strategies*) are not the only classical strategies available to Alice and Bob. They could instead employ a *mixed strategy* wherein they chose their A and B at random according to some arbitrary joint probability distribution. In this case, however, their probability of winning is then a convex combination of their winning probabilities using pure strategies, and so cannot exceed that of the best pure strategy. Thus we only need to consider pure strategies to get an upper bound for all classical strategies.

We now turn to Alice's and Bob's quantum strategy. Before receiving s and t from the referee, Alice and Bob share an EPR pair, i.e., the 2-qubit state

$$|\psi\rangle := \left|\Phi^+\right\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

the first qubit possessed by Alice, the second by Bob. It is best to conceive of each qubit as a spin-$\frac{1}{2}$ particle. After receiving s, Alice measures her particle's spin in a certain direction depending on the value of s. After receiving t, Bob measures his particle's spin in a certain direction depending on the value of t. If Alice sees spin-up ($\uparrow$), then she sends $a := +1$ to the referee; otherwise (if spin-down ($\downarrow$), she sends $a := -1$. Bob computes b using the same method; the only difference between Alice and Bob is which directions they choose to measure their respective spins.

Both spin measurements are in the x,z-plane. Generally, for any angle θ, a projective measurement of a spin in the direction having cartesian coordinates $(\sin\theta, 0, \cos\theta)$ (that is, clockwise from the upward direction through angle θ) corresponds to the csop

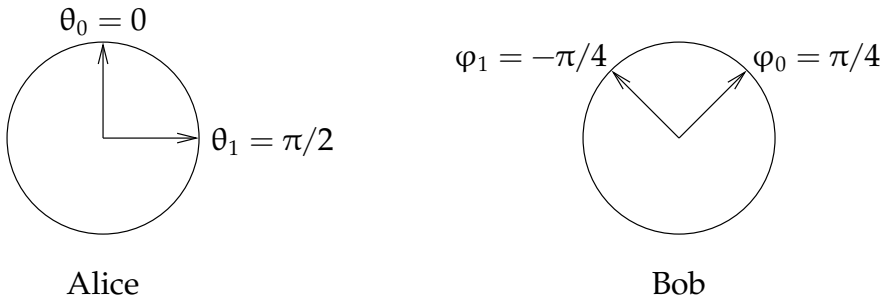
$$P_\uparrow(\theta) = \frac{1}{2}(I + (\sin\theta)X + (\cos\theta)Z)\,, \tag{131}$$

$$P_\downarrow(\theta) = \frac{1}{2}(I - (\sin\theta)X - (\cos\theta)Z)\,, \tag{132}$$

where I is the 2×2 identity matrix and X and Z are the usual Pauli matrices. Upon receiving s, Alice chooses angle θ_s to measure her spin. Upon receiving t, Bob chooses angle φ_t to measure his spin. The angles θ_s and φ_t are given by the following tables:

	$s=0$	$s=1$
θ_s	0	$\pi/2$

	$t=0$	$t=1$
φ_t	$\pi/4$	$-\pi/4$



To find the winning probability, we must compute, for any combination (s,t), the probability that $a = b$. Generally, if Alice measures her spin with angle θ and Bob measures his spin with angle φ, then the combined measurement corresponds to the 4-outcome csop

$$\{P_\uparrow(\theta)\otimes P_\uparrow(\varphi),\ \ P_\uparrow(\theta)\otimes P_\downarrow(\varphi),\ \ P_\downarrow(\theta)\otimes P_\uparrow(\varphi),\ \ P_\downarrow(\theta)\otimes P_\downarrow(\varphi)\}\,.$$

Let $C\otimes D$ be any of these four projectors (or let C and D be any 2×2 matrices generally). We have a handy formula for the probability of obtaining the corresponding outcome given state $|\psi\rangle$:

$$\begin{aligned}
\langle\psi|(C\otimes D)|\psi\rangle &= \langle\psi|\left(\begin{bmatrix} c_{11} & c_{12}\\ c_{21} & c_{22}\end{bmatrix}\otimes\begin{bmatrix} d_{11} & d_{12}\\ d_{21} & d_{22}\end{bmatrix}\right)|\psi\rangle\\
&= \frac{1}{2}\begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix} c_{11}d_{11} & c_{11}d_{12} & c_{12}d_{11} & c_{12}d_{12}\\ c_{11}d_{21} & c_{11}d_{22} & c_{12}d_{21} & c_{12}d_{22}\\ c_{21}d_{11} & c_{21}d_{12} & c_{22}d_{11} & c_{22}d_{12}\\ c_{21}d_{21} & c_{21}d_{22} & c_{22}d_{21} & c_{22}d_{22}\end{bmatrix}\begin{bmatrix}1\\0\\0\\1\end{bmatrix}\\
&= \frac{1}{2}(c_{11}d_{11}+c_{12}d_{12}+c_{21}d_{21}+c_{22}d_{22})\\
&= \frac{1}{2}\operatorname{tr}\left(C^{\mathsf{T}}D\right)
\end{aligned}$$

Applying this formula to the projectors given by (131) and (132), and noting that I, X, and Z are all symmetric and that X and Z have zero trace, we have

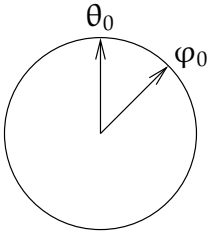
$$\Pr[\uparrow\uparrow] = \frac{1}{8}\operatorname{tr}\left((I+(\sin\theta)X+(\cos\theta)Z)(I+(\sin\varphi)X+(\cos\varphi)Z)\right) = \frac{1+\sin\theta\sin\varphi+\cos\theta\cos\varphi}{4}$$

$$\Pr[\downarrow\downarrow] \;=\; \frac{1}{8}\,\mathrm{tr}\,((I-(\sin\theta)X-(\cos\theta)Z)(I-(\sin\varphi)X-(\cos\varphi)Z)) = \frac{1+\sin\theta\sin\varphi+\cos\theta\cos\varphi}{4}$$

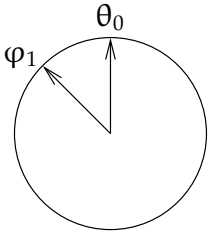
(There is no need for us to compute $\Pr[\uparrow\downarrow]$ or $\Pr[\downarrow\uparrow]$.) Thus

$$\Pr[a=b] = \Pr[\uparrow\uparrow]+\Pr[\downarrow\downarrow] = \frac{1+\sin\theta\sin\varphi+\cos\theta\cos\varphi}{2} = \frac{1+\cos(\theta-\varphi)}{2} = \cos^2\left(\frac{\theta-\varphi}{2}\right)\,.$$

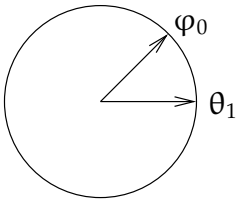
If $|\theta_s-\varphi_t|$ is small, then $\Pr[a=b]$ is close to 1; if $|\theta_s-\varphi_t|$ is close to π, then $\Pr[a=b]$ is close to 0. As the next picture illustrates, Alice's and Bob's measurements are chosen so that $|\theta_s-\varphi_t|$ is close to π if and only if $s=t=1$:



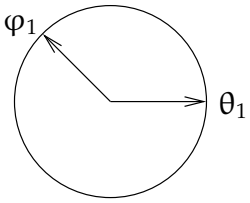
$(s,t)=(0,0)$



$(s,t)=(0,1)$



$(s,t)=(1,0)$



$(s,t)=(1,1)$

If $s=t=1$, then $|\theta_s-\varphi_t| = 3\pi/4$; otherwise, $|\theta_s-\varphi_t| = \pi/4$. Finally, we can compute Alice's and Bob's winning probability:

$$\begin{aligned}
\Pr\left[ab=(-1)^{st}\right] &= \Pr[a=b\wedge st=0]+\Pr[a\neq b\wedge st=1]\\
&= \Pr[st=0]\Pr[a=b\mid st=0]+\Pr[st=1]\Pr[a\neq b\mid st=1]\\
&= \frac{3}{4}\cos^2\left(\frac{\pi}{8}\right)+\frac{1}{4}\left(1-\cos^2\left(\frac{3\pi}{8}\right)\right)\\
&= \frac{3}{4}\cos^2\left(\frac{\pi}{8}\right)+\frac{1}{4}\sin^2\left(\frac{3\pi}{8}\right)\\
&= \frac{3}{4}\cos^2\left(\frac{\pi}{8}\right)+\frac{1}{4}\cos^2\left(\frac{\pi}{8}\right)\\
&= \cos^2\left(\frac{\pi}{8}\right)
\end{aligned}$$

as expected.

The Mermin game. This nonlocal game is based on a Bell inequality violation found by David Mermin, which in turn is based on work by him and Asher Peres. In this game, the referee chooses values $s, t \in \{0, 1, 2\}$ uniformly at random and independently, so that each pair (s, t) is chosen with probability $\frac{1}{9}$. Then, as with the CHSH game above, Alice and Bob produce values a and b in $\{+1, -1\}$, respectively, but now Alice and Bob win if and only if $ab = (-1)^{\delta_{st}}$, where $\delta_{st} = 1$ if $s = t$ and $\delta_{st} = 0$ otherwise. In words, if $s = t$, then Alice and Bob win iff $a \neq b$, and if $s \neq t$, they win iff $a = b$. We will show that using any classical strategy, Alice and Bob can win with probability no greater than $\frac{7}{9} \approx 0.778$, but using a quantum strategy, they can win with probability $\frac{5}{6} \approx 0.833$.

First, the limits of any classical strategy. As we noted in the discussion of the CHSH game above, we need only consider pure (deterministic) strategies for Alice and Bob. Any such pure strategy consists of two functions $A, B : \{0, 1, 2\} \to \{+1, -1\}$, one used by Alice and the other used by Bob. After Alice receives s, she outputs $a := A(s)$, and similarly, Bob outputs $b := B(t)$ upon receiving t.

The winning value of ab for every (s, t)-combination is given by the following 3×3 matrix:

$$W := \left[\; (-1)^{\delta_{st}} \; \right] = \begin{bmatrix} - & + & + \\ + & - & + \\ + & + & - \end{bmatrix},$$

where we again use "$+$" to mean $+1$ and "$-$" to mean -1. For each choice of A and B, the 3×3 matrix giving the ab-values is

$$M(A, B) := \begin{bmatrix} A(0) \\ A(1) \\ A(2) \end{bmatrix} \begin{bmatrix} B(0) & B(1) & B(2) \end{bmatrix}.$$

For a given A and B, the winning probability is $\frac{1}{9}$ times the number of entries of $M(A, B)$ that equal the corresponding entries of W. There are $2^3 = 8$ choices for each of A and B, making 64 matrices $M(A, B)$ in all. Rather than making an exhaustive table as we did for the CHSH game, we note that $M(A, B)$ always has rank 1, whereas it is easily checked that W has rank 3. Thus $M(A, B) \neq W$ for all A and B, and furthermore, changing any single entry of W still leaves two linearly independent columns (at least), resulting in a matrix with rank $\geqslant 2$. Thus $M(A, B)$ must differ from W in *at least* two places, giving a winning probability of at most $1 - \frac{2}{9} = \frac{7}{9}$. (Alice and Bob can achieve this probability by letting A be any nonconstant function and letting $B := -A$.)

For the quantum strategy, Alice and Bob share a pair of qubits in the Bell state

$$|\chi\rangle := \left|\Psi^-\right\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}}.$$

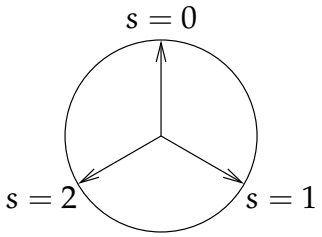
Again, think of Alice and Bob each having a spin-$\frac{1}{2}$ particle. After receiving $s \in \{0, 1, 2\}$ from the referee, Alice projectively measures her spin using the csop

$$\text{Alice:} \quad \left\{ P_\uparrow\left(\frac{2\pi s}{3}\right), \; P_\downarrow\left(\frac{2\pi s}{3}\right) \right\},$$

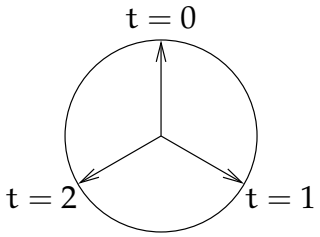
where $P_\uparrow(\theta)$ and $P_\downarrow(\theta)$ are defined by Equations (131) and (132) above for all $\theta \in \mathbb{R}$. Bob makes a similar measurement, except based on t:

$$\text{Bob:} \quad \left\{ P_\uparrow\left(\frac{2\pi t}{3}\right), \; P_\downarrow\left(\frac{2\pi t}{3}\right) \right\}.$$

Here are the three possible spin directions for each of Alice's and Bob's measurements:



Alice



Bob

If Alice sees spin-up ($\uparrow$), then she outputs $a := +1$; if spin-down ($\downarrow$), then she outputs $a := -1$. Likewise, if Bob sees spin-up ($\uparrow$), he outputs $b := +1$; if spin-down ($\downarrow$), he outputs $b := -1$. Generally, if Alice measures using angle θ and Bob measures using angle φ, then the probability of getting the same outcome is

$$\Pr[a = b] = \Pr[\uparrow\uparrow] + \Pr[\downarrow\downarrow] = \langle\chi|(P_\uparrow(\theta) \otimes P_\uparrow(\varphi))|\chi\rangle + \langle\chi|(P_\downarrow(\theta) \otimes P_\downarrow(\varphi))|\chi\rangle = \sin^2\left(\frac{\theta - \varphi}{2}\right) ,$$

where verifying the last equation is left as an exercise (Exercise 29.30, below). If $s = t$, then Alice and Bob measure their spins in the same direction, giving $\Pr[a = b] = \sin^2 0 = 0$, hence $a \neq b$ with certainty. If $s \neq t$, they measure their spins in directions differing by an angle of $2\pi/3$ (in either direction), and thus $\Pr[a = b] = \sin^2(\pi/3) = \frac{3}{4}$ in this case. Therefore, the winning probability is

$$\begin{aligned}
\Pr\left[ab = (-1)^{\delta_{st}}\right] &= \Pr[s = t \wedge a \neq b] + \Pr[s \neq t \wedge a = b] \\
&= \Pr[s = t]\Pr[a \neq b \mid s = t] + \Pr[s \neq t]\Pr[a = b \mid s \neq t] \\
&= \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot \frac{3}{4} \\
&= \frac{5}{6}
\end{aligned}$$

as desired.

Exercise 29.30 Show that, for all real θ and φ,

$$\left\langle\Psi^-\right|(P_\uparrow(\theta) \otimes P_\uparrow(\varphi))\left|\Psi^-\right\rangle = \left\langle\Psi^-\right|(P_\downarrow(\theta) \otimes P_\downarrow(\varphi))\left|\Psi^-\right\rangle = \frac{1}{2}\sin^2\left(\frac{\theta - \varphi}{2}\right) ,$$

where $|\Psi^-\rangle := (|01\rangle - |10\rangle)/\sqrt{2}$, and $P_\uparrow(\theta)$, $P_\uparrow(\varphi)$, $P_\downarrow(\theta)$, and $P_\downarrow(\varphi)$ are defined by (131–132).

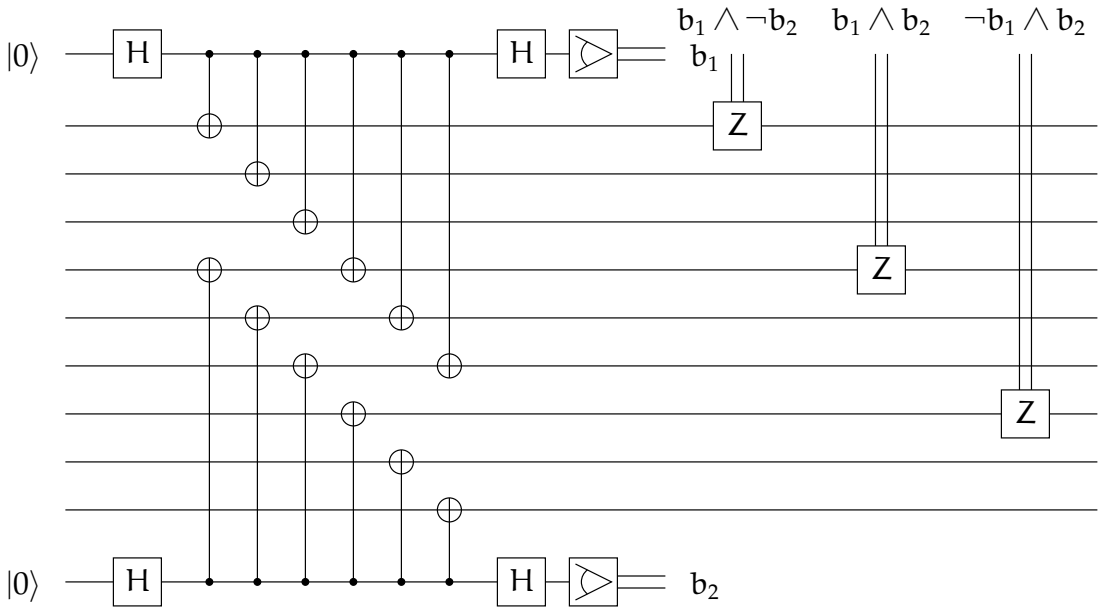
A Final Exam

Do all problems. Hand in your anwsers in at my office or in my mailbox by 5:00 pm on Wednesday, May 9. The same ground rules for the midterm apply here: any resources are at your disposal except discussion with humans other than me about the exam.

All questions with Roman numerals carry equal weight, but may not be of equal difficulty.

I) (A linear algebraic inequality) Suppose that A is any operator. Show that $A \geqslant 0$ if and only if $\mathrm{tr}(PA) \geqslant 0$ for all projectors P. EXTRA CREDIT: Show that $A \geqslant 0$ if and only if $\mathrm{tr}(PA) \leqslant \mathrm{tr}\, A$ for all projectors P. [Hint: The extra credit statement is a corollary of the previous statement.]

II) (A circuit identity) Look at Bob's phase-error recovery circuit for the Shor code in Figure 16. Show that the following alternative circuit does exactly the same thing:



Find a similar alternative for Bob's phase-error recovery circuit in Figure 14.

III) (The square root of SWAP)

(a) Show that if V is any unitary operator, then there exists a (not necessarily unique) unitary U such that $U^2 = V$. [Hint: All unitary operators are normal.]

(b) Find a two-qubit unitary U such that $U^2 = \mathrm{SWAP}$. The U that you find should fix the vectors $|00\rangle$ and $|11\rangle$.

This U is sometimes written as $\sqrt{\mathrm{SWAP}}$. It can be shown that $\sqrt{\mathrm{SWAP}}$, among many other two-qubit gates, is (by itself) universal for quantum computation. Also, there is currently some hope of implementing it flexibly using superconducting Josephson junctions.

IV) (Generalized Pauli gates and the QFT) For $n > 0$, let X_n and Z_n be n-qubit unitary operators such that, for all $x \in \mathbb{Z}_{2^n}$,

$$\begin{aligned} X_n|x\rangle &= |(x+1) \bmod 2^n\rangle, \\ Z_n|x\rangle &= e_n(x)|x\rangle, \end{aligned}$$

recalling that $e_n(x) := \exp(2\pi i x/2^n)$. X_n and Z_n are n-qubit generalizations of the Pauli X and Z gates, respectively.

(a) What are $X_n^* Z_n X_n$ and $Z_n^* X_n Z_n$? (Just show how each behaves on $|x\rangle$ for $x \in \mathbb{Z}_{2^n}$.)

(b) Draw an n-qubit quantum circuit that implements Z_n using only single-qubit conditional phase-shift gates $P(\theta)$ for various θ.

(c) Show that X_n and Z_n are unitarily conjugate via QFT_n.

(d) What are the eigenvalues and eigenvectors of X_n?

V) (The Schmidt Decomposition) You may either take the following on faith or read a proof of it in the textbook. (The Schmidt Decomposition is actually just the Singular Value Decomposition (Theorem B.9 in Section B.3) in disguise.)

Theorem A.1 (Schmidt Decomposition) *Let $\mathcal{H}$ and $\mathcal{J}$ be Hilbert spaces, and let $|\psi\rangle \in \mathcal{H} \otimes \mathcal{J}$ be any unit vector. There exists an integer $k > 0$, pairwise orthogonal unit vectors $|e_1\rangle, \ldots, |e_k\rangle \in \mathcal{H}$ and $|f_1\rangle, \ldots, |f_k\rangle \in \mathcal{J}$, and positive values $\lambda_1 \geqslant \cdots \geqslant \lambda_k > 0$ such that $\sum_{j=1}^k \lambda_j^2 = 1$ and*

$$|\psi\rangle = \sum_{j=1}^{k} \lambda_j (|e_j\rangle \otimes |f_j\rangle). \tag{133}$$

The vectors $|e_1\rangle, \ldots, |e_k\rangle$ and $|f_1\rangle, \ldots, |f_k\rangle$ are known collectively as a *Schmidt basis* for $|\psi\rangle$, although they may not span their respective spaces. The λ_j are called (the) *Schmidt coefficients* for $|\psi\rangle$, and k is called the *Schmidt number* of $|\psi\rangle$.

(a) Give full Schmidt decompositions for the Bell states $|\Phi^+\rangle := (|00\rangle + |11\rangle)/\sqrt{2}$ and $|\Phi^-\rangle := (|00\rangle - |11\rangle)/\sqrt{2}$ in terms of the two single-qubit spaces.

(b) Suppose $|\psi\rangle$ (given by Equation (133)) is projectively measured using the projectors $I_{\mathcal{H}} \otimes |f_1\rangle\langle f_1|, \ldots, I_{\mathcal{H}} \otimes |f_k\rangle\langle f_k|$, and $I_{\mathcal{H}} \otimes \left(I_{\mathcal{J}} - \sum_{j=1}^k |f_j\rangle\langle f_j|\right)$, where $I_{\mathcal{H}}$ and $I_{\mathcal{J}}$ are the identity operators in $\mathcal{L}(\mathcal{H})$ and $\mathcal{L}(\mathcal{J})$, respectively. The last projector corresponds to the default "none of the above" outcome. In terms of the λ_j, what is the probability of each of the $k+1$ outcomes? What is the post-measurement state for each possible outcome?

(c) It is implicit in the book's discussion on page 109 that k, the λ_j, and the Schmidt basis are unique, but they never come out and say it explicitly. Explain briefly why k and $\lambda_1, \ldots, \lambda_k$ are uniquely determined by $|\psi\rangle$. [Hint: Consider the density operator $|\psi\rangle\langle\psi|$ and trace out one of the spaces.]

(d) Show that the Schmidt basis is *not necessarily uniquely determined* by $|\psi\rangle$. Do this by finding a Schmidt basis for $|\Phi^+\rangle$ that is different from the one you found above. (Two Schmidt bases are considered the same if they are identical up to re-ordering and phase factors.)

VI) (Logical Pauli gates for the Shor code) Recall the nine-qubit Shor code defined by Equations (108) and (109).

(a) Show that the operator $Z_1Z_2Z_3Z_4Z_5Z_6Z_7Z_8Z_9$ (i.e., a Pauli Z gate applied to each of the nine qubits) implements the logical Pauli X gate X_S, such that $X_S|0_S\rangle = |1_S\rangle$ and $X_S|1_S\rangle = |0_S\rangle$.

(b) Find an operator that implements the logical Pauli Z gate Z_S, such that $Z_S|0_S\rangle = |0_S\rangle$ and $Z_S|1_S\rangle = -|1_S\rangle$.

B Background Results

Abstract

These results are background to the course CSCE 790S/CSCE 790B, Quantum Computation and Information (Spring 2007 and Fall 2011). Each result, or group of related results, is roughly one page long.

B.1 The Cauchy-Schwarz Inequality

This is one of the most versatile inequalities in all of mathematics.

Theorem B.1 (Cauchy-Schwarz) *For any real numbers* $a_1, \ldots, a_n$ *and* $b_1, \ldots, b_n$,

$$|a_1 b_1 + \cdots + a_n b_n| \leqslant \sqrt{(a_1^2 + \cdots + a_n^2)(b_1^2 + \cdots + b_n^2)}\,, \tag{134}$$

with equality holding iff the two vectors $(a_1, \ldots, a_n)$ *and* $(b_1, \ldots, b_n)$ *are linearly dependent.*

Proof. There are many, many ways of proving this. Here is a direct calculation. We have

$$\begin{aligned}
0 &\leqslant \sum_{1 \leqslant i < j \leqslant n} (a_i b_j - a_j b_i)^2 \\
&= \sum_{i<j} [a_i b_j (a_i b_j - a_j b_i) - a_j b_i (a_i b_j - a_j b_i)] \\
&= \sum_{i<j} [a_i b_j (a_i b_j - a_j b_i) + a_j b_i (a_j b_i - a_i b_j)] \\
&= \sum_{i<j} a_i b_j (a_i b_j - a_j b_i) + \sum_{i<j} a_j b_i (a_j b_i - a_i b_j) \\
&= \sum_{i<j} a_i b_j (a_i b_j - a_j b_i) + \sum_{j<i} a_i b_j (a_i b_j - a_j b_i) \\
&= \sum_{i \neq j} a_i b_j (a_i b_j - a_j b_i) \\
&= \sum_{i,j} a_i b_j (a_i b_j - a_j b_i) \\
&= \sum_{i,j} a_i^2 b_j^2 - \sum_{i,j} a_i b_i a_j b_j \\
&= \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{j=1}^n b_j^2 \right) - \left(\sum_{i=1}^n a_i b_i \right)^2 .
\end{aligned}$$

Adding $(\sum_i a_i b_i)^2$ to both sides then taking the square root of both sides (noting that the square root function is strictly monotone increasing) yields the inequality (134). Clearly, equality holds above iff $a_i b_j - a_j b_i = 0$ for all $i < j$, or equivalently, $a_i b_j = a_j b_i$ for all $i < j$. It is not hard to check that this condition is equivalent to $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ being linearly dependent. □

Note that (134) still holds if we remove the absolute value delimiters from the left-hand side. In that case, equality holds iff there exists a $\lambda \geqslant 0$ such that either $(a_1, \ldots, a_n) = \lambda(b_1, \ldots, b_n)$ or $(b_1, \ldots, b_n) = \lambda(a_1, \ldots, a_n)$.

Corollary B.2 (Triangle Inequality for Complex Numbers) *For any $z, w \in \mathbb{C}$, $|z + w| \leqslant |z| + |w|$.*

Proof. Writing $z = a_1 + a_2 i$ and $w = b_1 + b_2 i$ for real a_1, a_2, b_1, b_2, we have

$$\begin{aligned} |z+w|^2 &= (a_1+b_1)^2 + (a_2+b_2)^2 = a_1^2 + a_2^2 + b_1^2 + b_2^2 + 2(a_1 b_1 + a_2 b_2) \\ &\leqslant a_1^2 + a_2^2 + b_1^2 + b_2^2 + 2\sqrt{(a_1^2 + a_2^2)(b_1^2 + b_2^2)} \\ &= \left(\sqrt{a_1^2 + a_2^2} + \sqrt{b_1^2 + b_2^2}\right)^2 \\ &= (|z| + |w|)^2 . \end{aligned}$$

Taking the square root of both sides yields the corollary. □

Corollary B.3 *For any complex numbers $z_1, \ldots, z_n$ and $w_1, \ldots, w_n$,*

$$|z_1^* w_1 + \cdots + z_n^* w_n| \leqslant \sqrt{(|z_1|^2 + \cdots + |z_n|^2)(|w_1|^2 + \cdots + |w_n|^2)} . \tag{135}$$

Proof. We have

$$\begin{aligned} |z_1^* w_1 + \cdots + z_n^* w_n| &\leqslant |z_1^* w_1| + \cdots + |z_n^* w_n| && \text{(by Corollary B.2)} \\ &= |z_1||w_1| + \cdots + |z_n||w_n| \\ &\leqslant \sqrt{(|z_1|^2 + \cdots + |z_n|^2)(|w_1|^2 + \cdots + |w_n|^2)} . && \text{(by Theorem B.1)} \end{aligned}$$

□

Corollary B.4 *For any column vectors $u, v \in \mathbb{C}^n$,*

$$|\langle u, v \rangle| \leqslant \|u\| \|v\| .$$

B.2 The Schur Triangular Form and the Spectral Theorem

Theorem B.5 (Schur Triangular Form) *For every $n \times n$ matrix M, there exists a unitary U and an upper triangular T (both $n \times n$ matrices) such that $M = UTU^*$.*

Proof. We prove this by induction on n. The $n = 1$ case is trivial. Now supposing the theorem holds for $n \geqslant 1$, we prove it holds for $n+1$. Let M be any $(n+1) \times (n+1)$ matrix. We let A be the linear operator on $\mathbb{C}^{n+1}$ whose matrix is M with respect to some orthonormal basis. A has some eigenvalue λ with corresponding unit eigenvector v. Using the Gram-Schmidt procedure, we can

find an orthonormal basis $\{y_1, \ldots, y_{n+1}\}$ for $\mathbb{C}^{n+1}$ such that $y_1 = v$. With respect to this basis, the matrix for A looks like

$$N = \left[\begin{array}{c|c} \lambda & w^* \\ \hline 0 & N' \end{array}\right],$$

where w is some vector in $\mathbb{C}^n$ and N' is an $n \times n$ matrix. Since M and N represent the same operator with respect to different orthonormal bases, they must be unitarily conjugate, i.e., there is a unitary V such that $M = VNV^*$. N' is an $n \times n$ matrix, so we apply the inductive hypothesis to get a unitary W' and an upper triangular T' (both $n \times n$ matrices) such that $N' = W'T'W'^*$. Now we can factor N:

$$N = \left[\begin{array}{c|c} \lambda & w^* \\ \hline 0 & W'T'W'^* \end{array}\right] = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & W' \end{array}\right] \left[\begin{array}{c|c} \lambda & w^*W' \\ \hline 0 & T' \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & W'^* \end{array}\right] = WTW^*,$$

where

$$W = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & W' \end{array}\right] \qquad \text{and} \qquad T = \left[\begin{array}{c|c} \lambda & w^*W' \\ \hline 0 & T' \end{array}\right].$$

T is clearly upper triangular, and it's easily checked that $WW^* = I$, using the fact that W' is unitary. Thus W is unitary, and we get $M = VNV^* = VWTW^*V^* = UTU^*$, where $U = VW$ is unitary. □

A *Schur basis* for an operator A is an orthonormal basis that gives an upper triangular matrix for A.

Theorem B.6 *If an $n \times n$ matrix A is both upper triangular and normal, then A is diagonal.*

Proof. Suppose that A is upper triangular and normal, but not diagonal. Then there is some $i < j$ such that $[A]_{ij} \neq 0$. Let j be *least* such that there exists $i < j$ such that $[A]_{ij} \neq 0$. For this i and j, we get

$$[AA^*]_{ii} = \sum_{k=1}^{n} [A]_{ik}[A^*]_{ki} = \sum_{k=1}^{n} [A]_{ik}[A]^*_{ik} = \sum_{k=1}^{n} |[A]_{ik}|^2 = \sum_{k=i}^{n} |[A]_{ik}|^2 \geqslant |[A]_{ii}|^2 + \left|[A]_{ij}\right|^2 > |[A]_{ii}|^2. \tag{136}$$

The last inequality follows from the fact that $[A]_{ij} \neq 0$. Similarly,

$$[A^*A]_{ii} = \sum_{k=1}^{n} [A]^*_{ki}[A]_{ki} = \sum_{k=1}^{n} |[A]_{ki}|^2 = \sum_{k=1}^{i} |[A]_{ki}|^2 = |[A]_{ii}|^2. \tag{137}$$

The next to last equation holds because A is upper triangular, and the last equation holds because of our minimum choice of j and the fact that $i < j$. From (136) and (137), we have $[AA^*]_{ii} > [A^*A]_{ii}$. But A is normal, so these two quantities must be equal. From this contradiction we get that A must be diagonal. □

Corollary B.7 (Spectral Theorem for Normal Operators) *Every normal matrix is unitarily conjugate to a diagonal matrix. Equivalently, every normal operator has an orthonormal eigenbasis.*

B.3 The Polar and Singular Value Decompositions

Theorem B.8 (Polar Decomposition) *For every $n \times n$ matrix A there are is an $n \times n$ unitary matrix U and a unique $n \times n$ matrix H such that $H \geqslant 0$ and $A = UH$. In fact, $H = |A|$.*

Proof. First uniqueness. If $A = UH$ with U unitary and $H \geqslant 0$, then

$$|A| = \sqrt{A^*A} = \sqrt{H^*U^*UH} = \sqrt{H^*H} = |H| = H.$$

Now existence. Let $\{e_1, \ldots, e_n\}$ be the standard orthonormal basis for $\mathbb{C}^n$. We first prove the special case where $|A|$ is the diagonal matrix $\operatorname{diag}(s_1, s_2, \ldots, s_n)$ for some real values $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_n \geqslant 0$. Let $0 \leqslant k \leqslant n$ be largest such that $s_k > 0$ ($k = 0$ if $|A| = 0$). Thus we have

$$|A| = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right],$$

where D is the $k \times k$ nonsingular matrix $\operatorname{diag}(s_1, \ldots, s_k)$. If $j > k$, then $|A|e_j = 0$, and thus $0 = |A|e_j = |A|^2 e_j = A^*Ae_j$, whence $\|Ae_j\|^2 = \langle Ae_j, Ae_j\rangle = \langle e_j, A^*Ae_j\rangle = \langle e_j, 0\rangle = 0$, and so $Ae_j = 0$. This means that $A = \left[\begin{array}{c|c} B & 0 \end{array}\right]$, where B is some $n \times k$ matrix, and the last $n - k$ columns of A are 0. We have

$$\left[\begin{array}{c|c} B^*B & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c} B^* \\ \hline 0 \end{array}\right] \left[\begin{array}{c|c} B & 0 \end{array}\right] = A^*A = |A|^2 = \left[\begin{array}{c|c} D^2 & 0 \\ \hline 0 & 0 \end{array}\right],$$

and so $B^*B = D^2$. Let W be an $n \times (n-k)$ matrix whose columns are unit vectors orthogonal to all the columns of B and to each other. (There are many possibilities for W if $k < n$; the columns of W can be any orthonormal set in the orthogonal complement of the space spanned by the columns of B.) By our choice of W, we have $B^*W = 0$, $W^*B = 0$, and $W^*W = I$. Finally, define $U := \left[\begin{array}{c|c} BD^{-1} & W \end{array}\right]$. We claim that U is unitary and that $A = U|A|$. Noting that D^{-1} is Hermitean, we have

$$U^*U = \left[\begin{array}{c} D^{-1}B^* \\ \hline W^* \end{array}\right] \left[\begin{array}{c|c} BD^{-1} & W \end{array}\right] = \left[\begin{array}{c|c} D^{-1}B^*BD^{-1} & D^{-1}B^*W \\ \hline W^*BD^{-1} & W^*W \end{array}\right] = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array}\right] = I,$$

and therefore U is unitary. We also have

$$U|A| = \left[\begin{array}{c|c} BD^{-1} & W \end{array}\right] \left[\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} B & 0 \end{array}\right] = A.$$

Now for the general case. Since $|A| \geqslant 0$ (and hence normal), there is a unitary V such that $V|A|V^* = \operatorname{diag}(s_1, \ldots, s_n)$ for some real values $s_1 \geqslant \cdots \geqslant s_n \geqslant 0$. Since

$$V|A|V^* = V\sqrt{A^*A}V^* = \sqrt{VA^*AV^*} = \sqrt{(VAV^*)^*(VAV^*)} = |VAV^*|,$$

we see that VAV^* satisfies the special case, above, and so there is a unitary U such that $VAV^* = U|VAV^*|$. It follows that

$$A = V^*VAV^*V = V^*U|VAV^*|V = V^*UV|A|V^*V = V^*UV|A|,$$

which proves the theorem because V^*UV is unitary. □

Theorem B.9 (Singular Value Decomposition) *For any $n \times n$ matrix A there exist unique real values $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_n \geqslant 0$ such that there exist $n \times n$ unitary matrices V, W with $A = VDW$, where $D = \operatorname{diag}(s_1, \ldots, s_n)$. Furthermore, $s_1, \ldots, s_n$ are the eigenvalues of $|A|$.*

The $s_1, \ldots, s_n$ are known as the *singular values* of A.

Proof. For uniqueness, if $A = VDW$ as above, then

$$|A| = \sqrt{A^*A} = \sqrt{W^*DV^*VDW} = \sqrt{W^*D^2W} = W^*\sqrt{D^2}W = W^*DW,$$

and so the diagonal entries of D must be the eigenvalues of $|A|$. For existence, the Polar Decomposition gives a unitary U such that $A = U|A|$. Since $|A| \geqslant 0$ (and hence is normal), there exists a unitary Y such that $|A| = YDY^*$, where $D = \operatorname{diag}(s_1, \ldots, s_n)$ for some $s_1 \geqslant \cdots \geqslant s_n \geqslant 0$. Then $A = U|A| = UYDY^*$. Setting $V := UY$ and $W := Y^*$ proves the theorem. □

B.4 Sterling's Approximation

Theorem B.10 (Sterling's Approximation) $n! \sim \sqrt{2\pi n}(n/e)^n$.

Here, $f(n) \sim g(n)$ means that $\lim_{n\to\infty} f(n)/g(n) = 1$.

We'll prove a slightly weaker version of Theorem B.10 that nevertheless suffices for all our purposes, namely,

Theorem B.11 (Weak Sterling) *For all positive integers n,*

$$\frac{e}{\sqrt{2}}\sqrt{n}\left(\frac{n}{e}\right)^n \leqslant n! \leqslant e\sqrt{n}\left(\frac{n}{e}\right)^n.$$

Proof. We start with an integral approximation. The theorem clearly holds for $n = 1$, so assume $n \geqslant 2$. Since the log function is concave downward, we claim that for all i such that $2 \leqslant i \leqslant n$,

$$\frac{\log i + \log(i-1)}{2} \leqslant \int_{i-1}^{i} \log x \, dx \leqslant \log i - \frac{1}{2i}. \tag{138}$$

The left-hand side is the area of the trapezoid T_1 formed by the points $(i-1, 0), (i, 0), (i, \log i), (i-1, \log(i-1))$, and the right-hand side is the area of the trapezoid T_2 formed by the points $(i-1, 0), (i, 0), (i, \log i), (i-1, \log i - 1/i)$. Note that T_2's upper edge is the tangent line to the curve $y = \log x$ at the point $(i, \log i)$. By concavity of log, the region under the curve $y = \log x$ in the interval $[i-1, i]$ contains T_1 and is contained in T_2, hence the inequalities (138).

Now note that $\log(n!) = \sum_{i=1}^{n} \log i = \sum_{i=2}^{n} \log i$. Summing (138) from $i = 2$ to n and simplifying, we get

$$\log(n!) - \frac{\log n}{2} \leqslant \int_1^n \log x \, dx = n\log n - n + 1 \leqslant \log(n!) - \frac{1}{2}\sum_{i=2}^{n}\frac{1}{i}, \tag{139}$$

using the closed form $\int \log x \, dx = x \log x - x + C$. The sum on the right-hand side of (139) is the Harmonic series, which satisfies another integral approximation:

$$\sum_{i=2}^{n} \frac{1}{i} \geqslant \int_2^n \frac{dx}{x} = \log n - \log 2. \tag{140}$$

Equations (139) and (140) yield

$$\log n! - \frac{\log n}{2} \leqslant n \log n - n + 1 \leqslant \log(n!) - \frac{\log n}{2} + \frac{\log 2}{2},$$

and so

$$n \log n - n + 1 + \frac{\log n}{2} - \frac{\log 2}{2} \leqslant \log n! \leqslant n \log n - n + 1 + \frac{\log n}{2}. \tag{141}$$

Taking e to the power of all three quantities in (141) and simplifying, we have

$$\frac{e}{\sqrt{2}} \sqrt{n} \left(\frac{n}{e}\right)^n \leqslant n! \leqslant e \sqrt{n} \left(\frac{n}{e}\right)^n$$

as desired. □

B.5 Inequalities of Markov and Chebyshev

We only consider random variables that are real-valued and over discrete sample spaces. If X is such a random variable, then we let $E[X]$ and $\text{var}[X]$ respectively denote the expected value (mean) of X and the variance of X.

Theorem B.12 (Markov's Inequality) *Let X be a random variable with finite mean, and suppose $X \geqslant 0$. For every real $c > 0$,*

$$\Pr[X \geqslant c] \leqslant \frac{E[X]}{c}.$$

Proof. Let Ω be the sample space for X. We have

$$\begin{aligned}
E[X] &= \sum_{a \in \Omega} X(a) \Pr[a] \\
&= \sum_{a : X(a) \geqslant c} X(a) \Pr[a] + \sum_{a : X(a) < c} X(a) \Pr[a] \\
&\geqslant \sum_{a : X(a) \geqslant c} X(a) \Pr[a] \\
&\geqslant \sum_{a : X(a) \geqslant c} c \Pr[a] \\
&= c \Pr[X \geqslant c].
\end{aligned}$$

Dividing both sides by c proves the theorem. □

Theorem B.13 (Chebyshev's Inequality) *Let X be a random variable with finite mean μ and variance σ^2, and let $a > 0$ be real.*

$$\Pr[\,|X - \mu| \geqslant a\,] \leqslant \frac{\sigma^2}{a^2}.$$

Proof. We invoke Markov's Inequality with the random variable $Y = (X - \mu)^2$, letting $c = a^2$. Note that $Y \geqslant 0$, $\mathrm{E}[Y] = \sigma^2$, and $\Pr[\,|X - \mu| \geqslant a\,] = \Pr[Y \geqslant a^2]$. □

B.6 Relative Entropy

Let $p = (p_1, p_2, \ldots)$ and $q = (q_1, q_2, \ldots)$ be two probability distributions over some (finite or infinite) discrete sample space $\{1, 2, \ldots\}$. The *relative entropy of* q *with respect to* p is defined as

$$\mathrm{D}(p \,\|\, q) = \sum_i p_i \lg \frac{p_i}{q_i}, \tag{142}$$

Where the sum is taken over all i such that $p_i > 0$. If $q_i = 0$ and $p_i > 0$ for some i, then $\mathrm{D}(p \,\|\, q) = \infty$. Otherwise, the sum in (142) may or may not converge, but we always have the following regardless:

Theorem B.14 $\mathrm{D}(p \,\|\, q) \geqslant 0$*, with equality holding if and only if* $p = q$.

Proof. We use that fact that $\log x \leqslant x - 1$ for all $x > 0$, with equality holding iff $x = 1$. We have

$$\begin{aligned}
\mathrm{D}(p \,\|\, q) &= \sum_i p_i \lg \frac{p_i}{q_i} \\
&= -\sum_i p_i \lg \frac{q_i}{p_i} \\
&= -\frac{1}{\log 2} \sum_i p_i \log \frac{q_i}{p_i} \\
&\geqslant -\frac{1}{\log 2} \sum_i p_i \left(\frac{q_i}{p_i} - 1 \right) \\
&= \frac{1}{\log 2} \sum_i (p_i - q_i) \\
&= \frac{1}{\log 2} \left(1 - \sum_i q_i \right) \\
&\geqslant 0.
\end{aligned}$$

It is easy to see that equality holds above if and only if $p = q$. □

$\mathrm{D}(p \,\|\, q)$ is also known as the *Kullback-Leibler divergence* of p and q.

An important special case is when $q = (q_1, \ldots, q_n) = (1/n, \ldots, 1/n)$ is the uniform distribution on a sample space of size n (and $p = (p_1, \ldots, p_n)$ is arbitrary). In this case, we have

$$\mathrm{D}(p \,\|\, q) = \lg n - \mathrm{H}(p_1, \ldots, p_n). \tag{143}$$

If $(p, 1-p)$ and $(q, 1-q)$ are binary distributions, then we abbreviate $D((p, 1-p) \,\|\, (q, 1-q))$ by $d(p \,\|\, q)$, and we call $d(\cdot \,\|\, \cdot)$ the *binary relative entropy* function. Note that by (143), $d(p \,\|\, 1/2) = 1 - h(p)$.

B.7 A Standard Tail Inequality

It might be necessary to read Section B.6 before this one. In this section we give an upper bound on left tail the cumulative distribution function for the binomial distribution.

Let $0 < p < 1$ and let $n > 0$ be an integer. In this section, we give an upper bound for the sum $\sum_{i=0}^{t} \binom{n}{i} p^i (1-p)^{n-i}$, where $t \leqslant pn$. [For example, this sum is the probability of getting at most t heads among n flips of a p-biased coin (i.e., n identical Bernoulli trials with bias p). The expected number of heads among n flips is pn, and we want to show that the probability of getting significantly fewer than pn heads diminishes exponentially with n.]

Theorem B.15 *Let n be a positive integer. Let $0 < p < 1$ be arbitrary, and set $q = 1 - p$. If t is an integer such that $0 \leqslant t \leqslant pn$, then*

$$\sum_{i=0}^{t} \binom{n}{i} p^i q^{n-i} \leqslant 2^{-n\,d(t/n \,\|\, p)}, \tag{144}$$

where $d(\cdot \,\|\, \cdot)$ is the binary relative entropy defined in Section B.6.

Proof. If $t = 0$, then $d(t/n \,\|\, p) = d(0 \,\|\, p) = -\lg q$, and so both sides of (144) equal q^n and so the inequality is satisfied.

Now suppose $0 < t \leqslant pn$. Set $\lambda = t/n$, and let $\mu = 1 - \lambda$. Note that $0 < \lambda \leqslant p < 1$ and $0 < q \leqslant \mu < 1$. Define

$$C = \frac{p^t q^{n-t}}{\lambda^t \mu^{n-t}}.$$

For any $0 \leqslant i \leqslant t$, we have

$$p^i q^{n-i} = C \left(\frac{q}{p}\right)^{t-i} \lambda^t \mu^{n-t} \leqslant C \left(\frac{\mu}{\lambda}\right)^{t-i} \lambda^t \mu^{n-t} = C \lambda^i \mu^{n-i}.$$

Therefore, starting with the left-hand side of (144), we get

$$\sum_{i=0}^{t} \binom{n}{i} p^i q^{n-i} \leqslant C \sum_{i=0}^{t} \binom{n}{i} \lambda^i \mu^{n-i} \leqslant C \sum_{i=0}^{n} \binom{n}{i} \lambda^i \mu^{n-i} = C(\lambda + \mu)^n = C.$$

For the right-hand side of (144), we get

$$2^{-n\,d(t/n \,\|\, p)} = 2^{-n\,d(\lambda \,\|\, p)} = 2^{n[\lambda \lg(p/\lambda) + \mu \lg(q/\mu)]} = \left(\frac{p}{\lambda}\right)^{n\lambda} \left(\frac{q}{\mu}\right)^{n\mu} = \left(\frac{p}{\lambda}\right)^{t} \left(\frac{q}{\mu}\right)^{n-t} = C,$$

which proves the theorem. □