Quantum Information and Computation [*]

Zhiyuan Bai

Compiled on November 24, 2021

This document serves as a set of revision materials for the Cambridge Mathematical Tripos Part II course *Quantum Information and Computation* in Lent 2021. However, despite its primary focus, readers should note that it is NOT a verbatim recall of the lectures, since the author might have made further amendments in the content. Therefore, there should always be provisions for errors and typos while this material is being used.

Contents

[*] Based on the lectures under the same name taught by Prof. R. Jozsa in Lent 2021.

0 Overview

Why do we combine quantum mechanics with the theory of computation and information? First, what is information and what is computation? In a classical context, information exist as a string of Boolean variables ("bits"), and what computation does is a process of updating strings by a prescribed sequence of steps ("program"), and it does this by basic elementary Boolean operations ("gates") like AND, OR, NOT, SWAP, etc., with the property that each step takes fixed effort to perform, independent of the length of the string. But what precisely is a bit? In addition of being the unit of storage of a Boolean variable, it also has the property that we can identify the variable represented by it via distinguishing physical states (charge of an electron, etc.). Like R. Landauer said, "no information without representation." Consequently, we've come to the shocking conclusion that computation (and information processing) must corresponds to a physical evolution of the system representing the information. So, all possibilities and limitations of information storage, communication and processing must rest on laws of physics – not quite a popular view due to many reasons, but is somewhat based.

But of course, quantum physics is quite different from classical physics. It is true that, in principle, a quantum computer cannot compute anything that is not computable on a classical computer. The reason is quite simple: We can simulate Schrödinger's equation, hence any quantum system, with classical computer – however long it takes. Nonetheless, when we introduce quantum ideas to the "evolution of physical systems" that is computing, we can still achieve many more than classical computation.

First of all, quantum computers give possibilities of a stronger computing power, in both the space and time it needs to compute certain objects. For example, consider the following task: Given an integer N (with $n = O(\log N)$ digits), we want to find a factor of it quickly in the sense that the algorithm runs in polynomial time, i.e. the time needed to compute it is bounded by a polynomial of the "input size" n. One can use the obvious algorithm of trial division until $\sqrt{N}$, which however takes $O(\sqrt{N}) = O(2^{n/2})$ steps. Even the best known (classical) algorithm takes $2^{O(n^{1/3}(\log n)^{2/3})}$ steps, which is not in polynomial time. But on a quantum computer, what's known as Shor's algorithm can compute this is $O(n^3)$ time.

Secondly, quantum effects allow provably secure communication using quantum effects, which is totally impossible in classical contexts. One example of this is the BB84 quantum key distribution scheme. It also allows new kinds of commutation, e.g. quantum teleportation. We can also exploit the no-cloning theorem and quantum entanglements for further benefits.

Thirdly, it can get pass certain technological bottleneck. The so-called "Moore's Law" on the miniaturisation of computing components asserts a reduction of a factor of 4 every 3.5 years, a nice exponential trend that is hitting the wall after getting to the atomic level, where quantum physics kicks in. To seek even more computing power, more industrial sectors and nations are trying to build a quantum computer in order to get pass this barrier.

1 Principles of Quantum Mechanics

Let V be a finite dimensional complex vector space with a Hermitian inner product. The vectors in V are written as $|v\rangle$ and are called ket vectors (or simply ket). We often work with a 2-dimensional space $V = V_2$ with a chosen orthonormal basis $\{|0\rangle, |1\rangle\}$ (the "computational basis", "normal basis" or Z-basis). As a convention, kets are already written as column vectors with respect to any prescribed basis, so

$$a|0\rangle + b|1\rangle = \begin{pmatrix} a \\ b \end{pmatrix}$$

The conjugate transpose of a ket vector is called a bra vector $\langle v| = |v\rangle^\dagger$, which is of course a row vector. More formally, we can identify the bra vector $\langle v|$ as an element of the dual vector space V^* of V, namely the image of $|v\rangle$ under the canonical isomorphism $V \cong V^*$ of complex inner product spaces given by

$$|w\rangle \mapsto \text{inner product of } |v\rangle \text{ and } |w\rangle$$

Consequently, the inner product of $|v\rangle = a|0\rangle + b|1\rangle$ and $|w\rangle = c|0\rangle + d|1\rangle$ is

$$\langle v|w\rangle = \langle v|(|w\rangle) = |v\rangle^\dagger |w\rangle = \begin{pmatrix} a^* & b^* \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = a^*c + b^*d$$

So for example $\langle i, j\rangle = \delta_{ij}$ becomes the statement of orthonormality.

Definition 1.1. Suppose V, W are inner product spaces with orthonormal bases $\{|e_1\rangle, \ldots, |e_m\rangle\}, \{|f_1\rangle, \ldots, |f_n\rangle\}$, we construct their tensor product $V \otimes W$ as the dimension mn vector space with orthonormal basis $\{|e_i\rangle \otimes |f_j\rangle\}_{i,j}$.

We can of course extend $\otimes$ bilinearly, which gives a natural bilinear map $f : V \times W \to V \otimes W$ via $(|\alpha\rangle, |\beta\rangle) \mapsto |\alpha\rangle \otimes |\beta\rangle$ which is not surjective. We often abbreviate $|\alpha\rangle \otimes |\beta\rangle$ as $|\alpha\rangle|\beta\rangle$. Note that $|\alpha\rangle|\beta\rangle = |\beta\rangle|\alpha\rangle$ is almost never true; note also that f is almost never surjective.

Definition 1.2. Any $|\xi\rangle = |\alpha\rangle \otimes |\beta\rangle \in V \otimes W$ is called a product vector. Any $|\xi\rangle \in V \otimes W$ that is not a product vector is called an entangled vector.

Of course we take take tensor products again and again since it is easily seen to be associative. For a tensor product $V_1 \otimes \cdots \otimes V_n$, we set the product vectors to be vectors of the form $|\alpha_1\rangle \otimes \cdots \otimes |\alpha_n\rangle$ – that is, we want the notion of product vectors to depend on the specific (tensor) decomposition we are considering.
We will mostly be interested in k-fold tensor products $V_2^{\otimes n}$ of V_2 with itself, which has dimension 2^k and orthonormal basis $|i_1\rangle \otimes \cdots \otimes |i_k\rangle$ with $i_1, \ldots, i_k \in \{0, 1\}$. We often denote such a vector alternatively as $|i_1\rangle \cdots |i_k\rangle$ or simply $|i_1 \cdots i_k\rangle$.

Example 1.1. $|v\rangle = |00\rangle + |11\rangle \in V_2 \otimes V_2$ is entangled. Indeed, if $|v\rangle = (a|0\rangle + b|1\rangle)(c|0\rangle + d|1\rangle) = ac|00\rangle + ad|01\rangle + bc|01\rangle + bd|11\rangle$, so necessarily $ad = bc = 0, ac = bd = 1$ which means $0 = adbc = acbd = 1$, contradiction.
In general, one can show that $|v\rangle = a_{00}|00\rangle + a_{01}|01\rangle + a_{10}|10\rangle + a_{11}|11\rangle$ is entangled iff $\det(a_{ij}) \neq 0$. For higher dimensions, suppose $|1\rangle, \ldots, |m\rangle$ generate V and $|1\rangle, \ldots, |n\rangle$ generate W, then $\sum_{i,j} A_{ij}|i\rangle|j\rangle$ is product iff (A_{ij}) has rank 1.

Can we let $V \otimes W$ inherit an inner product? For product vectors $|\alpha_1\rangle |\beta_1\rangle$ and $|\alpha_2\rangle |\beta_2\rangle$, we can set their inner product to be $(\langle\beta_1| \langle\alpha_1|)(|\alpha_2\rangle |\beta_2\rangle) = \langle\alpha_1|\alpha_2\rangle \langle\beta_1|\beta_2\rangle$ (caution: we usually write tensor product of bra vectors the other way around). As product vectors include an orthonormal basis for $V \otimes W$, this extends to entangled states as well.

Postulate (QM1)**.** *State of any isolated physical system S are represented by unit vectors in a complex vector space V with an inner product.*

The simplest nontrivial case is the 2-dimensional space $V = V_2$. Fix a pair of orthonormal vectors $|0\rangle, |1\rangle$, then unit vectors in it are $|\psi\rangle = a|0\rangle + b|1\rangle, a, b \in \mathbb{C}, |a|^2 + |b|^2 = 1$. In this case, we say $|\psi\rangle$ is a superposition of $|0\rangle$ and $|1\rangle$ with amplitudes a, b respectively.
We can alternatively use the conjugate basis (or X-basis), namely

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

The Z-basis and the X-basis are called conjugate or mutually biased. In general,

Definition 1.3. A set of mutually unbiased bases in $\mathbb{C}^d$ is a set orthonormal bases such that for any distinct bases $\{e_i\}_i, \{f_j\}_j$ in this set we have $|\langle e_i|f_j\rangle|^2 = 1/d$ for all i, j.

An interesting quantity is $N = N(d)$ which is the maximum number of mutually unbiased bases in $\mathbb{C}^d$. We have $N(2) = 3$ and it's known in general that if d is a prime power then $N(d) = d+1$. However, N is very much unknown for other d. We don't even know $N(6)$, except that $3 \leq N \leq 7$. In general, for $d = p_1^{n_1} \cdots p_k^{n_k}, p_1 < \cdots < p_k$, we know that $p_1^{n_1} + 1 \leq N(d) \leq d+1$ which isn't exactly a very good bound.

Definition 1.4. A qubit is any quantum system with 2-dimensional state space.

Postulate (QM2)**.** *If systems S_1, S_2 have state space V_1, V_2 repsectively, then the joint system S_1S_2 has state space $V_1 \otimes V_2$.*

Example 1.2. Consider a system with n qubits, then the state space $V_2 \otimes n$ of this system has dimension 2^n with standard basis $|i_1 \cdots i_n\rangle$ labeled by a binary string of length n. A state $|\psi\rangle$ is a product state if it is a tensor product of n 1-qubit space (where we've generalised the notion of product state to keep track of how we obtained the product), otherwise it is entangled.

Remark. In classical physics, S_1S_2 has "state space" $V_1 \times V_2$, i.e. there is no correspondng notion of "entangled states" and the dimension (implying complexity) of the state space grows linearly instead of exponentially upon composition.

Let $|v\rangle, |w\rangle \in V$, we can form the ket-bra product $|v\rangle \langle w|$ which is then a rank 1 linear map $V \to V$ with $|v\rangle \langle w| (|x\rangle) = |v\rangle \langle w|x\rangle$. Note in particular that $\langle v|w\rangle = \operatorname{tr}(|v\rangle \langle w|)$
For a general linear map $A : V \to V$ with matrix (A_{ij}) (with indices starting at 0), we can always write $A = \sum_{i,j} A_{ij} |i\rangle \langle j|$, i.e. $\{|i\rangle \langle j|\}$ gives a basis for $V \otimes V^* \cong L(V, V)$.
Suppose $|v\rangle \in V$ is normalised, then $\pi_v = |v\rangle \langle v|$ is the projection map onto the (1-dimensional) subspace spanned by $|v\rangle$. In particular, π_v is idempotent

by either the general properties of projections or direct calculation $\pi_v \pi_v = |v\rangle \langle v|v\rangle \langle v| = |v\rangle \langle v| = \pi_v$. In general, if E is a d-dimensional subspace of V and $\{|e_1\rangle, \dots |e_d\rangle\}$ any orthonormal basis of E, then $\pi_E = |e_1\rangle \langle e_1| + \dots + |e_d\rangle \langle e_d|$ is the projection operator onto E, which can easily seen to be independent of the specific choice of basis.

We can also form the tensor product of linear maps. Suppose $A : V \to V', B : W \to W'$ are linear maps, then we define their tensor product $A \otimes B : V \otimes W \to V' \otimes W'$ to be the extension of $(A \otimes B)(|v\rangle |w\rangle) = (A|v\rangle)(B|w\rangle)$. In particular, if $A : V_2 \to V_2, B : V_2 \to V_2$, then $A \otimes B$ has matrix

$$(A \otimes B)_{ij} = \begin{pmatrix} A_{00}B_{00} & A_{00}B_{01} & A_{01}B_{00} & A_{01}B_{01} \\ A_{00}B_{10} & A_{00}B_{11} & A_{01}B_{10} & A_{01}B_{11} \\ A_{10}B_{00} & A_{10}B_{01} & A_{11}B_{00} & A_{11}B_{01} \\ A_{10}B_{10} & A_{10}B_{11} & A_{11}B_{10} & A_{11}B_{11} \end{pmatrix}$$

under the standard bases. Let I be the identity, then $A \otimes I, I \otimes A$ are the actions of A on the first and second components space, respectively.

Postulate (QM3)**.** *Any physical evolution of a quantum system is represented by a unitary linear operator.*

This is our version of Schrödinger's equation. Recall that Schrödinger's equation asserts that $i\hbar |\psi\rangle_t = H |\psi\rangle$ where H is the Hamiltonian which is assumed to be Hermitian. When H is independent of time, we can write down the solution $|\psi\rangle = e^{(-it/\hbar)H} |\psi_0\rangle$ which turns out to be unitary.

In case you forgot, $U : V \to V$ is unitary iff $UU^\dagger = I$ iff U preserves orthonormality iff the columns and rows of the matrix of U (under any orthonormal basis) form orthonormal bases.

We shall introduce yet another notation here: Any $|v\rangle \in V$ defines a linear map $V \otimes W \to W$ via $|a\rangle |b\rangle \mapsto \langle v|a\rangle |b\rangle$. This is known as the partial inner product with $|v\rangle$. We write $\langle v|\psi\rangle$ to denote the image of $|\psi\rangle \in V \otimes W$ under this partial inner product. This is also called the conditional state (or relative state) of $|\psi\rangle$ given v.

Example 1.3. For $V = V_2$ and $|\xi\rangle = a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle \in V \otimes V$, the partial inner product of $|0\rangle$ on $|\xi\rangle$ is

$$\langle 0|\xi\rangle = a\langle 0|0\rangle |0\rangle + b\langle 0|0\rangle |1\rangle + c\langle 0|1\rangle |0\rangle + d\langle 0|1\rangle |1\rangle = a|0\rangle + b|1\rangle$$

One can also perform the partial inner product with $|0\rangle$ in the second component in the tensor product, which gives $\langle 0|\xi\rangle = a|0\rangle + c|1\rangle$.

Postulate (QM4.1: Basic Born Rule)**.** *Consider a single instance of a quantum state $|\xi\rangle$ for a physical system with state space V of dimension n.*

1. Let $B = \{|e_1\rangle, \dots, |e_n\rangle\}$ be any orthonormal basis of V, and $|\psi\rangle = \sum_i a_i |e_i\rangle$ a state in V. When we make a measurement on $|\psi\rangle$ relative to the basis B (a "complete" measurement) where the possible outcomes are $j \in \{1, \dots, n\}$, the probability $\mathbb{P}(j)$ of the outcome j equal to $|a_j|^2$. If outcome j is seen, then the state collapses from $|\psi\rangle$ to $|e_j\rangle$.

2. Let $\{E_1, \dots, E_d\}$ be decompositions of V into mutually orthogonal subspaces (so $V = E_1 \otimes^\perp \cdots \otimes^\perp E_d$). If we make a measurement on $|\psi\rangle$ relative to this decomposition (an "incomplete" measurement) with possible outcomes

$j \in \{1, \ldots, d\}$, then the probability for outcome j is $\mathbb{P}(j) = \langle\psi| \pi_j^\dagger \pi_j |\psi\rangle = \langle\psi| \pi_j |\psi\rangle$ where π_j is the projection operator onto V_j. Post-measurement state ("collapsed" vector) after the measurement j is obtained is the value $|\psi_j\rangle = \pi_k |\psi\rangle / \sqrt{\mathbb{P}(j)}$.

Example 1.4 (Parity measurement on 2-qubits)**.** The parity of a 2-bit string b_1b_2 of qubits is $b_1 + b_2 \pmod 2$. The parity measurement on this system of 2 qubits is an incomplete measurement given by the orthogonal decomposition $E_0 = \text{span}\{|00\rangle, |11\rangle\}, E_1 = \text{span}\{|01\rangle, |10\rangle\}$. For state $|\psi\rangle = a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle$, the probability of the observation 0 is $\mathbb{P}(0) = |a|^2 + |d|^2$ and the post-measurement state of 0 is $(a|00\rangle + d|11\rangle)/\sqrt{|a|^2 + |d|^2}$

There is also a notion of the measurement of "quantum observables".

Definition 1.5. A quantum observable θ is a Hermitian operator on V.

By general spectral theory, we know that θ has real eigenvalues and V is an orthogonal decomposition of eigenspaces Λ_j of θ. The measurement of a quantum observable θ is just the incomplete measurement of Λ_j, which usually are labelled by the eigenvalues λ_j (instead of the indices) instead – which however is not the convention we want to use.

Postulate (QM4.2: Extended Born Rule)**.** *Suppose $|\psi\rangle$ is a state in the state space $V \otimes W$ in the composite system S_1S_2. Let $B_V = \{|e_1\rangle, \ldots, |e_n\rangle\}$ be an orthonormal basis of V. Suppose $|\psi\rangle = \sum_i |e_i\rangle |\xi_j\rangle$ for $\{\xi_j\}$ not necessarily normalised nor orthogonal (in fact $|\xi_j\rangle = \langle e_i|\psi\rangle$). As $|\psi\rangle$ is normalised, we have $\sum_i \langle\xi_i|\xi_i\rangle = \langle\psi|\psi\rangle = 1$.*
Suppose we perform a measurement of $|\psi\rangle$ relative to the basis B_V with outcomes labelled by $i \in \{1, \ldots, n\}$, i.e. with respect to the orthogonal decomposition $V = \bigoplus_i^\perp E_i, E_i = \text{span}\{e_i \otimes |\xi\rangle : |\xi\rangle \in W\}$, then the corresponding projections are $\pi_u = |e_i\rangle\langle e_i| \otimes I_W$ and the probability of getting result i is $\langle\xi_i|\xi_i\rangle$ with post-measurement state $|\psi_i\rangle = |e_i\rangle |\xi_i\rangle / \sqrt{\langle\xi_i|\xi_i\rangle}$.

Remark. 1. According to QM4 (i.e. putting QM4.1 and QM4.2 together), two different states with guaranteed (i.e. true with probability q) different outcomes for some measurement must be orthogonal. Non-orthogonal states, although physically distinct, cannot be reliably distinguished by any quantum processes.
2. If $|\psi\rangle$ has dimension n, then any measurement has at most n outcomes – but we can get more outcomes by adjoining an ancilla $|A\rangle$ independent of $|\psi\rangle$ and measuring $|\psi\rangle |A\rangle$ jointly.

2 Quantum Information

2.1 Quantum Information and Quantum Process

Information in the classical sense is just a distinguishable state of a physical system. Classically, any two distinct physical states must be distinguishable. However, two quantum states are distinguishable if and only if they are orthogonal. For example, a qubit is then the simplest system that can reliably encode a classical bit (even though it has infinitely many different states) since it has the smallest possible dimension that allows an orthogonal pair of states.

For a general quantum state $|\psi\rangle$, we cannot really identify it from any observation with certainty due to QM4, so it is not really a piece of classical information. But our quantum formality allows such distinction of quantum states, even though we cannot access it physically. So we call the "information" that's represented by a quantum state a quantum information.
Given some quantum information $|\psi\rangle$, we can perform some processes on it by way of operations within one of three categories:

Definition 2.1. The ancilla operation involves taking a known quantum state $|A\rangle$ (called the ancilla) and adjoining it to get a new state $|\psi\rangle |A\rangle$.
The unitary operation is, well, a unitary operation, i.e. applying a unitary operator U of our choice on $|\psi\rangle$.
The measure operation is a (possibly incomplete) measurement on $|\psi\rangle$, recording the result and using the post-measurment state for further quantum information purposes.

If we do use the post-measurement state for further purposes, we usually choose the further operations adaptively depending on the outcome. So the final output is generally a probability mixture over successive measurements and/or other operations.
One important class of unitary operations is quantum gates.

Definition 2.2. A qubit gate is a unitary operation on qubits.

We first look at 1-qubit gates.

Example 2.1. 1. The Hadamard gate

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

with the defining property that $H|0\rangle = |+\rangle, H|1\rangle = |-\rangle$. Note also that $H^2 = I$ and we might view H as a reflection across the "line with inclination $\pi/8$".
2. The Pauli gates are

$$I = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \sigma_x = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \sigma_y = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}, \sigma_z = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

which, notably, are also Hermitian in addition of being unitary. All of them have eigenvalues ±1 and they satisfy $\sigma_x\sigma_y = -\sigma_y\sigma_z = i\sigma_z$ (up to cyclic permutations of the symbols x, y, z). We write $X = \sigma_x, Z = \sigma_z, Y = i\sigma_y$ because (according to the lecturer) complex-valued matrices intimidate computer scientists.
We have $X|0\rangle = |1\rangle, X|1\rangle = |0\rangle$ or, to use a shorthand notation that, as per usual, does not make life easier at all, $X|k\rangle = |k \oplus 1\rangle$ where $\oplus$ denotes addition modulo 2. This inspires us (well, some people at least) to call it the "quantum not" operation. The eigenstates of X are precisely the conjugate basis $|+\rangle, |-\rangle$.
As for Z, we have $Z|0\rangle = |0\rangle, Z|1\rangle = -|1\rangle$ or $Z|k\rangle = (-1)^k|k\rangle$ and Z has eigenstates $|0\rangle, |1\rangle$. This is the reason why we call $|0\rangle, |1\rangle$ the Z-basis and $|+\rangle, |-\rangle$ the X-basis.
3. The phase gate with parameter θ is

$$P(\theta) = \begin{pmatrix}1 & 0\\ 0 & e^{i\theta}\end{pmatrix}$$

with the special value $Z = P(\pi)$.

How about 2-qubit gates?

Example 2.2. 1. The controlled-X (or controlled-NOT, CX, C-NOT) is given by (the extension of) $CX\,|j\rangle\,|k\rangle = |j\rangle\,|j \oplus k\rangle = |j\rangle\,X^j\,|k\rangle$ (for $j, k \in \{0, 1\}$). So $CX\,|0\rangle\,|\alpha\rangle = |0\rangle\,|\alpha\rangle\,, CX\,|1\rangle\,|\alpha\rangle = |1\rangle\,X\,|\alpha\rangle$ and CX has the matrix

$$CX = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

under the standard basis.
For $CX\,|\alpha\rangle\,|\beta\rangle$ we call $|\alpha\rangle$ the control qubit and $|\beta\rangle$ the target qubit. Sometimes we might want to take the control qubit as the second entry. We keep track of this by writing $CX_{1,2}$ as this CX and $CX_{2,1}$ as $CX_{2,1}\,|j\rangle\,|k\rangle = CX_{1,2}\,|j\rangle\,|k\rangle$.
2. The controlled-Z is defined by $CZ\,|j\rangle\,|k\rangle = |j\rangle\,Z^j\,|k\rangle = (-1)^{jk}\,|j\rangle\,|k\rangle$. Notably, this is symmetric in j, k, so we don't have to specify a control qubit. It has matrix

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

2.2 No-Cloning Theorem

Shockingly, we can't actually duplicate a piece of unknown quantum information, i.e. performing any process of the form $(b)(0)(M) \mapsto (b)(b)(M_b)$. In particular, we can't even perform CNOT (i.e. $b_1 b_2 \mapsto b_1(b_1 \oplus b_2)$) on quantum information (note that the gate CX we discussed earlier does not achieve this, in case you're wondering). To make these precise,

Definition 2.3. Given quantum information $|\psi\rangle_A$ (the subscript A means that the register/space A carries the information), a cloning process is a quantum evolution on registers A, B, M (with $\dim A = \dim B$) that achieves

$$|\psi\rangle_A\,|\psi_0\rangle_B\,|M_0\rangle_M \mapsto |\psi\rangle_A\,|\psi\rangle_B\,|M_\psi\rangle_M$$

where $|\psi_0\rangle_B$ is any chosen fixed known state in B.

Theorem 2.1 (No-Cloning Theorem)**.** *Let S be any (known) set of quantum states of A that contains at least one pair $|\xi\rangle \neq |\eta\rangle$ of non-orthogonal states, then no unitary process that achieves the cloning of all states in S exists.*

Remark. 1. The theorem remains true if we replace "unitary process" by "sequence of arbitrary quantum operations" (where, in the case of measure, we will insist the cloning to occur on all probability branches). For ancilla, this is easy to see by regarding the ancilla as part of M. For measure, the situation is slighly complicated but one can use a trick of replacing the measurement collapse with the introduction of a further quantum register for the measurement pointer with orthonormal basis labelled by the measurement outcome.
2. If $|\xi\rangle = |\eta\rangle$ is known, then we can trivially clone it. If $|\xi\rangle \perp |\eta\rangle$ are known, then we can rotate them to $|0\rangle\,, |1\rangle$, clone them with CX, and then rotate back.

Proof. Let $|\xi\rangle \neq |\eta\rangle$ be non-orthogonal states and suppose we can clone them with a unitary process U. Then the process must achieve $|\xi\rangle_A |\psi_0\rangle_B |M_0\rangle_M \mapsto |\xi\rangle_A |\xi\rangle_B |M_\xi\rangle_M, |\eta\rangle_A |\psi_0\rangle_B |M_0\rangle_M \mapsto |\eta\rangle_A |\eta\rangle_B |M_\eta\rangle$.
Since unitary operators preserve inner products, we have

$$\langle\xi|\eta\rangle \langle\psi_0|\psi_0\rangle \langle M_0|M_0\rangle = \langle\xi|\eta\rangle \langle\xi|\eta\rangle \langle M_\xi|M_\eta\rangle$$

Taking absolute values, we see that $1 = |\langle\xi|\eta\rangle| |\langle M_\xi|M_\eta\rangle|$ as ξ and η are not orthogonal. But $|\langle\xi|\eta\rangle| < 1$, so $|\langle M_\xi|M_\eta\rangle| > 1$ which is already a contradiction to Cauchy-Schwarz inequality. □

Remark. The no-cloning theorem was formulated in 1982, by Wootters and Zurek and by Dieks (independently), although it was recently discovered that the theorem was already proved (in a different context) by Park (1970). The work of Wootters, Zurek and Dieks are to refute a proposal of Herbert (1980) for superluminal signaling in quantum mechanics.
Herbert's (failed) attempt at this is as follows: Suppose Alice (A) and Bob (B) are distantly separated in space, each holding a qubit such that the composite system of the two qubits is in the entangled state

$$|\varphi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A |0\rangle_B + |1\rangle_A |1\rangle_B) = \frac{1}{\sqrt{2}}(|+\rangle_A |+\rangle_B + |-\rangle_A |-\rangle_B)$$

Notably, this is a situation that is known to be experimentally possible to produce. A wants to communicate a yes/no decision at noon, with B expecting it. For the "yes" decision, she measures her qubit in the computational basis $\{|0\rangle, |1\rangle\}$ (i.e. measures the whole system in the decomposition $V_2 \otimes V_2 = (|0\rangle_A \otimes V_2) \oplus^\perp (|1\rangle_A \otimes V_2)$); For the "no" decision, she measures it in the conjugate basis $\{|+\rangle, |-\rangle\}$. Then the (extended) Born rule tells us that the quantum state of Bob's qubit instantaneously collapses, to one of $|0\rangle, |1\rangle$ (each with probability 1/2 depends on A's measurement) in the case of "yes" and $|+\rangle, |-\rangle$ (ditto) in the case of "no".
These are all valid, hence, as one can expect, doesn't really give Bob any information at all no matter how he measures his qubit now. Indeed, suppose π_i is the projection operator for the outcome i in Bob's measurement, then the probability of outcome i happening in the "yes" case is the trace of $\pi_i((|0\rangle\langle 0| + |1\rangle\langle 1|)/2)$; And that in the "no" case is the trace of $\pi_i((|+\rangle\langle +| + |-\rangle\langle -|)/2)$. These two probabilities, sadly, are the same. So A's attempt of transmitting information is in vain.
However, suppose (falsely) that B can clone his qubit. He does it at time noon$+\epsilon$ to make a lot of copies of his qubit. In the "yes" case, either all qubits are $|0\rangle$ or all are $|1\rangle$; Similarly, in the "no" case, either all qubits are $|+\rangle$ or all are $|-\rangle$. These can, now, be distinguished with high probability by measuring all these qubits in $\{|0\rangle, |1\rangle\}$. In the "yes" case, he will see either all zeros or all ones; In the "no" case, he will see (with the number of qubit copies approaching ∞) approximately half zeros and half ones. So he can read A's message with probability arbitrarily close to 1.

2.3 Distinguish Non-Orthogonal States

How would we attempt to distinguish non-orthogonal states with a quantum process? As we've discussed, we can't do it perfectly. But how well can we do?

Consider the following situation: We have a unknown quantum state $|\psi\rangle$ (say in some Hilbert space H of dimension d) that is either $|\alpha_0\rangle$ or $|\alpha_1\rangle$. The problem is to determine whether $|\psi\rangle = |\alpha_0\rangle$ or $|\psi\rangle = |\alpha_1\rangle$. When $\langle\alpha_0|\alpha_1\rangle \neq 0$, we already know that we cannot do this with certainty. On the other hand, we can also solve this task with probability $1/2$ by, well, tossing a coin. So we will attempt to do at least better than that, especially when $|\alpha_0\rangle, |\alpha_1\rangle$ are already "orthogonal enough" as measured by a small $\langle\alpha_0|\alpha_1\rangle$ (as we will see later, this is the only quantity that affects our ability to distinguish them).

Suppose we have a quantum process on $|\psi\rangle$ for this task, then by pushing all ancilla to the start and delaying all measurements to the end (there's a slight technicality here which we shall omit, but it should be easy to sketch how to do this), we can assume WLOG that the process is simply adjoining an ancilla $|A\rangle$ to $|\psi\rangle$, applying a unitary operation U to it, and then measure it.

The ancilla operation solely increases the dimension of the space, since all it does is replacing the problem of discriminating $|\alpha_0\rangle$ and $|\alpha_1\rangle$ with the problem of discriminating $|\alpha_0\rangle|A\rangle$ and $|\alpha_1\rangle|A\rangle$. $\langle\alpha_0|\alpha_1\rangle = \langle A|\langle\alpha_0|\alpha_1\rangle|A\rangle$, so this doesn't make them "more orthogonal" either. Consequently, we can replace the first operation with the assumption that the dimension of $|\psi\rangle$ is big enough.

In addition, if we apply a unitary operator U and then measure it (say with projection operator π_i for outcome i), then it can easily seen to be equivalent (in the sense that the probabilities of outcomes stay the same) to simply a measurement operation with projection operators $U^\dagger \pi_i U$.

So we can in fact assume WLOG that this quantum process is just a measurement on the quantum state which can be assumed to have arbitrarily large dimension. We also don't really care about the post-measurement state for this particular problem, so we can assume that the measurement has two possible outcomes 0 (i.e. $|\psi\rangle = |\alpha_0\rangle$) and 1 (i.e. $|\psi\rangle = |\alpha_1\rangle$) with projection operators π_0, π_1. The task now is to choose the best pair of projections π_0, π_1 that does the best job.

Definition 2.4. The success probability in the settings above is defined as

$$\begin{aligned}
P_s &= \frac{1}{2}\mathbb{P}(\text{Output } 0 \mid \text{Input } |\alpha_0\rangle) + \frac{1}{2}\mathbb{P}(\text{Output } 1 \mid \text{Input } |\alpha_1\rangle) \\
&= \frac{1}{2}\langle\alpha_0|\pi_0|\alpha_0\rangle + \frac{1}{2}\langle\alpha_1|\pi_1|\alpha_1\rangle = \frac{1}{2}\langle\alpha_0|\pi_0|\alpha_0\rangle + \frac{1}{2}\langle\alpha_1|I - \pi_0|\alpha_1\rangle \\
&= \frac{1}{2} + \frac{1}{2}(\langle\alpha_0|\pi_0|\alpha_0\rangle - \langle\alpha_1|\pi_0|\alpha_1\rangle) = \frac{1}{2} + \frac{1}{2}\operatorname{tr}(\pi_0(|\alpha_0\rangle\langle\alpha_0| - |\alpha_1\rangle\langle\alpha_1|))
\end{aligned}$$

Optimising this expression clearly requires us to study $D = |\alpha_0\rangle\langle\alpha_0| - |\alpha_1\rangle\langle\alpha_1|$. Easily D is Hermitian, so it has real eigenvalues and there is an orthonormal basis of H consisting of eigenvectors of D. Also, whenever $|\beta\rangle$ is orthogonal to both $|\alpha_0\rangle$ and $|\alpha_1\rangle$, we have $D|\beta\rangle = 0$, so D can have at most two nonzero eigenvalues, and their corresponding eigenvectors have to be contained in $V = \operatorname{span}\{|\alpha_0\rangle, |\alpha_1\rangle\}$. We also have $\operatorname{tr} D = 0$, therefore these two possibly nonzero eigenvalues are $\pm\delta$ for some $\delta \geq 0$ (note that $\delta = 0$ iff $D = 0$ iff $|\alpha_0\rangle = e^{i\theta}|\alpha_1\rangle$ for some $\theta \in \mathbb{R}$). Consequently, there is some $|p\rangle, |m\rangle \in H$ nonzero with $D = \delta|p\rangle\langle p| - \delta|m\rangle\langle m|$.

What is δ? Say we are in the non-degenerate case that $\dim V = 2$, then since we already know that D is zero in $V^\perp$, it suffices to restrict everything to V. Choose a state $|\alpha_0^\perp\rangle$ orthogonal to $|\alpha_0\rangle$. Say $|\alpha_1\rangle = c_0|\alpha_0\rangle + c_1|\alpha_0^\perp\rangle$, then D

would have the matrix

$$\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1 & 0\end{pmatrix}-\begin{pmatrix}c_0\\c_1\end{pmatrix}\begin{pmatrix}c_0^* & c_1^*\end{pmatrix}=\begin{pmatrix}|c_1|^2 & -c_0c_1^*\\ -c_1c_0^* & -|c_1|^2\end{pmatrix}$$

under the basis $\{|\alpha_0\rangle, |\alpha_0^\perp\rangle\}$. From here it's then easy to see that $\delta = \sin\theta$ where θ is chosen such that $|c_0| = |\langle\alpha_1|\alpha_0\rangle| = \cos\theta, |c_1| = \sin\theta$. Then

$$P_s = \frac{1}{2} + \frac{\sin\theta}{2}(\langle p|\pi_0|p\rangle - \langle m|\pi_0|m\rangle)$$

Now for any projector π and state $|\xi\rangle$, we have $\langle\xi|\pi|\xi\rangle \in [0,1]$, so P_s is maximised if we can find π_0 such that $\langle p|\pi_0|p\rangle = 1$ and $\langle m|\pi_0|m\rangle = 0$, in which case $P_s = (1+\sin\theta)/2 \in [1/2, 1]$. This is possible as we can simply choose π_0 to be the projection onto a subspace containing $|p\rangle$ and is orthogonal to $|m\rangle$ which is possible since $|p\rangle, |m\rangle$ are orthogonal. We conclude this result as the following theorem:

Theorem 2.2 (Helstrom-Holevo bound for pure states)**.** *Any quantum process can distinguish $|\alpha_0\rangle$ and $|\alpha_1\rangle$ with success probability at most*

$$\frac{1+\sin\theta}{2} = \frac{1+\sqrt{1-|\langle\alpha_1|\alpha_0\rangle|^2}}{2}$$

and the bound is achievable.

Our description of the process tells us that we don't really need to adjoint an ancilla at all. In particular, if $|\alpha_0\rangle$ and $|\alpha_1\rangle$ are qubits, we can work entirely in the 2-dimensional space and the optimal solution to do the discrimination will be a complete measurement of the quantum observable $D = |\alpha_0\rangle\langle\alpha_0| - |\alpha_1\rangle\langle\alpha_1|$.

Remark. There are other discrimination scenarios possible. For example, one can attempt a so-called unambiguous state discrimination, namely a measurement with outcomes $0, 1$ and "fail" where the outcomes $i = 0, 1$ indicates the certain result $|\psi\rangle = |\alpha_i\rangle$ (although the outcome "fail", which means exactly what it looks like, generally loses us all information about $|\psi\rangle$).

2.4 No-Signaling Principle

We are interested in the following scenario: Alice and Bob are distantly seperated in space with local quantum systems A, B. Initially, A, B are in some joint quantum state (which is usually known and likely entangled) and Alice and Bob can apply only local actions their own systems. Suppose Alice does a complete measurement on her system, then, by Born rule, for each measurement outcome k for A, the state at B will instantaneously change to a corresponding post-measurement state $|\beta_k\rangle$ due to measurement collapse. If Bob can notice this change, then we can in theory achieve some kind of superluminal signalling, so one would guess that he can't.
Let's formalise what we call a local action.

Definition 2.5. Let H_A, H_B be state spaces of A, B and let $H_{AB} = H_A \otimes H_B$. A local unitary operation U_A by Alice on H_A is represented by $U_A \otimes I_B$ on H_{AB}.

Adjoining local ancilla is simply enlarging their locally held system.
If Alice performs local measurement on A with orthogonal subspaces $\{E_a\}$ with projections $\{\pi_a\}$, it is of course a measurement on the joint system AB with subspaces $E_a \otimes H_B$ and projections $\pi_a \otimes I_B$.

Of course the situation is symmetric if we replace Alice by Bob.
Note that even if the local measurement of Alice is complete, the induced measurement on the whole system would still be incomplete (as long as $\dim H_B > 1$).
If Bob also does a local measurement with subspaces $\{F_b\}$ and projections $\{\varpi_b\}$, then the joint probability $P(a,b)$ obtained by performing both measurement turns out to be independent of who does the the measurement first or whether the measurements are done independently. This is because of the identity $(\pi_a \otimes I)(I \otimes \varpi_b) = (I \otimes \varpi_b)(\pi_a \otimes I) = \pi_a \otimes \varpi_b$.

Theorem 2.3 (No-Signaling Principle)**.** *No local action of Alice can change the output probability distribution of any local quantum process by Bob.*

We shall demonstrate the case when all Alice and Bob do are complete measurements.

Proof. Say Bob's local quantum process is a complete measurement wrt the bases $\{|b\rangle_B\}$ with the outcomes labelled by b.
For $|\psi\rangle_{AB} \in H_{AB}$, we can write $|\psi\rangle_{AB} = \sum_b |\xi_b\rangle_A |b\rangle_B$ where $|\xi_b\rangle_A = \langle b|\psi\rangle$ is the conditional state of $|\psi\rangle$ given b. Then the probability of outcome b is just $\langle \xi_b|\xi_b\rangle$.
Suppose now that Alice performs a complete measurement wrt the basis $\{|a\rangle\}$ with outcomes labelled by a. Write $\pi_a = |a\rangle\langle a|$ to denote the projection. Suppose Alice got the outcome a which happens with probability

$$\mathbb{P}(a) = \left\| \sum_b (\pi_a |\xi_b\rangle)_A |b\rangle_B \right\|^2$$

Then the post-measurement state of the joint system would be

$$|\psi_a\rangle_{AB} = \frac{1}{\sqrt{\mathbb{P}(a)}} \sum_b (\pi_a |\xi_b\rangle)_A |b\rangle_B$$

If Bob does his measurement after Alice's measurement, then the probability of him getting the outcome b is

$$\mathbb{P}(b) = \sum_a \mathbb{P}(a)\mathbb{P}(b \mid a) = \sum_a \mathbb{P}(a) \frac{\|\pi_a |\xi_b\rangle\|^2}{\mathbb{P}(a)} = \sum_a \langle \xi_b| \pi_a |\xi_b\rangle = \langle \xi_b|\xi_b\rangle$$

which is the same as before. □

Remark. The above arguments clearly generalises to the case of incomplete measurements of Alice and Bob. The good ol' trick of pushing ancillas to the front, delaying measurements to the end, and conjugating any unitary actions into the final measurement (like what we did for no-cloning) then allows us to deduce the general case.

This is quite a spooky conclusion: The Born rules suggest that we can instantaneously affect arbitrarily distant quantum states by performing local actions on another quantum states that's entangled with it. However, the theorem means that this effect cannot be physically noticed in any way, so although something has happened with this distant state, there is no classical information that one can extract from this incident. This can prompt one to question of the completeness of this quantum formalism that we adapt: Maybe there is another hidden underlying theory that will reproduce all the statistics that we've seen here and is local (i.e. local actions cannot affect distant objects)? This is known as the local hidden variable theory, which however is defied by Bell's theorem (1964).

3 Quantum Communication

3.1 Quantum Dense Coding

We already know that single qubit that is sent from a sender to a receiver can reliably encode just one classical bit of information. However, if this qubit is entangled nicely with a second qubit (i.e. the joint system of the two qubits is in an entangled state) that is already held by the receiver, then it actually has two bits of classical information capacity! The method of doing this is called quantum dense coding, which is a space-efficient quantum communication protocol designed by Bennett and Weisner (1992).

Definition 3.1. The Bell states is an orthonormal basis for the state space of 2 qubits given by

$$\begin{aligned}\left|\varphi^+\right\rangle &= \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle), \left|\varphi^-\right\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)\\ \left|\psi^+\right\rangle &= \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle), \left|\psi^-\right\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)\end{aligned}$$

Notably, we have

$$\begin{cases}\left|\varphi^-\right\rangle = Z \otimes I \left|\varphi^+\right\rangle = I \otimes Z \left|\varphi^+\right\rangle\\ \left|\psi^+\right\rangle = X \otimes I \left|\varphi^+\right\rangle = I \otimes X \left|\varphi^+\right\rangle\\ \left|\psi^-\right\rangle = Y \otimes I \left|\varphi^+\right\rangle = -I \otimes Y \left|\varphi^+\right\rangle\end{cases}$$

That is, we can make all the Bell states from just $\left|\varphi^+\right\rangle$.
In fact, for any 1-qubit gate U, we have $U \otimes I \left|\varphi^+\right\rangle = I \otimes U^\top \left|\varphi^+\right\rangle$ and $U \otimes U \left|\psi^-\right\rangle = (\det U) \left|\psi^-\right\rangle$. Note that $\det U$ is a phase (i.e. has unit modulus), so (as a physical state) $\left|\psi^-\right\rangle$ is invariant under coincidental joint changes to the two qubits. A similar phenomenon is that $U \otimes U \left|\varphi^+\right\rangle = \left|\varphi^+\right\rangle$ when U is a real rotation.
We can also make the Bell states from the computational basis. For example, $\left|\varphi^+\right\rangle = CX_{12}(H \otimes I(|0\rangle |0\rangle))$. The other Bell states can be made similarly (by starting with the other compuational basis) or from $\left|\varphi^+\right\rangle$ as described earlier. By reversing this process, we can also make the computational basis from the Bell states.
The quantum dense coding protocol is as follows: Suppose we have our usual

protagonists Alice and Bob distantly separated in space, each possessed one component qubit of a $|\varphi^+\rangle$ state. Alice wants to send 2 bits of (classical) information to Bob. For the messages $00, 01, 10, 11$, she applies the gates I, Z, X, Y to make the state $|\varphi^+\rangle, |\varphi^-\rangle, |\psi^+\rangle, |\psi^-\rangle$ respectively, and then send her qubit to Bob. Upon receipt of the qubit, Bob can then measure the whole system with respect to the basis given by the Bell state to distinguish the message.

3.2 Quantum Teleportation

Quantum teleportation is a quantum communication protocol in the so-called LOCC (local operations & classical communication) paradigm, where we impose the restriction that quantum operations can only be done locally and only classical bit (but not quantum states) can be communicated between different parties. The necessity of such restriction might be due to a hostile communication channel, i.e. somebody might try to contaminate the quantum state in transit, which may lose us all quantum information contained in it. Meanwhile, we have enough mechanism to detect and handle the contamination of classical bits. Generally, for such communication to work, we will need to make use of a entangled state with components seperated across space. This is based on the idea that physical effects that take place due to quantum entanglement is not affected by any physical actions out there.

Alice and Bob are, as before, distantly separated in space, and they can communicate classical bits. They each hold a qubit, the joint state of which is in the state $|\varphi^+\rangle$. Suppose Alice has another qubit in some state $|\alpha\rangle$ possibly unknown to her and she wants to transfer the quantum information in this qubit to Bob. Obviously she cannot try to identify $|\alpha\rangle$ since she won't achieve this with certainty but her observations will make it collapse. We don't allow communication of quantum states, so how would she do it?

Label the register containing $|\alpha\rangle$, the qubit Alice holds, and the qubit Bob holds as qubits $1, 2, 3$ respectively. So the initial state of the three-qubit system is $|\alpha\rangle_1 |\varphi^+\rangle_{23}$ (with Alice holding $1, 2$ and Bob holding 3). Write $|\alpha\rangle = a|0\rangle + b|1\rangle$, then by calculation

$$|\alpha\rangle_1 |\varphi^+\rangle_{23} = \frac{a}{\sqrt{2}}|000\rangle + \frac{a}{\sqrt{2}}|011\rangle + \frac{b}{\sqrt{2}}|100\rangle + \frac{b}{\sqrt{2}}|111\rangle$$

The quantum teleportation protocol is as follows:

Firstly, Alice applies CX_{12} to 12, H to 1, and measures her 2 qubits in the computational basis to get a 2-bit string. Notably, this is equivalent to a measurement in the basis of Bell states.

What would be the effect of this step? After applying the unitary operations, the joint state becomes

$$\frac{1}{2}(|00\rangle(a|0\rangle + b|1\rangle) + |01\rangle(a|1\rangle + b|0\rangle) + |10\rangle(a|0\rangle - b|1\rangle) + |11\rangle(a|1\rangle - b|0\rangle))$$

as one can verify easily with calculation. Note that we can write this alternatively as

$$\frac{1}{2}(|00\rangle I|\alpha\rangle + |01\rangle X|\alpha\rangle + |10\rangle Z|\alpha\rangle + |11\rangle XZ|\alpha\rangle)$$

So after Alice does the measurement, each 2-bit string outcome ij would be seen with probability $1/4$ and would result in a post-measurement state of

$|ij\rangle_{12} X^j Z^i |\alpha\rangle_3$.

The second step of the protocol is then Alice sending the two-bit outcome (which is classical information) over to Bob. Notably, since we know that each 2-bit string occurs with probability $1/4$ regardless of $|\alpha\rangle$, eavesdropping on this transmission of classical information is completely useless.

Lastly, after Bob receives the string ij, he can apply the unitary operation $Z^i X^j = (X^j Z^i)^{-1}$ on his qubit which guarantees to make it in state $|\alpha\rangle$.

The whole process can be summarised in the quantum circuit diagram as shown in Figure 3.1.

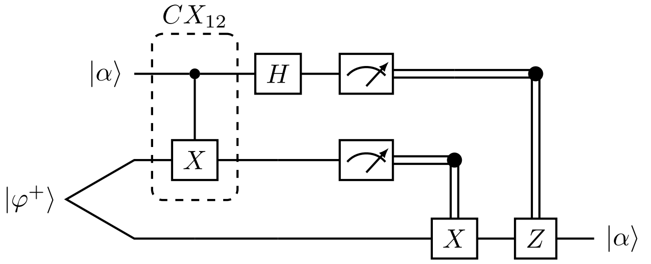

Figure 3.1: Quantum Circuit Diagram for Quantum Teleportation

Here, the single lines connect operations in quantum processes and double lines denote classical processes (in this case, the transmission of the classical bits obtained from the observations to Bob). I (the author) am going to assume that the reader is smart enough to work out what other components in this diagram means.

After teleportation, A is left with a 2-qubit state that is one of the computational basis. As we've discussed, its specific identity is uniformly random. One implication of this is that no information at all about $|\alpha\rangle$ remains with A. This is consistent with no-cloning (in fact a even stronger version of that, which says that if a copy is created then there is no information about the state that can be found anywhere else).

In addition, before Alice's measurement, Bob's qubit is the second component (the "right hand" half) of a $|\varphi^+\rangle$ state. One can show that for any complete measurement on this particular qubit, then any outcome occurs with probability $1/2$, which follows from either direct calculation or no-signaling. After Alice's measurement, B's qubit is in a state that's an equal probabilistic mixure of $|\alpha\rangle, X|\alpha\rangle, Z|\alpha\rangle, XZ|\alpha\rangle$. Averaging over these tells us that the outcome of any complete measurement must also have probability $1/2$, the same as before. This is consistent with no-signaling.

Quantum teleportation is, of course, distinct from e.g. Star Trek teleportation. The physical system embodying $|\alpha\rangle$ is not transferred. What happened was actually the reconstruction of the identity of the quantum state $|\alpha\rangle$ in a different physical system at the cost of destructing the original copy.

How did the (quantum) information actually get across? Only 2 classical bits got across the space, but the information transmitted, i.e. $|\alpha\rangle$ should in theory bear infinitely more information than what we've sent since it depends on continuous parameters. Which channel does this huge amount of quantum information get across? If you squint hard enough when looking at the space-time diagram of this whole process, you'll realise that there is another path that $|\alpha\rangle$

could've moved along: When $|\varphi^+\rangle$ was first created, the two components of it went to Alice and Bob through physical processes P_A and P_B respectively. The Bob component of it then stayed with Bob and went through the last part of the protocol to be come $|\alpha\rangle$, a process we shall denote as P_C. In a sense, the information $|\alpha\rangle$ could have travelled "backwards in time" via P_A^{-1}, then P_B and P_C to reach its final position. This is of course somehow counterintuitive, since it means that the information travelled even before Alice got it, but there are tonnes of weird stuff in the quantum realm already, so why not this too?

But of course, we want some justification and/or formalisation of this way of thinking this teleportation business. Normally if we want to detect if some information have travelled through a channel, we'd put some sort of observer on the route and see if there is anything that actually went through. This however doesn't work in the quantum case since observing a state would ruin it. What we could do, on the other hand, is to put a gate on the path and see if it affects the result at the end.

Suppose for simplicigy that the Bell measurement (i.e. the first step of the protocol done by Alice) yields 00 as the outcome. Then Bob needs to do nothing after receiving the bits. Now if we add a qubit gate U on P_C, then Bob is simply going to get $U|\alpha\rangle$. One can view this as the same teleportation but we are using $I \otimes U|\varphi^+\rangle$ as the entangled pair instead. Similarly, you'll get $U|\alpha\rangle$ if U were placed on P_B. So it looks like $|\alpha\rangle$ did run through the $P_B P_C$ path.

Interesting things happen when you put U on P_A: This will result in the same teleporation process with $U \otimes I|\varphi^+\rangle = I \otimes U^\top|\varphi^+\rangle$, so one will get $U^\top|\alpha\rangle$ which (with some quantum mechanical justification) turns out to be the correct time-reversed gate of U! This means that we can indeed view teleporation as $|\alpha\rangle$ going through $P_A^{-1} P_B P_C$.

This idea can also be applied to quantum dense coding to justify how the second bit got across.

Of course, physically "going back in time" would usually cause causality paradoxes. But in this case, we don't really have those. The teleportation process actually splits $|\alpha\rangle$ into two parts: The classical two bits i, j and the uniform probabilistic mixture of $|\alpha\rangle, X|\alpha\rangle, Z|\alpha\rangle, XZ|\alpha\rangle$. Both objects are completely random, so even if the latter travelled back in time, no information can be extracted by anyone in the past. This is somehow analogous to an idea in classical information theory to encrypt messages: For any fixed bit b and random bit r, both $b \oplus r$ and r are fully random, but one can reproduce b by $b = (b \oplus r) \oplus r$.

A natural problem with this teleporation protocol is that of implementation. In fact, since 1993 (when the protocol was first introduced), it was implemented for a single qubit in various instances. For example, it was successfully implemented in the Canary islands in 2012, with a distance separation of 90 miles. The entangled pair in that particular experiment was polarised photons, that were sent to the two communicating locations via optical fibres (to avoid perturbation of the qubits). The main technical difficulty was the Bell measurement, which has taken a lot of works to overcome.

The best experimental works on quantum information so far were those done by Jian-Wei Pan's team. They launched a satellite (called Micius) in 2016, which achieved a quantum teleportation over 1400 kilometers in 2017.

Of course, another question that always gets asked is whether we are able to teleport a human with this idea. Expectedly, there are a crap ton of technical difficulties: The very least you need to do is to access the $\approx 10^{30}$ atoms in a human body on a quantum level. You'll also need to send $\approx 10^{30}$ bits of classical information (which would take a magnitude of $\approx 10^8$ times the age of the universe with the fastest transmission rate so far), and apply the same magnitude of delicate reconstruction. There are also many philosophical questions about this, e.g. is a human really equivalent to the full information of their quantum state? Nevertheles, quantum teleportation will be useful for quantum internet, distributed quantum computing, etc..

3.3 Quantum Cryptography

The study of cryptography is not in any sense a completely modern thing. Screcy might as well be amongst the earliest inventions of humanity. The earliest systematic way to encode a message might be the substitution cipher (or permutation cipher), which takes a string of message and apply the same permutation of the alphabet to its letters. Examples of this include Caesar's method of moving every letter three steps ahead in the alphabet, modulo the size of the alphabet. These kinds of encryption require the receiver and the sender to share a secret, namely the details of the permutation, which is not really secure: It is hard to guess the permutation, sure. But one can compile a table of frequencies of symbols and use th characteristic of language (e.g. vowels are much more common than consonants) to deduce the permutation.

There more sophisticated (classical) encryption schemes than that. The ones are used most are probably the family of public key cryptography systems like Diffie-Hellman, RSA, elliptic curves, etc.. The advantage of these systems is that we don't really need a shared secret. The receiver can publish the encryption key ("public key") but only the receiver can decrypt it easily using their knowledge of a "private key". The secrecy is usually based on unproven but widely believed difficulty of calculating the private key from the public key. For example, RSA is based on the assertion that factorising an integer cannot be done in polynomial time. At the end of the day, they are still not provably secure: No one knows what sort of algorithms will be discovered in the future. In fact, quantum computing has already provided new algorithms that solve the problem (e.g. Shor's algorithm solves the factorisation problem) for – coincidentally – all problems of calculating private keys from public keys in such cryptography systems that's currently in use.

The only provably secure classical encryption method is the so-called one-time pad: Assume the message is a bit string M of length n. Our protagonists Alice and Bob will share a secret private key K which is a freshly generated uniformly random bit string, of length n as well. Alice can then send $C = K \oplus M$ to Bob using whatever method, who can recover M by $M = K \oplus C$. If K is uniformly random, then so is C. An eavesdropper (called Eve) can indeed obtain C without leaving a trace, but C is uniformly random since K is, so it tells Eve nothing at all. This achieves the desired secrecy. But of course, if the same key is used several times (say $C_1 = M_1 \oplus K, C_2 = M_2 \otimes K$), then Eve can obtain $M_1 \oplus M_2 = C_1 \oplus C_2$, which is why we need a freshly generated K each time. Classically, the distribution of this private key is done by a trusted middle party, which has the fatal problem of being not so trustable at times.

So, the only provably secure method of cryptography is "provably insecure" when done classically, and the ones that "looks secure" are susceptable to quantum attacks. There has been studies of "post-quantum cryptography", which aims at finding classical encryption schemes that prevent quantum attacks, but there hasn't so far been many successes. So where should encryption go with the development of quantum technologies? Quantum giveth, quantum taketh away (or more like the other way around): Quantum information theory can actually provide new encryption protocols that allows provably secure communication!

The basic principle of quantum cryptography is to encode messages with non-orthgonal states. As we've seen, we cannot reliably distinguish non-orthogonal states, and any attempt to read the messag would cause irreversible damage to the quantum state. This is a problem for both the receiver of the message and the eavesdropper, but the former can circumvent this situation by making use a more elaborate protocol that's inaccessible to the eavesdropper, which will usually allow unconditionally provably secure communication.

The most studied ways of doing this are quantum key distribution schemes. Namely, we will take the classical idea of one-time pad, but the middle party is replaced by quantum mechanisms to make it secure. One method of doing this is called the Bennett-Brassard 1984 (BB84) scheme, which is what we will discuss here.

Our situation is like before: Alice and Bob are distantly seperated in space and communicate only over classical and quantum channels. This time, there is an eavesdropper Eve who also has access to these channels. Alice and Bob want to create a random bit to share. Write

$$|\psi_{00}\rangle = |0\rangle, |\psi_{10}\rangle = |1\rangle, |\psi_{01}\rangle = |+\rangle, |\psi_{11}\rangle = |-\rangle$$

They form a pair of orthonormal bases, namely the Z-basis $B_0 = \{|0\rangle, |1\rangle\}$ and the X-basis $B_1 = \{|+\rangle, |-\rangle\}$, which are conjugate bases.

Alice first generates two uniformly random binary strings of length m, say $X = x_1 x_2 \cdots x_m$ and $Y = y_1 y_2 \cdots y_m$. Then, she prepares m qubit states $|\psi_{x_1 y_1}\rangle |\psi_{x_2 y_2}\rangle \cdots |\psi_{x_m y_m}\rangle$ and send them over to Bob. This is sometimes called (quantum) conjugate coding.

When Bob receives the m qubits, they might or might not have been eavesdropped/corrupted on the way by Eve (there might also be noises in the channel, which we shall include as a form of eavesdropping/corruption). Assume first that Eve slept through the event and the quantum states were pristine. Bob then generate a uniformly random bit string $Y' = y'_1 y'_2 \cdots y'_m$ and measures the k^{th} qubit in the basis $B_{y'_k}$ which gets him a string of outcomes $X' = x'_1 x'_2 \cdots x'_m$. Of course, if $y'_i = y_i$ (which happens with probability 1/2), then $x'_i = x_i$. If $y'_i \neq y_i$, then x'_i would not relate to x_i at all.

Next, Alice and Bob publish and compare their choices of bases, i.e. Y and Y', and discard all entries of X, X' for which $y_i \neq y'_i$, leaving a shorter string (shared by Alice and Bob) $\tilde{X} = \tilde{X'}$ of expected length $n/2$, which would provide the desired shared private key.

What if Eve were awake and had done something to the qubits Alice sent? We of course want to leave out the manipulated qubits, a process called information reconciliation. The first thing to do is to estimate the number bits of $\tilde{X}$ that differs from $\tilde{X}$. To do this, they publically compare a random sample of their strings, say half of the bits chosen (publicly) at random positions, and of course discard all announced bits. Assume the remaining bits have about the same proportion of errors (albeit at unknown positions).

Next, they want to correct these errors to obtain two strings that agree at high proportion of positions with high probability. Sounds crazy, but it is actually possible without giving everything away (assuming the bit error rate is not too much – if it is indeed too much, as measured by the estimation earlier, Alice and Bob can just start over). It's actually quite complicated to describe in full generality so we'll leave that there.

The last step is the privacy amplification, namely trying to prevent Eve from knowing much about the string has she been eavesdropping. From the estimation done in the last step, Alice and Bob supposedly know the approximated proportion of eavesdropped qubits. From this estimate, they can use only public discussion (with well-established classical information theory) to obtain a possibly shorter string that Eve practically has no knowledge of.

There are many possible attacks by Eve. She can perform a intercept-resend attack, i.e. she intercepts each passing qubit separately, measure them, and then send the post-measurement qubits to Bob. She can also perform the general coherent attack, where she take a (large) quantum probe system that unitarily interacts with each passing qubit. Then at any point afterwards (possibly even after Alice and Bob's public discussions), E can measure the probe to obtain (possibly joint) information about the qubits.

However, it is provable that this scheme is indeed secure and immune of these attacks. We don't exactly have the budget to provide the details of the last two steps or the full proof of security, but we'll go through some examples. Let's first see what a certain form of intercept-resend attack would do to the protocol.

Example 3.1. Assume that the communication channel is noiseless except that Eve intercepts each qubit and measures it in the Breidbart basis

$$|\alpha_0\rangle = \cos(\pi/8)|0\rangle + \sin(\pi/8)|1\rangle, |\alpha_1\rangle = -\sin(\pi/8)|0\rangle + \cos(\pi/8)|1\rangle$$

and sends on the post-measurement state. If the outcome is $|\alpha_0\rangle$, then Eve concludes that the bit Alice held (in the string X) at that position is 0; If the outcome is $|\alpha_1\rangle$, she concludes that the bit is 1. The choice of $\pi/8$ is of course due to the fact that $|1\rangle, |+\rangle$ are reflected to $|0\rangle, |-\rangle$ via the "line with inclination $\pi/8$". This is useful in the sense that the square overlaps of both $|0\rangle, |+\rangle$ with $|\alpha_0\rangle$ are $\cos^2(\pi/8)$, as were that of $|1\rangle, |-\rangle$ with $|\alpha_1\rangle$. So in any case Eve's measurement correctly guesses Alice's bit with probability $\cos^2(\pi/8) = (1+1/\sqrt{2})/2$, so it's quite a good eavesdropping strategy.

Let's now calculate the bit error rate of $\tilde{X}$ against $\tilde{X}'$ has Eve eavesdropped in this way. We know that in the substring $\tilde{X}'$, Bob was using the same bases that Alice used to encode. There are a few cases. Suppose that Alice sent $|0\rangle$ (so $\tilde{X}$ has entry 0 at that position), then Bob gets the wrong answer $|1\rangle$ with probability

$$\sin^2(\pi/8)\cos^2(\pi/8) + \cos^2(\pi/8)\sin^2(\pi/8) = \frac{1}{4}$$

after some calculations. The other cases are similar, so in any case the bit error rate is 1/4.

We'll also give some idea of how certain ways of information reconciliation would work, which is strictly speaking just classical information theory. Knowing the bit error rate estimate, Alice and Bob first apply a (public) random permutation to both strings (i.e. randomise the positions of errors). Then, they break strings into blocks of suitable lengths determined by the bit error rate so that (with high probability) each block contains at most one error. For each block, both of them compute the sum of the digits modulo 2 (the parity) and accept blocks that have the same parity due to our assumption. For block with disagreeing parities, they further break the blocks into halves and repeat the process until the block size is reduced to 1, in which case Bob flips it. For this new, possibly amended string, Alice and Bob repeat the whole process with a most likely smaller bit error rate. Eventually, they stop when the bit error rate is small enough (or the block size is big enough), in which case they accept the strings.

Note that more information would be leaked to Eve (i.e. the parities and flipped digits) if this protocol were adopted, but it's not a big problem, since (if the block size is k) the magnitude of the leaked information has the order of mere $\log_2 k$.

How would privacy amplification be done then? Assume that information reconciliation is done perfectly so that Alice and Bob are confident that the string they are holding are the same (or close enough to work).

We'll make use of the fact that if Eve knows bit x but not bit y, then $x \oplus y$ is uniformly random to Eve. More generally, if Eve knows some, but not all, bits in a string, then she has no knowledge of the parity of the string. So, suppose the shared string is $x_1x_2x_3$ and Eve only knows one bit in it, then she still has no information at all about the string z_1z_2 given by $z_1 = x_1 \oplus x_3, z_2 = x_2 \oplus x_3$. In $\mathbb{Z}/2\mathbb{Z}$, we can write this as

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = G \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, G = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

It takes a bit of work to construct such a modification which produces a string that's completely unknown to Eve in general. Of course, not everything works. Suppose we have a 4-bit string $x_1x_2x_3x_4$ from which Eve knows only 2 bits, then $z_1 = x_1 \oplus x_2 \oplus x_3$ and $z_2 = x_2 \oplus x_3 \oplus x_4$ are both unknown to Eve, but $z_1 \oplus z_2 = x_1 \oplus x_4$ will be known to Eve if the two bits she has knowledge of are x_1 and x_4.

Expectedly, we do have a theorem on this.

Definition 3.2. A code of dimension n is a linear subspace of $(\mathbb{Z}/2\mathbb{Z})^n$. The ($n$-dimensional) code generated by an $m \times n$ matrix G is one that is spanned by the rows (the "codewords") of G.

The Hamming weight of a bit string is the number of 1's in it. The minimum Hamming weight of a code is the least possible Hamming weight of codewords in it.

The Hamming distance of two codewords is the Hamming weight of their sum.

Theorem 3.1. *Suppose Alice and Bob share a string of length n which Eve knows $k < n$ bits from, then the $m \times n$ Boolean matrix G with $m < n$ will produce our desired secure string in the way described above if and only if the minimum Hamming weight of the code generated by G is strictly greater than k.*

Theorem 3.2 (Universal Hashing)**.** *Given an n bit string x and $m < n$ (possibly determined by bit error rate) and a random Boolean matrix G, then with high probability E will have no information about the string $z = Gx$.*

4 Quantum Computation

4.1 Computational Complexity

Write $B = B_1 = \mathbb{Z}/2\mathbb{Z}, B_n = (\mathbb{Z}/2\mathbb{Z})^n, B^* = \bigcup_n B_n$.

Definition 4.1. Given the input of an integer n (the "input size"), a bit string $x \in B_n$, and a "language" $L \subset B^*$, a decision problem is a computational task to give a one-bit output that decides whether $x \in L$.

There are many possible choices of (classical) computational model, e.g. Turing machines, circuit model, etc.. An important common feature of these models is that they are all processes with discrete steps with each of them requiring constant effort to perform. In the circuit model, a class of examples of this consists of elementary 1- or 2-bit Boolean gates: For input $x = i_1 \cdots i_n$ which we extend with extra trailing zero entries, the basic computational steps are the operations of AND, OR or NOT gate applied to specified bits. It can be proved that these three gates are universal in the sense that any Boolean function can be constructed as the composition of a sequence of them.

Definition 4.2. A classical computation (or algorithm) is a family prescribed sequence C_n indexed by the input size n consisting of AND, OR and NOT.

There's also the notion of a randomised (or probabilistic) classical computation, which is simply classical computation but we extend x in addition by a trailing sequence of (newly generated) random bits. We usually require the output of this to be correct with suitably high probability.
We want to know how much more resources would be needed if we increase n. This usually consists of two parts: The time $\mathrm{T}(n)$ which is the size of C_n and the space $\mathrm{Sp}(n)$ which is the amount of memory (length of the extended trailing string) needed. The growth of $\mathrm{T}(n)$ is known as the time complexity of the algorithm, and that of $\mathrm{Sp}(n)$ the space complexity of it.

Definition 4.3. The algorithm is said to be polynomial-time (aka poly-time, efficient) if $\mathrm{T}(n) = O(n^k)$ for some k.

Polynomial-time algorithms are recognised as practically computable. Superpolynomial algorithms (i.e. algorithms whose $\mathrm{T}(n)$ grows faster than any polynomial) are computable in principle but usually not feasible in practice. If $\mathrm{T}(n)$ is grows even quicker, e.g. if it grows exponentially, then it is basically impossible to compute before the end of humanity for large n.
By the way, the reason why we choose polynomial instead of functions with even faster growth (e.g. linear/quadratic) as benchmark is that it makes the notion independent from the specific computation model, e.g. whether the input is in decimal or binary, whether we are using circuits or Turing machines (which can be single-taped, multi-taped, etc.), or which set of universal gates we are using.
Another nice property about polynomials is that they are closed under elementary algebraic operations, so once we know that these operations can be done in polynomial time (which is pretty easy to see) and our algorithm consists of polynomial-many such operations, then we can also conclude that the whole algorithm is polynomial-time.
We classify decision problems by their complexity classes.

Definition 4.4. The class P (poly-time) consists of the decision problems that admit a deterministic polynomial-time algorithm.
The class BPP (bounded-error probabilistic poly-time) consists of those which admit a probabilistic polynomial-time algorithms such that the probability of a correct answer uniformly exceeds a fixed constant that is strictly greater than 1/2.

This constant for BPP is conventionally taken to be 2/3, but as one would have expected it doesn't matter at all because of the simple reason that one can run the algorithm repeatedly (and take the "majority vote"). BPP is regarded as the set of practically solvable problems (by classical means).
Clearly $\mathsf{P} \subset \mathsf{BPP}$, but whether $\mathsf{P} = \mathsf{BPP}$ remains unknown.

Example 4.1. The problem of primality testing asks whether a given integer x is prime. The naïve algorithm of trial division until $\sqrt{x}$ takes exponential time. Knocking off the non-primes in this trial division doesn't push it to polynomial time either. One might turn to seek probabilistic algorithms by testing a randomly chosen subset of these potential divisors, but most fails when e.g. x is the product of two large primes.
There are of course more sophisticated algorithms. The Solovay-Strassen algorithm (1977) gives a probabilistic primality test which shows that the problem is in BPP. In 2004, an unconditional deterministic primality test algorithm was found by Agrawal, Kayal and Saxena, so the problem is actually in P.

As widely acknowledged, there are probably more complexity classes than interesting computational problems. One has the (in)famous class NP which we will talk about later, and PSPACE which includes problems admitting a deterministic algorithms with $\mathrm{Sp}(n)$ growing like polynomial. Curiously, space is "more powerful" than time: Clearly $\mathsf{P} \subset \mathsf{PSPACE}$, and in fact we have $\mathsf{P} \subset \mathsf{NP} \subset \mathsf{PSPACE}, \mathsf{P} \subset \mathsf{BPP} \subset \mathsf{PSPACE}$. It is unknown whether any of these inclusions is strict, and we know very little about the relation between NP and BPP.

4.2 Circuit Model of Quantum Computation

For input $x = i_1 \cdots i_n \in B_n$, we start with a system of qubits in the form $|i_1\rangle \cdots |i_n\rangle |0\rangle \cdots |0\rangle$. Computational steps in the model we are considering are quantum gates on designated qubits, which are usually chosen from the gates $H, X, Z, P(\theta), CX, CZ$. The output is a final quantum measurement in the compuational basis $\{|0\rangle, |1\rangle\}$ on specific qubits. Recall that we can push all measurements to the end of a quantum process, so we don't really need to restrict ourselves to quantum circuits with all measurements done at the end when talking about quantum computation.
How would we relate the classical sense of computation to the its quantum counterpart? For a Boolean function $f : B_m \to B_n$, one can consider $\tilde{f} : B_{m+n} \to B_{m+n}$ given by $(x, y) \mapsto (x, y \oplus f(x))$ (with the identification $B_{m+n} = B_m \times B_n$) where $\oplus = \oplus^n$ is the bitwise modulo 2 addition. We can recover f from $\tilde{f}$ as $f(x)$ is the last n bits of $\tilde{f}(x, 0)$. Easily $\tilde{f}$ is self-inverse, in particular invertible.
Invertible Boolean functions can actually be modelled within the framework of quantum circuits. For invertible $g : B_k \to B_k$, the linear map on k-qubits defined

(on the computational basis) by $A_g |x\rangle = |g(x)\rangle$ is unitary. Apply this on $\tilde{f}$ gives a unitary operator $A_{\tilde{f}} = U_f$ on $(m+n)$-qubits with $U_f |x\rangle |y\rangle = |x\rangle |y \oplus f(x)\rangle$. Note that if we know a state is Boolean (i.e. in the computational basis), then we can decide which one it is with absolute certainty, hence we can represent any Boolean function in the framework of quantum computing. Also, linearly of U_f shows that it provides the mapping

$$(H^{\otimes m} |0\rangle^{\otimes m}) |0\rangle^{\otimes n} = \left(\frac{1}{\sqrt{2^m}} \sum_{x \in B_m} |x\rangle \right) |0\rangle^{\otimes n} \mapsto |f\rangle = \frac{1}{\sqrt{2^m}} \sum_{x \in B_m} |x\rangle |f(x)\rangle$$

That is, one single evaluation of U_f (at a state that can be produced in linear time) can produce $|f\rangle$ which supposedly would contain all (exponentially many) values of f!
In classical computing, we have a set of universal gates, namely AND, OR and NOT, in the sense that a sequence of them can represent any Boolean functions. It is very tempting to study a quantum analogue of such universality.

Definition 4.5. A set of quantum gates is universal if we can reproduce any unitary operation on the string of qubits from a sequence of them.

Example 4.2. The set of gates consisting of CX and all 1-qubit gates is universal.

However, we can only use a finite set of quantum gates in practice, but it can be seen easily from countability arguments that there is no finite universal set of quantum gates. This asks for a slightly weaker sense of universality.

Definition 4.6. A set of gates $\mathcal{G}$ on n-qubits is approximately universal if, for any $\epsilon > 0$ and any unitary operator W on n-qubits, there is a circuit $\tilde{W}$ made from gates in $\mathcal{G}$ such that $\sup_{\langle\psi|\psi\rangle=1} \| W |\psi\rangle - \tilde{W} |\psi\rangle \| < \epsilon$.

One can rewrite the last condition as $\|W - \tilde{W}\|_{\text{op}} < \epsilon$ where $\| \cdot \|_{\text{op}}$ is the operator norm $\|A\|_{\text{op}} = \sup_{\langle\psi|\psi\rangle=1} \|A |\psi\rangle \|$.

Example 4.3. The set $\mathcal{G} = \{H, CX, P(\pi/4)\}$ is approximately universal.

The (worst case) size of the circuit $\tilde{W}$ produced in this way is in general exponential in n, but it behaves much better as $\epsilon \to 0$.

Theorem 4.1 (Solovay-Kitaev). *Suppose $G \in \mathcal{G} \implies G^{-1} \in \mathcal{G}$. For each fixed n, there is a polynomial P such that for all unitary operator W on n-qubits, the smallest possible size of a circuit $\tilde{W}$ with $\|W - \tilde{W}\|_{\text{op}} < \epsilon$ is bounded by $P(-\log \epsilon)$.*

The introduction of quantum circuits also allows us to define a new complexity class of decision problems.

Definition 4.7. The complexity class BQP (bounded-error quantum poly-time) of decision problems consists of those that can be solved with a polynomial-sized quantum circuit family (indexed by the input size as before) with success probabilty exceeding $2/3$.

Like before, one can replace $2/3$ by any number in $(1/2, 1)$ and end up with the exact same complexity class. Turns out, BQP is also independent of the choice of finite approximately universal gate set.
One can show that $\mathsf{BPP} \subset \mathsf{BQP}$. Whether or not the inclusion is strict is, unsurprisingly, still open.

4.3 Oracle Problems

Suppose we have a different scenario: Instead of inputting a bit string or a string of computational basis, we input a black box (or an "oracle") O_f that computes some Boolean function $f : B_m \to B_n$ whose form we might have a priori knowledge about. Each use of the oracle (each "query") counts as one computational step and the oracle is our only access to f. The family of problems we are interested here are the decisions of whether f has certain properties. WLOG the computation always starts on $|0\rangle \cdots |0\rangle$.
In this scenario, we introduce

Definition 4.8. The query complexity of the algorithm is the number of times that the oracle needs to be used in order to solve the problem. The total time complexity is the total size of the circuit, counting each oracle use as a single gate.

Example 4.4 (Balanced vs. constant problem)**.** We input an oracle for a Boolean function $f : B_n \to B_1$ that is either constant or "balanced" in the sense that exactly half of its values is 0. The problem is to decide whether f is constant with certainty.
Classically, $2^{n-1} + 1$ queries is obviously sufficient and necessary. So the best query complexity in classical situation grows exponentially in n, which is bad.

The amazing thing is that there actually exists a quantum algorithm that solves the problem much, much quicker! This is known as the Deutsch-Jozsa algorithm (1992).
By setting the output register (the last digit) of U_f to $|-\rangle = H|1\rangle = HX|0\rangle$, we have

$$U_f(|x\rangle |-\rangle) = |x\rangle \left(\frac{|f(x)\rangle - |1 \oplus f(x)\rangle}{\sqrt{2}} \right) = (-1)^{f(x)} |x\rangle |-\rangle$$

That is, we have encoded f with $|+\rangle, |-\rangle$ instead of $|0\rangle, |1\rangle$ (a "phase kickback"). Linearity then gives

$$U_f \left(\frac{1}{\sqrt{2^n}} \sum_{x \in B_n} |x\rangle |-\rangle \right) = |\xi_f\rangle |-\rangle, |\xi_f\rangle = \frac{1}{\sqrt{2^n}} \sum_{x \in B_n} (-1)^{f(x)} |x\rangle$$

So one query to the oracle (i.e. one application of U_f) gives us the state $|\xi_f\rangle$. We shall show that this state already allows us to decide whether f is constant or balanced with certainty. The key observation here is that $|\xi_f\rangle$ is orthogonal to $|\xi_g\rangle$ whenever f is constant and g is balanced, so in theory we can perfectly distinguish these quantum states, i.e. a certain measurement on $|\xi_f\rangle$ would tell whether f is constant or balanced.
Not too fast though – in the context of quantum computation, we only allow measurements in standard basis, so we still need one more step to specify how the measurement shall be done. Indeed, if f is constant, then $H^{\otimes n} |\xi_f\rangle = \pm |0\rangle^{\otimes n}$. A measurement of $H^{\otimes n} |\xi_f\rangle$ then gives all zero if and only if f is constant (with probability 1).
The quantum circuit diagram representationn of the algorithm is shown in Figure 4.1. This whole process uses 1 query and the total circuit size is $O(n)$, which is a gigantic improvement from the classical case.

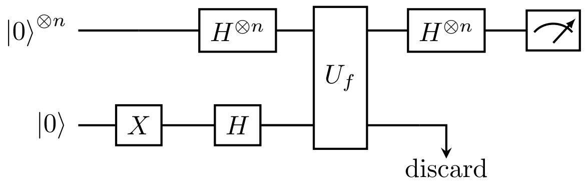

Figure 4.1: Quantum Circuit Diagram for the Deutsch-Jozsa Algorithm

The class of functions $f = f_a$ that has $|\eta_{f_a}\rangle = |a\rangle$ for some $a \in B_n - \{0 \cdots 0\}$ (explicitly, $f_a(x) = x \cdot s = x_1 a_1 \oplus \cdots \oplus x_n a_n$) is particularly interesting in the sense that the final measurement in the algorithm essentially recovers a. The recovery of a given the oracle of f_a is known as the Bernstein–Vazirani problem.
Can we decide any yes/no question about a Boolean function $f : B_n \to B_1$ by a quantum algorithmm on $|f\rangle$ or small enough number (i.e. polynomial many) of queries to U_f? The answer, sadly, is no. An example of this is the Boolean satisfiability problem (the SAT problem): Is there an x with $f(x) = 1$? One can prove that any quantum algorithm solving this with probability $1 - \epsilon$ for any $\epsilon > 0$ needs at least $O(\sqrt{2^n})$ queries, and of course any classical algorithm needs at least $O(2^n)$ queries.
If we tolerate an error in the constant vs. balanced problem (i.e. answer being correct with probability $1 - \epsilon$ for small $\epsilon > 0$), then we do have a classical algorithm with query complexity constant in n as well. Indeed, for $K \in \mathbb{N}$ and $x_1, \ldots, x_K$ chosen uniformly randomly from B_n, one can simply evaluate $f(x_1), \ldots, f(x_K)$ and conclude that f is constant if these are all the same. In this way, the probability of being wrong is at most $1/2^{K-1}$ which is less than ϵ whenever $K > 1 - \log_2 \epsilon$. Note that the query complexity is $O(-\log \epsilon)$ as $\epsilon \to 0$.
One might say that, in this sense, the Deutsch-Jozsa algorithm doesn't really exhibit a huge improvement of query complexities since one has constant query complexity with classical approach to this problem as well with the only setback being a possible error rate which can be made arbitrarily small.
However, as you will see in example sheet, there're some more problems where quantum algorithms have a provably exponential complexity improvement from the classical case even with the allowance of bounded error. Specifically, one would consider Simon's problem: Consider an oracle for $f : B_n \to B_n$ which is either one-to-one or is periodic, i.e. has $f(x \oplus \xi) = f(x)$ for some fixed bit string $\xi \in B_n$. The problem is to decide which case is it. One can show that any classical algorithm would have exponential query complexity even with an allowance of bounded error, but on example sheet you'll see a quantum algorithm (Simon's algorithm) solving this with $O(n)$ queries.
Simon's algorithm was inspired by Deutsch-Jozsa, and the fact that Simon's problem is a kind of periodicity determination further fueled the invention of Shor's factoring algorithm.
Does these superior quantum algorithms fully demonstrate the supermacy of quantum computing? One possible criticism of showing the supremacy with this is that these are all oracle problems, which is somehow unrealistic. We would of course like such a provable complexity seperation for standard computational task, but sadly no provable ones have been found yet. The problem is just that it is very very hard to prove a lower bound for the classical complexity of a standard problem.

What if we replace the oracle with an algorithm that computes f? If the algorithm halts in polynomial time, then any polynomial time algorithm for the oracle case surely stays polynomial. But don't you get to analyse the algorithm and maybe do something with it that's not equivalent to a query? Sadly, the algorithm can be hugely obfuscated. It is not known yet whether one can produce a quicker classical algorithm with this idea.

4.4 Quantum Fourier Transform; Periodicity Problem

Simon's problem hints that quantum computing does pretty well on periodicity problems. Let's try to develop this idea further.
We are going to move away from just qubits and consider a general quantum state (with dimension possibly not a power of 2).

Definition 4.9. Suppose we have a N-dimensional state space with computational basis $\{|n\rangle : n \in \mathbb{Z}/N\mathbb{Z}\}$. The quantum Fourier transform modulo N (also written as $\mathrm{QFT}, \mathrm{QFT}_N$) on the space is the linear operator takes a basis state $|a\rangle$ to

$$\frac{1}{\sqrt{N}} \sum_{b\in\mathbb{Z}/N\mathbb{Z}} e^{2\pi i ab/N} |b\rangle$$

The matrix of QFT_N in the computational basis is then, by definition, $[\mathrm{QFT}_N]_{ab} = N^{-1/2}\omega_N^{ab}$ where $\omega_N = e^{2\pi i/N}$. Clearly, $\sqrt{N}\,\mathrm{QFT}_N$ is symmetric and its first row and first column are populated with 1 only. Also, each row and column of it is a geometric series. As a special case, $\mathrm{QFT}_2 = H$. But of course $\mathrm{QFT}_4 \neq H \otimes H$ and in general $\mathrm{QFT}_{2^n} \neq H^{\otimes n}$ for $n > 1$.

Proposition 4.2. QFT_N *is unitary.*

Proof. Just compute. □

It is a fact that there is an implementation of QFT_N with time complexity polynomial in $\log N$. One possible implementation for the case where N is a power of 2 (which has complexity $O((\log N)^2)$) is based on the fact that

$$\mathrm{QFT}_{2^n} |x\rangle = \frac{1}{\sqrt{2^n}} \sum_{y=0}^{2^n-1} e^{2\pi i xy/2^n} |y\rangle$$

is always a product state of n qubits. The resulting quantum circuit is actually analogous to the classical fast Fourier transform, albeit faster.
We will use it to deal with the periodicity problem: Given a function $f : \mathbb{Z}/N\mathbb{Z} \to Y$, it is automatically periodic as $f(x+N) = f(x)$. Let r (called the period of f) be the minimal positive value such that $f(x+r) = f(x)$ holds for all x, then clearly $r \mid N$. Suppose in addition that f is one-to-one in $\{0, \ldots, r-1\}$. The problem is to find r with a positive constant level of probability given an oracle of f.
Classically, $O(\sqrt{N})$ queries are necessary and sufficient (example sheet). However, we can do it with quantum computing with $O(\log\log N)$ queries. The total complexity would also be a polynomial in $\log N$.
We will see later that we can reduce the factorisation problem of an integer K to the periodicity problem with some $N \approx K^2$, which would give a polynomial-time factoring algorithm.

How would the algorithm work? WLOG $Y = \mathbb{Z}/M\mathbb{Z}$ for some M, then we can have the quantum oracle $U_f \left|x\right\rangle \left|y\right\rangle = \left|x\right\rangle \left|y + f(x)\right\rangle$ analogous to what we did before. The first thing we do is to make the state

$$\frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} \left|x\right\rangle = \mathrm{QFT}_N \left|0\right\rangle$$

with quantum Fourier transform. Then we can use one query on this to get the state

$$\left|f\right\rangle = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} \left|x\right\rangle \left|f(x)\right\rangle$$

Then, we measure the second register of $\left|f\right\rangle$ under the computational basis and get some value $y = f(x_0)$ with x_0 chosen minimally. Each outcome occurs with uniform probability $1/r$. Set $A = N/r$, then the collapsed state would be $\left|\mathrm{per}\right\rangle \left|y\right\rangle$ where

$$\left|\mathrm{per}\right\rangle = \frac{1}{\sqrt{A}} \sum_{j=0}^{A-1} \left|x_0 + jr\right\rangle$$

Simply measuring this would give us no information about r nor x_0 at all. The trick here is to use QFT_N on it. We have

$$\begin{aligned}\mathrm{QFT}_N \left|\mathrm{per}\right\rangle = \frac{1}{\sqrt{NA}} \sum_{j=0}^{A-1} \left(\sum_{l=0}^{N-1} \omega_N^{(x_0+jr)l} \left|l\right\rangle \right) = \frac{1}{\sqrt{NA}} \sum_{l=0}^{N-1} \omega_N^{x_0 l} \left(\sum_{j=0}^{A-1} \omega_N^{jlr} \right) \left|l\right\rangle \\ = \frac{1}{\sqrt{NA}} \sum_{k=0}^{r-1} \omega_N^{x_0 kA} A \left|kA\right\rangle = \frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} \omega_N^{x_0 kN/r} \left|kN/r\right\rangle\end{aligned}$$

We now perform a measurement on it under the computational basis, which would give an outcome C that has the form $k_0 N/r$ for some $0 \le k_0 < r$. Each of these outcomes is observed with probability $1/r$.

Lastly, we attempt to recover r from C. If k_0 is coprime to r, then one can just reduce $C/N = k_0/r$ in the simplest form and r would sit in the denominator. This can be done with Euclid's algorithm whose complexity is polynomial in $\log N$. In general, we are just going to do exactly the same thing, which would give some $\tilde{r}$ that would be a factor of r. $\tilde{r}$ would be the correct answer as long as k_0 (which is uniformly random in $\{0, \ldots, r-1\}$) is coprime to r.

We don't need to worry about a wrong answer $\tilde{r} \neq r$ since we can spot it right away by sending queries to the oracle at 0 and $\tilde{r}$ (as $\tilde{r} \mid r \implies \tilde{r} \le r$). What we do need to take a look at is the success probability, which is the probability of $\gcd(k_0, r) = 1$.

Theorem 4.3.

$$\liminf_{r\to\infty} \frac{\varphi(r)}{r} \log\log r = e^{-\gamma}$$

This doesn't quite guarantee a constant success probability with only one attempt as above. However, we can improve it to that level with $O(\log\log N)$ attempts (as $\log\log N > \log\log r$). Overall, since each attempts involves a constant number of queries (one to get $\tilde{r}$, two to check if it's the correct answer), the overall query complexity would be $O(\log\log N)$ and the overall time complexity would be polynomial in $\log N$ (fixing any desired success probability in $(0,1)$). With some more number theory arguments, we can further optimise the algorithm and only need to repeat the step constantly many steps. But the algorithm in its current form is sufficient for our discussion, so we'll just leave it there.

4.5 Search Problems

The principal issue of search problems is to search for some "good items" in a search space, either to understand what it is or simply whether it exists.
Usually, given any item in the search space, it should be easy (i.e. can be done in poly-time) to check if it's good or not. We will assume that this is the case, since otherwise there is little thing we can do.
For interestingness, the search space is almost always large in the sense that it has superpolynomial size (so that the brute-force way cannot be done in poly-time).

Example 4.5. 1. Simultaneous constraint satisfaction problems such as making a timetable.
2. The SAT problem that asks whether 1 is in the image of $f : B_n \to B_1$ which can either be computed in polynomial time or has an oracle.

In fact, every search problem can be reduced to some form of the SAT problem.
We introduce the (in)famous complexity class known as NP.

Definition 4.10. A decision problem (say with language L) is in the complexity class NP (non-deterministic poly-time) if there is a computation $V(x,c)$ (the "poly-time verifier") that:
(a) Is poly-time in the length of x.
(b) Halts with "accept" for some c if $x \in L$ and "reject" for all c if $x \notin L$.

The variable c serves as a "certificate" of the membership of x in L. Clearly $\mathsf{P} \subset \mathsf{NP}$, but it is (yet again) not known whether the inclusion is strict. The SAT problem is also in NP, so is the test for whether a number is composite.
An alternative definition is the based on "what the computer does". A deterministic algorithm is one that looks like a straight line which you know what happens at each step; A non-deterministic algorithm is one that can split into two branches at any step (although no probabilistic behaviour is involved). A non-deteministic algorithm is poly-time if each particular path halts in poly-time (note that different path can give distinct accept/reject answers for any given input). We say the non-deterministic computation accepts an input x if at least one path accepts the input, and rejects x if all paths rejects the input.

Definition 4.11 (Alternative Definition of NP)**.** A decision problem is in NP if it is the acceptance set of a non-deterministic poly-time computation.

Theorem 4.4. *These two definitions coincide.*

Sketch of Proof. Given a polynomial verifier $V(x,c)$, since V runs in poly-time, the number of bits of c that can be relevant is polynomial in the length of x. Branching over possibilities of c gives the desired non-deterministic algorithm.
Conversely, given a non-deterministic algorithm, one can encode the string c in the branching of the algorithm. So we can simply take $V(x,c)$ to be the resulting path corresponding to the branching encoded in c. □

There is a certain connection between the class NP and quantum computing, since quantum process looks (superficially) non-deterministic computation as one can input a superposed state and run all branches at once. There is, however, a catch: We cannot, in general, distinguish the last state we would obtain efficiently!

For example, if we want to solve SAT of $f : B_n \to B_1$ with this idea, i.e. applying its quantum oracle U_f to the uniform superposition

$$H^{\otimes n} |0\rangle^{\otimes n} = \frac{1}{\sqrt{2^n}} \sum_{x \in B_n} |x\rangle$$

Then we get

$$|f\rangle = \frac{1}{\sqrt{2^n}} \sum_{x \in B_n} |x\rangle |f(x)\rangle$$

But we cannot easily decide SAT from $|f\rangle$, despite the fact that it is indeed a superposition of the outcome of all branches of the obvious non-deteministic algorithm for SAT. Just take $f_0 \equiv 0$ against $f_1 = 1_{x_1}$ (for some fixed $x_1 \in B_n$), then $|f_0\rangle$ and $|f_1\rangle$ would be exponentially close to each other, so exponentially many samples would be needed to distinguish them.

Of course, the search problem can be greatly simplified if the search space has some sort of structure to it. For example, if the space (say with size 2^n) is a linearly ordered database and we are looking for the position of a certain datum, then n queries would be sufficient (and also necessary) using classical binary search. In the quantum context, the optimal number queries is between $0.220n$ and $0.526n$, which is faster than the classical case but not by too much.

Unstructured search (i.e. search problems where the query of one item does not tell one anything about other items in the search space) are of course harder, and therefore more interesting. Suppose we have such a search problem whose search space (say with size $N = 2^n$) has at most one good item. The problem of interest is to find the good item (if it exists) with some constant probability in $(0, 1)$. Classically, N queries is necessary and sufficient. In the quantum realm, Grover's algorithm (1996), which we'll talk about in a moment, gives a solution with $O(\sqrt{N})$ queries. It can be shown that this is optimal, but we'll not do that here.

Represent the database as an oracle for $f : B_n \to B_1$ (so a lookup is just a query) with $f(x) = 1$ iff x is good. We can simply consider the case where there does exist some good $x_0 \in B_n$, as we can easily check if any output $x \in B_n$ of an algorithm indeed has $f(x) = 1$.

We will make use of the gate on n-qubits given by

$$I_f = I_{|x_0\rangle} |x\rangle = \begin{cases} |x\rangle & \text{if } f(x) = 0 \\ -|x\rangle & \text{if } f(x) = 1 \end{cases}$$

Making such a gate might take a long time, but let's ignore that problem for now.

Note that if we knew x_0, then we can write $I_{|x_0\rangle} = I - |x_0\rangle\langle x_0|$. Also, if $U_f |x\rangle |y\rangle = |x\rangle |y \oplus f(x)\rangle$ is the usual quantum oracle for f, then we have $U_f(|x\rangle |-\rangle) = (I_{|x_0\rangle} |x\rangle) |-\rangle$. In general, for any state $|\alpha\rangle$, $I_{|\alpha\rangle} = I - 2 |\alpha\rangle \langle\alpha|$ is the reflection in the hyperplane that is the orthogonal complement of $\operatorname{span}\{|\alpha\rangle\}$.

Grover's algorithm goes as follows: As usual we initialise at the uniform superposition $|\psi_0\rangle = H^{\otimes n}|0\rangle^{\otimes n}$. Consider the Grover iteration operator on n-qubits given by $Q = -H^{\otimes n}I_0H^{\otimes n}I_f$ (where $I_0 = I_{|0\rangle^{\otimes n}}$) which takes one query to the oracle to produce.
Both $|\psi_0\rangle$ and Q have real components, so we can use our geometric intuition. Let $P(x_0)$ be the real plane spanned by $|x_0\rangle$ and $|\psi_0\rangle$, then Q preserves $P(x_0)$ and is a rotation on it with angle 2α where $\sin\alpha = 1/\sqrt{N} = \langle x_0|\psi_0\rangle$. Indeed, for any unitary U one has $UI_{|\psi\rangle}U^\dagger = I - 2U|\psi\rangle\langle\psi|U^\dagger = I_{U|\psi\rangle}$. In particular, $Q = -H^{\otimes n}I_0H^{\otimes n}I_f = -I_{|\psi_0\rangle}I_{|x_0\rangle}$. Note also that $I_{|\psi\rangle}|\xi\rangle = |\xi\rangle - 2|\psi\rangle\langle\psi|\xi\rangle$, i.e. $I_{|\psi\rangle}$ modifies $|\psi\rangle$ by a multiple of $|\xi\rangle$. Applying this to the form of Q we obtained earlier shows that Q preserves $P(x_0)$. It is also clear that Q acts as $-I$ in the orthogonal complement of $P(x_0)$. On $P(x_0)$, $I_{|x_0\rangle}$ is the reflection across the line perpendicular to $|x_0\rangle$ and $I_{|\psi_0\rangle}$ is the one across the line perpendicular to $|\psi_0\rangle$. Elementary Euclidean geometry then shows that Q is in fact the rotation with angle 2α.
So what we'll do is to repeatly apply Q to $|\psi_0\rangle$ to rotate it "near" x_0 and then measure it. The initial angle β between $|x_0\rangle$ and $|\psi_0\rangle$ is β with $\cos\beta = \langle x_0|\psi_0\rangle = 1/\sqrt{N}$. To move $|\psi_0\rangle$ within α of $|x_0\rangle$, we just need to apply Q approximately

$$\frac{\arccos(1/\sqrt{N})}{2\arcsin(1/\sqrt{N})} = \frac{\beta}{2\alpha}$$

times. This is asymptotically $(\pi/4)\sqrt{N}$ by an elementary estimation. After the process, we measure the state under the standard basis, which will give the correct answer $|x_0\rangle$ with probability at least $\cos^2\alpha = 1 - 1/N$ since we must arrive at some state that's at most α away from $|x_0\rangle$.
It's not hard to make the justification entirely algebraic (which in fact was how Grover did it at first), but that's not as inspiring and hence left as an exercise.

Example 4.6. If we take $N = 4$ (so $n = 2$), then the initial angle is $\beta = \arccos(1/\sqrt{4}) = \pi/3$ and $2\alpha = 2\arcsin(1/2) = \pi/3 = \beta$, so one iteration of Q will map $|\psi_0\rangle$ exactly onto $|x_0\rangle$, so we can complete the task with certainty with mere one query!

Of course, one can generalise the algorithm to more general search situations. If there are multiple good items $x_1, \ldots, x_k$, then we can still consider

$$I_f|x\rangle = \begin{cases} -|x\rangle & \text{if } f(x) = 1, \text{ i.e. } x \in \{x_1, \ldots, x_k\} \\ |x\rangle & \text{otherwise} \end{cases}$$

and make use of certain properties of $Q = -H^{\otimes n}I_0H^{\otimes n}I_f$. One can show that Q preserves the plane spanned by $|\psi_0\rangle$ and $k^{-1/2}\sum_{i=1}^k|x_i\rangle$ on which it is the rotation with angle 2α with $\sin\alpha = \sqrt{k/N}$. For large N and $k << N$, it takes $O(\sqrt{N/k})$ iterations for a good item to pop up in the measurement with high probability.
The value of k does not affect the algorithm itself, so we can run it without knowledge of k. One do need some more probabilistic trick to effectively bound the error, but it is provable that one can still achieve $O(\sqrt{N})$ complexity.
This can also help us to prepare interesting states. If, say, x is good iff it divides N, then we finish with something that would be quite close to the uniform

superposition of all divisors of N, which is tricky to produce on its own.
The details of these can be seen in example sheet.
Grover's algorithm is indeed optimal as one can prove

Theorem 4.5. *Let A be a quantum algorithm that solves the unique search problem with constant probability $1-\epsilon$ with T queries. Then T is at least $O(\sqrt{N})$. In fact, $T > c\sqrt{N}$ for any $c < \pi/4$.*

5 Shor's Quantum Factoring Algorithm

5.1 Factorisation as a Periodicity Problem

Given an integer N (and $n = O(\log N)$ digits), we want to find a factor $K \in \{2, \ldots, N-1\}$ of it with high success probability and polynomial time complexity. The best known classical algorithm has runtime $\exp(O(n^{1/3}(\log n)^{2/3}))$, which is still superpolynomial.
As we've mentioned, the idea is to reduce the problem to a periodicity problem. How would we do this? The first step is to choose $1 < a < N$ uniformly random and compute $\gcd(a, N)$ (which can be done efficiently by Euclid's algorithm). If this is not 1, then we are already done. Otherwise, $f : \mathbb{Z} \to \mathbb{Z}/N\mathbb{Z}$ by $f(k) = a^k \pmod{N}$ would have period r where r is the least positive integer such that $a^r \equiv 1 \pmod{N}$ (i.e. r is the order of a modulo N). Clearly f is one-to-one in each period. Also, we can compute f in time polynomial in the number of digits of k by the trick of binary expansion and repeated squaring.
Finding r is classically hard. However, if r is known, even, and $a^{r/2}+1$ isn't divisible by N, then one can factor N in polynomial time: We have $a^r - 1 = (a^{r/2}+1)(a^{r/2}-1)$ is divisible by N but $a^{r/2}-1$ isn't (by definition of r), so $\gcd(N, a^{r/2} \pm 1)$ are nontrivial factors of N.
What is the probability of this happening? We have the following theorem from number theory:

Theorem 5.1. *If N is odd and not a prime power, and $a \in \{2, \ldots, N-1\}$ is chosen randomly, then the probability of the above happening is at least $1/2$ (in fact at least $1 - 2^{-m}$ where m is the number of prime divisors of N).*

Hence, suppose that N is as in the theorem and we repeat the process K times, then the probability of us failing to get a factor in this way is at most $1/2^K$. The requirement of N being even, of course, doesn't matter at all. The assumption that N is not a prime power can also be remedied: It is well-known that there's a poly-time classical algorithm that computed c given $N = c^l$.
So, if we can find an efficient algorithm that computes r, then we can solve the problem. This is a periodicity problem.

5.2 The Algorithm

Casting the quantum period-finding algorithm directly doesn't quite work: The domain of f is $\mathbb{Z}$, which is infinite. We definitely need to truncate the domain, but as we don't a priori have knowledge of r we can't guarantee that this truncation doesn't mess up any period.
We choose the domain $D = \{0, 1, \ldots, 2^m - 1\}$ where 2^m is the least power of 2 that exceeds N^2. The reason for such a choice will become clear later.

Write $2^m = Br + b$ for some $0 \leq b < r$. So we have B full period and one corrupted period with length b. $B > N$ as $r < N$ and $2^m > N^2$.
Let's apply the period-finding algorithm blindly and see what happens. Recall that we first made the superposition

$$|f\rangle = \frac{1}{\sqrt{2^m}} \sum_{x \in B_m} |x\rangle \, |f(x)\rangle$$

Next, we would measure the second register which would give some $y = f(x_0)$ and collapse the state into $|\text{per}\rangle \, |y\rangle$ with

$$|\text{per}\rangle = \frac{1}{\sqrt{A}} \sum_{k=0}^{A-1} |x_0 + kr\rangle \, , A = A(x_0) = \begin{cases} B+1 & \text{if } x_0 < b \\ B & \text{if } x_0 \geq b \end{cases}$$

We continue this journey blind and further apply QFT_{2^m} to it, which gives some $\text{QFT}_{2^m} \, |\text{per}\rangle = \sum_{c \in D} \tilde{f}(c) \, |c\rangle$ for some $\tilde{f}$. To get $\tilde{f}$, we simply recall that

$$\text{QFT}_N \, |x_0 + kr\rangle = \frac{1}{\sqrt{2^m}} \sum_{c \in D} \exp\left(\frac{2\pi i(x_0 + kr)c}{2^m}\right) |c\rangle$$

which gives

$$\tilde{f}(c) = \frac{\omega^{cx_0}}{\sqrt{A}\sqrt{2^m}} (1 + \alpha + \cdots + \alpha^{A-1}), \omega = e^{2\pi i/2^m}, \alpha = e^{2\pi i cr/2^m}$$

Which $c \in D$ will we get with good probability if we measure $\text{QFT}_{2^m} \, |\text{per}\rangle$ under the computational basis?
When we derived the period-finding algorithm, we found that for $r \mid 2^m$ we have $\tilde{f}(c) = 0$ except when c is a multiple of $2^m/r$. Geometrically c is indeed a multiple of $2^m/r$, then $\alpha = 1$ and the summation $1 + \alpha + \alpha^2 + \cdots + \alpha^{A-1}$ happens on a straight line from 0 (which somewhat looks like a constructive interference). When this is not the case, then the sum is the sum of the vertices of a regular A-gon on the unit circle which eventually cancels out perfectly (a destructive interference).
Of course, we don't have the luxury of such nice situations in general. Nonetheless, we can attempt to use this geometric intuition to visualise the general case. For those c with $cr/2^m$ close to the integers $0, 1, \ldots, r-1$, the powers $1, \alpha, \ldots, \alpha^{A-1}$ are expected to be close to the real axis, which gives a good chance of constructive interference. When would this happen? As c increases, $cr/2^m$ increases by steps with small (but uniform) step size $r/2^m$. So for $k \in \{0, 1, \ldots, r-1\}$, we can choose a uniqiue $c = c_k$ such that $\xi = c_k r/2^m - k \in [-r/2^{m+1}, r/2^{m+1})$. We have $\alpha^A = e^{2\pi i(k+\xi)A} = e^{2\pi i \xi A}$ which is almost always in the right half-plane, whcih hints that $1 + \alpha + \cdots + \alpha^{A-1}$ almost never cancels and is expected to be quite large. One can formalise this (exercise!) as the following theorem.

Theorem 5.2. *Suppose* $\text{QFT}_{2^m} \, |\text{per}\rangle$ *is measured, then the probability of the outcome* c_k *(that is,* $|\tilde{f}(c)|^2$*) must exceed* γ/r *where* $\gamma = 4/\pi^2 \approx 0.4$.

That is, the measurement will give some c_k with constant probability approximately 0.4.

How would we get r out of the mere information about the value of some $c = c_k$? We have the estimate

$$\left|\frac{c_k}{2^m} - \frac{k}{r}\right| \leq \frac{1}{2^{m+1}} < \frac{1}{2N^2}$$

We claim that there is at most one fraction q that satisfies $|c_k/2^m - q| < 1/2N^2$ and has minimal denominator strictly smaller than N. Indeed, suppose we have two such fractions, say $q_1 = k_1/r_1$ and $q_2 = k_2/r_2$ with $r_1, r_2 < N$, then either they are the same or

$$\frac{1}{N^2} > \left|\frac{c_k}{2^m} - q_1\right| + \left|\frac{c_k}{2^m} - q_2\right| \geq \left|\frac{k_1}{r_1} - \frac{k_2}{r_2}\right| = \frac{|k_1r_2 - k_2r_1|}{r_1r_2} \geq \frac{1}{r_1r_2} > \frac{1}{N^2}$$

which is a contradiction. So, for k coprime to r, we can in theory extract r from the known values c_k, m, N. As we will see later, there is a polynomial-time algorithm to do this.
How many tries do we need to make the probability of reaching some k coprime to r constant in N? We of course use Theorem 4.3 to conclude that it has order at most $O(\log\log N)$, which is good enough.

5.3 Continued Fractions

Given $c < 2^m$, we still need an efficient algorithm to compute k/r such that $|c/2^m - k/r| < 1/2N^2$ and $r < N$. The brute-force way is, of course, a terrible idea since it would take $O(N^2)$ time which is exponential in $\log N$. Amazingly, the theory of continued fractions gives an efficient way to do this.
Given a real number θ, its complete quotients (a_n) and partial quotients (θ_n) are computed in the following way: Take $\theta_0 = \theta, a_0 = \lfloor\theta\rfloor$. Once θ_n, a_n are found and $\theta_n \neq a_n$, we take $\theta_{n+1} = 1/(\theta_n - a_n)$ and $a_{n+1} = \lfloor\theta_{n+1}\rfloor$. If $\theta_n = a_n$, then we simply stop the algorithm.
The theory of continued fractions of irrational numbers is beautiful, but that's not going to help us here. We only need the trivial case of continued fractions of rational numbers.
Easily, for rational θ, the algorithm is simply the Euclidean algorithm on its numerator and denominator, and hence must terminate (and in fact its time complexity is linear in the logarithmic height of θ). Suppose it terminates at the l^{th} step, then we can write

$$\theta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_{l-1} + \cfrac{1}{a_l}}}}}$$

We write formulas in the form of the right hand side of the above identity as $[a_0; a_1, \ldots, a_l]$. When $a_0 = 0$ (which we will henceforth assume to be the case), we simply write the formula as $[a_1, \ldots, a_l]$.

Definition 5.1. Suppose $\theta \in (0, 1)$ has complete quotients (a_n). The k^{th} convergent of θ is the (reduced) fraction $p_k/q_k = [a_1, \ldots, a_k]$.

Example 5.1. $29/51 = [1,1,3,7]$ has convergents $[1] = 1, [1,1] = 1/2, [1,1,3] = 4/7, [1,1,3,7] = 29/57$.

Turns out, the convergents of $c/2^m$ will help us recover k/r. Expectedly, there's an efficient way to compute them.

Lemma 5.3. *For real numbers $a_1, \ldots, a_l$, we have $[a_1, \ldots, a_k] = p_k/q_k$ where*

$$p_0 = 0, q_0 = p_1 = 1, q_1 = a_1, p_k = a_k p_{k-1} + p_{k-2}, q_k = a_k q_{k-1} + q_{k-2}$$

Furthermore, $q_k p_{k-1} - p_k q_{k-1} = (-1)^k$, in particular $\gcd(p_k, q_k) = 1$.

Proof. Induction with the fact that $[a_1, \ldots, a_k, a_{k+1}] = [a_1, \ldots, a_k + a_{k+1}^{-1}]$. □

Theorem 5.4. *If $0 \leq s < t$ are (WLOG coprime) m-bit integers and $s/t = [a_1, \ldots, a_l]$. Then $l = O(m)$ and all convergents $p_k/q_k = [a_1, \ldots, a_k]$ can be computed in $O(m^3)$ time.*

Proof. Computing the complete quotients takes $O(m)$ time so it doesn't affect anything.
Necessarily all a_k, p_k, q_k are positive integers, so $p_k \geq 2p_{k-2}$ and $q_k \geq 2q_{k-2}$ by the recurrence. Consequently, p_k and q_k must both grow exponentially in k, hence $l = O(m)$. The addition and multiplication of m-bit integers both take $O(m^2)$ time, so p_k, q_k can be computed in $O(m)O(m^2) = O(m^3)$ time. □

As $m = O(\log N)$, we can compute all these quantities in $O((\log N)^3)$ time.

Theorem 5.5. *For $\theta \in (0,1)$, $|\theta - p/q| < 1/(2q^2)$ only when p/q is a convergent of θ.*

Proof. Omitted. □

Remark. Although it is not what we care about here, continued fractions of irrational numbers are also quite fun, and the preceding theorem is true in both cases.
For example, when $\theta = (\sqrt{5}+1)/2$ is the golden ratio, then its complete quotients are simply $[1; 1, 1, 1, \ldots]$ and its convergents, which are fractions of successive fibonacci numbers $1/1, 2/1, 3/2, 5/3, 8/5, 13/8, \ldots$, are the "best" rational sequence approximating it by the theorem.
We also have $\sqrt{2} = [1; 2, 2, 2, \ldots]$. In both of these cases, the sequence of complete fractions is (eventually) periodic. In fact, a theorem of Lagrange says that a real number has periodic complete quotients if and only if it is the root of a rational quadratic.
As for other irrationals, we have $e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, \ldots]$. Euler actually showed that we can write e as another kind of continued fractions

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{2}{3 + \cfrac{3}{4 + \cfrac{4}{5 + \cfrac{5}{\ddots}}}}}}$$

Let's get back to Shor's algorithm. How would our knowledge about the convergents of $c/2^m$ help us to find k/r? We have

$$\left|\frac{c}{2^m} - \frac{k}{r}\right| < \frac{1}{2N^2} < \frac{1}{2r^2}$$

So, by the preceding theorem, k/r must be a convergent of $c/2^m$. We don't need any more fancy tricks: Just try all of the convergents (during the iteration that finds them), and one of them must be the desired k/r.

5.4 A Worked Example and Final Remarks

Let's factor $N = 39$ with Shor's algorithm. Say, possibly after some tries, we have chosen $a = 7$ which is coprime with 39. As $2^{10} < 39^2 < 2^{11}$, we take $m = 11$. Suppose the measurement of $\mathrm{QFT}_{2^m} \left|\mathrm{per}\right\rangle$ gives $c = 853$, then by our discussion c has a good probability to satisfy $|853/2048 - k/r| < 1/4096$. Is $c = 853$ lucky? We don't know, but we are gonna try. We have $853/2048 = [2, 2, 2, 42, 4]$ and its convergents are $[2] = 1/2, [2, 2] = 2/5, [2, 2, 2] = 5/12, [2, 2, 2, 42] = 212/509, [2, 2, 2, 42, 2] = 853/2048$, amongst which $k/r = 5/12$ satisfies the desired inequality. With good probability k and r are coprime, so if we are lucky then $r = 12$. We can check $7^{12} \equiv 1 \pmod{39}$, so we are indeed lucky here. Hence 39 divides $(7^6 + 1)(7^6 - 1)$. One can compute $\gcd(7^6 + 1, 39) = 13$ and $\gcd(24, 39) = 3$ which are, luckily, nontrivial factors of N.

If we were unlucky at any of these steps, we just start over with another random value of a.

Overall, Shor's algorithm has $O((\log N)^3)$ complexity for a fixed requirement on success probability. Funnily enough, the slowest part in Shor's algorithm is the classical part.

After Shor formulated this algorithm in 1994, many people have tried to make use of the same technique on other problems. In Shor's original paper, he used it to solve the discrete logarithm problem (i.e. breaking RSA) which you will come across in example sheet.

There's also the hidden subgroup problem: Given the oracle for $f : G \to Y$ (where G is a group) such that there is a subgroup $H < G$ such that f is constant on H and is distinct on distinct cosets of H, we want to find a description of H in time polynomial in $\log |G|$. The periodicity problem is a special case of this by taking $G = (\mathbb{Z}/N\mathbb{Z}, +), H = \{0, r, 2r, \ldots, (N/r - 1)r\}$. Sadly, the hidden subgroup problem is not quite solvable with this idea for nonabelian groups, since the Fourier transforms associated to them are usually expensive to compute.