Quantum Algorithms and Learning Theory
Notes and Exercises

Faris Sbahi

Abstract

In this document, we provide a presentation of the latest results in quantum learning theory alongside theoretical extensions. We also provide experimental analyses of quantum feature maps which can be used for supervised learning.

The first part of our paper is a review: First, we present an overview of quantum computation and information. Next, we present a review of the brief history of quantum machine learning. The subsequent part of our paper is an analysis of recent results in quantum learning theory: (1) information theoretic bounds on quantum computation learning, (2) supervised learning using hybrid quantum-classical circuits, and (3) Tang's [35] idea of least-square sampling providing parallel classical algorithms for quantum machine learning algorithms that solve singular value transformation problems.

The last part of our paper provides new results on quantum feature maps which seek to solve the quantum encoding problem by encoding data inputs into a quantum state that implicitly performs the feature map given by a kernel function. Therefore, if the kernel is sufficiently difficult to evaluate classically, then there may exist a quantum advantage. Hence, we provide a geometric analysis of the properties of a kernel that may provide quantum advantage, and provide experimental results to demonstrate the robustness of particular candidate maps.

Contents

1 Quantum Mechanics

Reference: Chapter 2 of [29]

Exercise 1.1. *(2.11) Find the eigenvectors and eigenvalues of the Pauli matrices.*

Proof.

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

$$X - \lambda I = \begin{bmatrix} -\lambda & 1 \\ 1 & -\lambda \end{bmatrix}$$

$$\lambda^2 - 1 = 0$$

$$\lambda_\pm = \pm 1$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} v_+ = 0$$

$$\Rightarrow v_+ = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

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

$$\Rightarrow v_- = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Similarly, Y has eigenvalues ± 1 with respective eigenvectors $\left\{ \begin{bmatrix} 1 \\ i \end{bmatrix}, \begin{bmatrix} 1 \\ -i \end{bmatrix} \right\}$. Z has eigenvalues ± 1 with respective eigenvectors $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ □

1.1 Postulates of Quantum Mechanics

First, we cover the fundamental postulates of quantum mechanics.

1.1.1 State Space and State Vector

Associated with an isolated physical system is a Hilbert space, $\mathcal{H}$. A Hilbert space is a complete inner-product vector space. Note that completeness holds trivially in a finite-dimensional vector space because we have closure with respect to all sequences (and hence any Cauchy sequence in the vector space must converge to a vector in the same space). Nevertheless, the state space of a physical system may be infinite-dimensional.

A system is completely described by a unit vector $u \in \mathcal{H}$ called the state vector.

For example, consider a system given by a single qubit, which has a two-dimensional state space. Let $|0\rangle$ and $|1\rangle$ be an orthonormal basis for this space. Hence, a state vector in this space is given by

$$|\psi\rangle = a\,|0\rangle + b\,|1\rangle$$

where $a, b \in \mathbb{C}$ and $|a|^2 + |b|^2 = 1$.

1.1.2 Evolution

The evolution of a closed quantum system is described by a unitary transformation. Recall that an operator U is unitary iff $U^\dagger U = I = UU^\dagger$ (and hence preserves inner products[1]).

So, let the state of a system at time t_1 be given by $|\psi\rangle$ and $|\psi'\rangle$ at t_2. Hence,

$$|\psi'\rangle = U\,|\psi\rangle$$

Exercise 1.2. *(2.51) Verify that H is unitary*

$$\begin{aligned} HH^\dagger &= 1/2 \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= 1/2 \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \\ &= I = H^\dagger H \end{aligned}$$

[1] and furthermore has a spectral decomposition because it is normal

Exercise 1.3. *(2.52) Verify that $H^2 = I$*
Because H is hermitian we have that

$$\begin{aligned} H^2 &= HH^\dagger \\ &= I \end{aligned}$$

from H being unitary.

Exercise 1.4. *(2.53) What are the eigenvalues and eigenvectors of H?*

$$\begin{aligned} \det \frac{1}{\sqrt{2}} \begin{pmatrix} 1-\lambda & 1 \\ 1 & -1-\lambda \end{pmatrix} &= -(1-\sqrt{2}\lambda)(1+\sqrt{2}\lambda) - 1 = 0 \\ &= -1 + 2\lambda^2 - 1 \\ \lambda &= \pm 1 \\ \frac{1}{\sqrt{2}} \begin{pmatrix} 1-\sqrt{2} & 1 \\ 1 & -1+\sqrt{2} \end{pmatrix} v_1 &= 0 \\ (1-\sqrt{2})v_{11} + v_{12} &= 0 \\ v_{11} - (1-\sqrt{2})v_{12} &= 0 \\ v_1 &= \begin{pmatrix} 1+\sqrt{2} \\ 1 \end{pmatrix} \\ \frac{1}{\sqrt{2}} \begin{pmatrix} 1+\sqrt{2} & 1 \\ 1 & -1-\sqrt{2} \end{pmatrix} v_2 &= 0 \\ v_2 &= \begin{pmatrix} 1-\sqrt{2} \\ 1 \end{pmatrix} \end{aligned}$$

1.1.3 Evolution in Continuous Time

Schrodinger's equation provides the time evolution of the state of a quantum system

$$i\hbar \frac{d\,|\psi\rangle}{dt} = H\,|\psi\rangle \tag{1}$$

where H is the (Hermitian) Hamiltonian of the closed system. Because the Hamiltonian is Hermitian it has spectral decomposition

$$H = \sum_E E\,|E\rangle\,\langle E|$$

where E is the energy eigenvalue corresponding to energy eigenstate $|E\rangle$.

For example, consider the Hamiltonian $H = \hbar\omega X$ (recall that $X = \sigma_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ from A.1). Hence, we solve for its eigenvalues and eigenvectors

$$\begin{aligned}
\det\left\{\hbar\omega\begin{pmatrix}-\lambda & 1\\ 1 & -\lambda\end{pmatrix}\right\} &= 0\\
\lambda^2 - 1^2 &= 0\\
\lambda &= \pm 1\\
\Rightarrow E_\pm &= \pm\hbar\omega\\
\hbar\omega\begin{pmatrix}-1 & 1\\ 1 & -1\end{pmatrix}|E_+\rangle &= 0\\
|E_+\rangle &= \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \coloneqq |+\rangle\\
|E_-\rangle &= |-\rangle
\end{aligned}$$

Onwards, notice that we can solve Schrodinger's equation (1) and have

$$|\psi(t_2)\rangle = \exp\left[\frac{-iH(t_2-t_1)}{\hbar}\right]|\psi(t_1)\rangle$$

and equivalently from 1.1.2 we can represent this transformation with unitary operator $U = \exp\left[\frac{-iH(t_2-t_1)}{\hbar}\right]$. This holds in general and so we can consider the two descriptions from 1.1.2 and 1.1.3 interchangeably (the authors prefer the latter).

Exercise 1.5. *(2.54) Suppose* $[A, B] = 0$ *and* A, B *are Hermitian. Prove that* $\exp(A)\exp(B) = \exp(A+B)$

Proof. From Theorem 2.2, A and B are simultaneously diagonalizable. Hence, there is a common set of orthonormal eigenvectors $\{|i\rangle\}$. Hence,

$A = \sum_i a_i |i\rangle\langle i|, B = \sum_i b_i |i\rangle\langle i|$. So,

$$\exp(A)\exp(B) = \sum_{k'=0}^{\infty}\sum_{i'}\frac{(b_{i'}|i'\rangle\langle i'|)^{k'}}{k'!}\sum_{k=0}^{\infty}\sum_{i}\frac{(a_i|i\rangle\langle i|)^k}{k!}$$

By orthonormality,

$$
\begin{aligned}
&= \sum_i \Big[\sum_{k'=0}^{\infty} \frac{b_i^{k'} \left|i\right\rangle \left\langle i\right|}{k'!} \sum_{k=0}^{\infty} \frac{a_i^k \left|i\right\rangle \left\langle i\right|}{k!} \Big] \\
&= \sum_i \sum_{k'=0}^{\infty} \sum_{k=0}^{\infty} \frac{a_i^k b_i^{k'} \left|i\right\rangle \left\langle i\right|}{k!k'!} \\
&= \sum_i \sum_{l=0}^{\infty} \sum_{k=0}^{l} \frac{a_i^k b_i^{l-k} \left|i\right\rangle \left\langle i\right|}{k!(l-k)!} \\
&= \sum_i \sum_{l=0}^{\infty} \frac{1}{l!} \sum_{k=0}^{l} \binom{k}{l} a_i^k b_i^{l-k} \left|i\right\rangle \left\langle i\right| \\
&= \sum_i \sum_{l=0}^{\infty} \frac{(a_i + b_i)^l}{l!} \left|i\right\rangle \left\langle i\right| \\
&= \exp(A+B)
\end{aligned}
$$

□

Exercise 1.6. *(2.55) Prove that $U(t_1, t_2)$ is unitary*

Proof. Using the result of 2.54,

$$
\begin{aligned}
UU^\dagger = U^\dagger U &= \exp\left[\frac{-iH(t_2 - t_1)}{\hbar}\right] \exp\left[\frac{iH(t_2 - t_1)}{\hbar}\right] \\
&= \exp\big(\hat{0}\big) \\
&= I
\end{aligned}
$$

□

Exercise 1.7. *(2.56) Use the spectral decomposition to show that $K \coloneqq -i \log U$ is Hermitian for any unitary U and thus $U = \exp(iK)$ for some Hermitian K*

Proof. The eigenvalues of U can be given as $\exp(i\theta)$ by unitary. Furthermore, from spectral theorem, U is diagonalizable as $U = V\Lambda V^\dagger$ where V is unitary[2]. Hence, $U = V\Lambda V^\dagger$ where diagonal matrix Λ has elements of the form $\exp(i\theta)$ across the diagonal.

Furthermore, $(V\Lambda V^\dagger)^n = V\Lambda V^\dagger V\Lambda V^\dagger \cdots V\Lambda V^\dagger = V\Lambda^n V^\dagger \Rightarrow \exp\big(V\Lambda V^\dagger\big) = V\exp(\Lambda)V^\dagger$. Therefore, let $\Lambda' = \log(\Lambda)$ which therefore has elements of the form $i\theta$. Hence, $U = \exp\big(V\Lambda' V^\dagger\big)$

$$
\begin{aligned}
K &= -i \log U \\
&= -i \log\big(\exp\big(V\Lambda' V^\dagger\big)\big) \\
&= -iV\Lambda' V^\dagger \\
&= V\Theta V^\dagger
\end{aligned}
$$

[2]Quick proof: U can be written as $U = VTV^\dagger$ where V is unitary and T is upper triangular by Schur Decomposition. However, $UU^\dagger = U^\dagger U, VV^\dagger = I = V^\dagger V \Rightarrow T$ is normal $\Rightarrow T$ is diagonal.

where $\Theta = -i\Lambda'$ has elements of the form θ (and hence the elements are real along the diagonal and zero elsewhere $\Rightarrow \Theta^\dagger = \Theta$). Therefore, $K^\dagger = V^\dagger \Theta^\dagger V = V\Theta V^\dagger = K$. $\square$

1.1.4 Quantum Measurement

Quantum measurements are described by a collection of measurements operators $\{M_m\}$ (where the index m refers to the potential measurement outcomes of the experiment) which act on the state space of the system being observed.

Hence, if the pre-measurement state is $|\psi\rangle$, then

$$p(m) = \langle\psi| M_m^\dagger M_m |\psi\rangle$$

and the post-measurement state is

$$\frac{M_m |\psi\rangle}{\sqrt{p(m)}}$$

Furthermore, $\{M_m\}$ satisfy the completeness equation

$$\sum_m M_m^\dagger M_m = I$$

An important example of a measurement is the measurement of a qubit in the computational basis. This is a measurement of a qubit with two outcomes defined by the measurement operators $M_0 = |0\rangle\langle 0|$ and $M_1 = |1\rangle\langle 1|$.

Now, we see an interesting implication. If we seek to distinguish our physical system from a set of orthogonal states, then we can reliably do so by simply defining each measurement operator to be the outer product of our states of interest. We add a final operator defined to be the remaining complement of the identity in order to satisfy the completeness equation.

On the flipside, two non-orthogonal states $|\psi_1\rangle$ and $|\psi_2\rangle$ necessarily share a parallel component in their orthogonal decomposition. Hence, a measurement outcome that corresponds to the pre-measurement state being $|\psi_1\rangle$ with probability $p = 1$ has a probability $p' > 0$ of having been in state $|\psi_2\rangle$.

Exercise 1.8. *(2.57) Suppose $\{L_l\}$ and $\{M_m\}$ are two sets of measurement operators. Show that a measurement defined by the measurement operators $\{L_l\}$ followed by a measurement defined by the measurement operators $\{M_m\}$ is physically equivalent to a single measurement defined by measurement operators $\{N_{lm}\}$ with the representation $N_{lm} = M_m L_l$.*

Proof. Let $|\varphi\rangle$ be our initial state and recall that if l is measured then the post-measurement state is given by $\frac{L_l|\psi\rangle}{\sqrt{p(l)}}$. Furthermore, if we then measure m we have $\frac{M_m(L_l|\psi\rangle)}{\sqrt{p(m)}\sqrt{p(l)}} = \frac{N_{lm}|\psi\rangle)}{\sqrt{p(m)p(l)}}$.

Now,

$$
\begin{aligned}
p(m)p(l) &= \langle\psi| L_l^\dagger L_l |\psi\rangle \frac{\langle\psi| L_l^\dagger}{\sqrt{p(m)}} M_m^\dagger M_m \frac{L_l |\psi\rangle}{\sqrt{p(m)}} \\
&= p(l) \frac{\langle\psi| L_l^\dagger M_m^\dagger M_m L_l |\psi\rangle}{p(l)} \\
&= \langle\psi| N_{lm}^\dagger N_{lm} |\psi\rangle \\
&= p(lm)
\end{aligned}
$$

Hence, $\frac{N_{lm}|\psi\rangle}{\sqrt{p(m)p(l)}} = \frac{N_{lm}|\psi\rangle}{\sqrt{p(lm)}}$. Therefore, the representation is physically equivalent. □

1.1.5 Projective Measurements

There exists a special class of quantum measurements known as projective measurements. These measurements can be described by an observable M, a hermitian operator on the state space being observed. M has spectral decomposition

$$M = \sum_m m P_m$$

where P_m is the projector onto the eigenspace of M with eigenvalues m. Furthermore, if the pre-measurement state is $|\psi\rangle$, then

$$p(m) = \langle\psi| P_m |\psi\rangle$$

and the post-measurement state is

$$\frac{P_m |\psi\rangle}{\sqrt{p(m)}}$$

This simplifies the formula for the expected value of a measurement

$$
\begin{aligned}
\langle M\rangle &= \sum_m m p(m) \\
&= \langle\psi| \left(\sum_m m P_m\right) |\psi\rangle \\
&= \langle\psi| M |\psi\rangle
\end{aligned}
$$

Exercise 1.9. *(2.58) Suppose we prepare a quantum system in an eigenstate $|\psi\rangle$ of some observable M, with corresponding eigenvalue m. What is the average observed value of m and the standard deviation?*

Proof. First,

$$\begin{aligned}\langle M\rangle &= \langle\psi| M |\psi\rangle \\ &= \langle\psi| m |\psi\rangle = m\end{aligned}$$

Furthermore,

$$\begin{aligned}\langle M^2\rangle - \langle M\rangle^2 &= \langle\psi| M^2 |\psi\rangle - m^2 \\ &= \langle\psi| M^\dagger M |\psi\rangle - m^2 \\ &= m^2 - m^2 = 0\end{aligned}$$

□

For example, consider projective measurements on the system given by single qubits with observable Pauli matrix Z. Hence, Z has eigenvalues $+1$ and -1 and eigenstates $|0\rangle$ and $|1\rangle$, respectively. So, consider state $|\psi\rangle = \frac{|0\rangle+|1\rangle}{\sqrt{2}} = |+\rangle \Rightarrow p(+1) = \langle+|0\rangle\langle 0|+\rangle = \frac{1}{2}$. Similarly, $p(-1) = \frac{1}{2}$.

More generally, suppose v is an arbitrary 3-d vector. We can define an observable

$$v \cdot \sigma = v_1\sigma_x + v_2\sigma_y + v_3\sigma_z$$

Exercise 1.10. *(2.59) Suppose we have a qubit in the state $|0\rangle$, and we measure the observable X. What is the average value of X? What is the standard deviation of X?*

Proof. X has eigenvalues $+1$ and -1 and eigenstates $|+\rangle$ and $|-\rangle$, respectively. Hence,

$$\begin{aligned}\langle X\rangle &= \langle\psi| X |\psi\rangle \\ &= \langle\psi| (|+\rangle\langle+| - |-\rangle\langle-|) |\psi\rangle \\ &= \langle 0|+\rangle\langle+|0\rangle - \langle 0|-\rangle\langle-|0\rangle \\ &= \frac{1}{2} - \frac{1}{2} = 0\end{aligned}$$

Furthermore,

$$\begin{aligned}\langle M^2\rangle - \langle M\rangle^2 &= \langle\psi| M^2 |\psi\rangle - 0 \\ &= \langle\psi| (|+\rangle\langle+| + |-\rangle\langle-|) |\psi\rangle \\ &= \frac{1}{2} + \frac{1}{2} = 1\end{aligned}$$

□

Exercise 1.11. *(2.60) Show that $v \cdot \sigma$ has eigenvalues ± 1 and that the projectors onto the corresponding eigenspaces are given by $P_\pm = (I \pm v \cdot \sigma)/2$.*

Proof. First, $v \cdot \sigma$ is Hermitian so it's spectral decomposition is given by $v \cdot \sigma = U \Lambda U^\dagger$ for some unitary U, diagonal matrix Λ. Hence, using $(v \cdot \sigma)^2 = I$ we have

$$\begin{aligned} I = (v \cdot \sigma)^2 &= (U \Lambda U^\dagger)^2 \\ &= U \Lambda^2 U^\dagger \\ \Rightarrow U^\dagger I U &= \Lambda^2 \\ I &= \Lambda^2 \end{aligned}$$

Therefore, Λ must have diagonal entries ± 1.

Next, $P_i P_j = \delta_{ij} P_j$ since if $i \neq j$ then $(I + v \cdot \sigma)(I - v \cdot \sigma) = I - (v \cdot \sigma)^2 = I - I = 0$. Furthermore, $P_+ + P_- = (I + v \cdot \sigma)/2 + (I - v \cdot \sigma)/2 = I$.

Finally, $(+1)P_+ + (-1)P_- = (I + v \cdot \sigma)/2 - (I - v \cdot \sigma)/2 = v \cdot \sigma$. □

Exercise 1.12. *(2.61) Calculate the probability of obtaining result $+1$ for a measurement of $v \cdot \sigma$, given that the state prior to measurement is $|0\rangle$. What is the state of the system after measurement if $+1$ is obtained?*

Proof. First,

$$\begin{aligned} p(+1) &= \langle \psi | P_+ | \psi \rangle \\ &= \langle \psi | (I + v \cdot \sigma)/2 | \psi \rangle \\ &= 1 + \frac{1}{2}[v_1 \langle 0 | X | 0 \rangle + v_2 \langle 0 | Y | 0 \rangle + v_3 \langle 0 | Z | 0 \rangle] \\ &= 1 + \frac{1}{2}[v_1 \langle 0 | 1 \rangle + i v_2 \langle 0 | 1 \rangle + v_3 \langle 0 | 0 \rangle] \\ &= 1 + \frac{v_3}{2} \end{aligned}$$

Furthermore, after measurement of $+1$ we have

$$\begin{aligned} (I + v \cdot \sigma)/2 |0\rangle &= |0\rangle + \frac{1}{2}[v_1 |1\rangle + i v_2 |1\rangle + v_3 |0\rangle] \\ &= \left[\left(\frac{v_3}{2} + 1 \right) |0\rangle + \frac{v_1 + i v_2}{2} |1\rangle \right] / \sqrt{1 + \frac{v_3}{2}} \end{aligned}$$

□

1.1.6 POVM measurements

POVMs are best viewed as a special case of the general measurement formalism, providing the simplest means to study post-measurement statistics without knowledge of the post measurement state.

From above, $p(m) = \langle \psi | M_m^\dagger M_m | \psi \rangle$ so if we define $E_m \coloneqq M_m^\dagger M_m$ (which is hence positive from A.3) then these E_m's are sufficient for the purpose of computing probabilities. We denote $\{E_m\}$ as a POVM. POVMs also satisfy the completeness relation.

Note that projective operators are the special case of being equivalent to their respective POVM element because $E_m = P_m^\dagger P_m = P_m$.

Exercise 1.13. *(2.62) Show that any measurement where the measurement operators and the POVM elements coincide is a projective measurement*

Proof. We would then have $M_m = E_m = M_m^\dagger M_m$. Furthermore, E_m is a positive operator $\Rightarrow M_m = M_m^\dagger M_m = M_m M_m^\dagger = M_m^\dagger$ so M_m is Hermitian. Hence, $M_m = M_m^2$ so the measurement is projective. □

Nevertheless, the POVM formalism is a useful guide in for our intuition in quantum information. Consider if Alice prepares some state for Bob that is either $|\psi_1\rangle = |0\rangle$ or $|\psi_2\rangle = \frac{|0\rangle+|1\rangle}{\sqrt{2}}$. Recall from 1.1.4, Bob can't determine which state was prepared with full certainty (because of the shared orthogonal component $|0\rangle$). Still, we can define a POVM[3]

$$\begin{aligned}
E_1 &= \frac{\sqrt{2}}{1+\sqrt{2}} |1\rangle\langle 1| \\
E_2 &= \frac{\sqrt{2}}{1+\sqrt{2}} \frac{(|0\rangle - |1\rangle)(\langle 0| - \langle 1|)}{2} \\
E_3 &= I - E_1 - E_2
\end{aligned}$$

Now, notice what happens.

$$\begin{aligned}
\langle\psi_1| E_1 |\psi_1\rangle &= \langle 0| \frac{\sqrt{2}}{1+\sqrt{2}} |1\rangle\langle 1|0\rangle \\
&= 0 \\
\langle\psi_2| E_1 |\psi_2\rangle &= \frac{\langle 0| + \langle 1|}{\sqrt{2}} \frac{\sqrt{2}}{1+\sqrt{2}} |1\rangle\langle 1| \frac{|0\rangle + |1\rangle}{\sqrt{2}} \\
&= \frac{\sqrt{2}}{2\sqrt{2}+2} > 0
\end{aligned}$$

Hence, if we observe E_1 after the measurement described by $\{E_1, E_2, E_3\}$, then Alice must've prepared $|\psi_2\rangle$. Similarly,

$$\begin{aligned}
\langle\psi_1| E_2 |\psi_1\rangle &= \langle 0| \frac{\sqrt{2}}{1+\sqrt{2}} \frac{(|0\rangle - |1\rangle)(\langle 0| - \langle 1|)}{2} |0\rangle \\
&= \frac{\sqrt{2}}{2\sqrt{2}+2} > 0 \\
\langle\psi_2| E_2 |\psi_2\rangle &= \frac{\langle 0| + \langle 1|}{\sqrt{2}} \frac{\sqrt{2}}{1+\sqrt{2}} \frac{(|0\rangle - |1\rangle)(\langle 0| - \langle 1|)}{2} \frac{|0\rangle + |1\rangle}{\sqrt{2}} \\
&= 0
\end{aligned}$$

so if we observe E_2, then Bob concludes that Alice prepared $|\psi_1\rangle$. Our routine is imperfect because we may observe E_3 and hence would infer nothing of the original

[3] verify that completeness and these being positive operators holds

state. Still, we would never *incorrectly* guess given that we allow ourselves to abstain when we see E_3.

Exercise 1.14. *(2.63) Suppose a measurement is described by measurement operators M_m. Show that there exist unitary operators U_m such that $M_m = U_m\sqrt{E_m}$ where E_m is the POVM associated to the measurement.*

Proof. From SVD, we have that $M_m = UDV$ for U, V unitary and D real, diagonal. Hence,

$$\begin{aligned}\sqrt{E_m} &= \sqrt{M_m^\dagger M_m} = \sqrt{V^\dagger D U^\dagger U D V} \\ &= \sqrt{V^\dagger D^2 V} \\ &= V^\dagger D V = V^\dagger U^\dagger U D V \\ &= U_m^\dagger M_m\end{aligned}$$

where $U_m \coloneqq UV$. Therefore, there exists the unitary transformation of interest. □

Exercise 1.15. *(2.64) Suppose Bob is given a quantum state chosen from a set $S = |\psi_1\rangle, \cdots, |\psi_m\rangle$ of linearly independent states. Construct a POVM $\{E_1, \cdots, E_{m+1}\}$ such that if outcome E_i occurs, $1 \leq i \leq m$, then Bob knows with certainty that he was given state $|\psi_i\rangle$.*

Proof. To distinguish the states we require $\langle\psi_i| E_j |\psi_i\rangle = p_i \delta_{ij}$ where $p_i > 0$ and $1 \leq i, j \leq m$.

So, we can use the Gram-Schmidt process using S as our linearly independent set. This will give us an orthonormal set $U = |\varphi_1\rangle, \cdots, |\varphi_m\rangle$ that spans the same subspace as S. Next, we can represent each $|\psi_i\rangle$ in this orthonormal basis, U. Finally, for each i we can find a vector $|\psi_i'\rangle$ in the span of U that is orthogonal to all $|\psi_j\rangle, j \neq i$. Hence, we can define $E_i = |\psi_i'\rangle\langle\psi_i'|, 1 \leq i \leq m$. Finally, take $E_{m+1} = I - \sum_m E_i$.

Creating an optimal POVM is much trickier (in the sense of minimizing the probability p_{m+1}). □

From this exercise, we see that POVMs present a reliable way to distinguish non-orthogonal (but linearly independent) states given that we allow for the slack of an "inconclusive" measurement (E_{m+1}).

Exercise 1.16. *(2.65) Express the states $(|0\rangle + |1\rangle)/\sqrt{2}$ and $(|0\rangle - |1\rangle)/\sqrt{2}$ in a basis in which they are not the same up to relative phase shift.*

Proof. Trivially, the $|+\rangle$ and $|-\rangle$ suffices as a basis where they are not the same up to relative phase shift. □

1.1.7 Composite Systems

The state space of a composite physical system is the tensor product of the state spaces of the component physical systems.

Exercise 1.17. *(2.66) Show that the average value of the observable $X_1 Z_2$ (X acting on the first qubit and Z on the second) for a two qubit system measured in the state $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$ is zero.*

Proof. Let observable $M = X_1 Z_2$. Hence,

$$\begin{aligned}
\langle M \rangle &= \frac{\langle 00| + \langle 11|}{\sqrt{2}} M \frac{|00\rangle + |11\rangle}{\sqrt{2}} \\
&= \frac{\langle 00| + \langle 11|}{\sqrt{2}} \frac{X_1 |0\rangle Z_2 |0\rangle + X_1 |1\rangle Z_2 |1\rangle}{\sqrt{2}} \\
&= \frac{\langle 00| + \langle 11|}{\sqrt{2}} \frac{|1\rangle |0\rangle - |0\rangle |1\rangle}{\sqrt{2}} \\
&= 0
\end{aligned}$$

□

Interestingly, we can show that a general quantum measurement (as described in 1.1.4) can be implemented as a projective measurement coupled with unitary dynamics.

Consider a quantum system with state space Q and measurements M_m on this system. We can introduce an *ancilla* system M with orthonormal basis $|m\rangle$ which is in one-to-one correspondence with the possible outcomes of the measurement we wish to implement.

So, let $|0\rangle$ be a fixed state of M and define an operator U on $|\psi\rangle |0\rangle$ (with $|\psi\rangle$ as a state of Q) by

$$U |\psi\rangle |0\rangle \coloneqq \sum_m M_m |\psi\rangle |m\rangle$$

Hence,

$$\langle \varphi| \langle 0| U^\dagger U |\psi\rangle |0\rangle = \sum_m \sum_{m'} \langle \varphi| M_m^\dagger M_{m'} |\psi\rangle \langle m|m'\rangle$$

So, because the states $|m\rangle$ are orthonormal

$$= \sum_m \langle \varphi| M_m^\dagger M_m |\psi\rangle$$

and finally by the completeness of M_m

$$= \langle \varphi|\psi\rangle$$

This tells us that U preserves inner products between states of the form $|\psi\rangle |0\rangle$. Furthermore, we can show that U can be extended to a unitary operator on $Q \otimes M$ (exercise).

Exercise 1.18. *(2.67)*

Hence, consider a projective measurement on the two systems ($U|\psi\rangle|0\rangle$) given by projectors $P_m \coloneqq I_Q \otimes |m\rangle\langle m|$. So,

$$\begin{aligned} p(m) &= \langle\psi|\langle 0|U^\dagger P_m U|\psi\rangle|0\rangle \\ &= \sum_{m'}\sum_{m''}\langle\psi|M_{m'}^\dagger\langle m'|(I_Q\otimes|m\rangle\langle m|)M_{m''}|\psi\rangle|m''\rangle \\ &= \langle\psi|M_m^\dagger M_m|\psi\rangle \end{aligned}$$

which agrees with the general result from 1.1.4. Similarly, the post-measurement state is as expected. Hence, we've shown that unitary dynamics, projective measurements, and ancillary systems can be used together to describe any general measurement.

Exercise 1.19. *(2.68) Prove that $|\psi\rangle \neq |a\rangle|b\rangle$ for all single qubit state $|a\rangle$ and $|b\rangle$ where $|\psi\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}$.*

Proof. First, decompose the qubit state in their basis, $|a\rangle = \alpha_0|0\rangle + \alpha_1|1\rangle$ and $|b\rangle = \beta_0|0\rangle + \beta_1|1\rangle$. Now, we prove by contradiction

$$\begin{aligned} |a\rangle|b\rangle &= \alpha_0\beta_0|00\rangle + \alpha_0\beta_1|01\rangle + \alpha_1\beta_0|10\rangle + \alpha_1\beta_1|11\rangle \\ &= \frac{|00\rangle+|11\rangle}{\sqrt{2}} \end{aligned}$$

which would imply that either α_0 or β_1 are zero in order to remove the $|01\rangle$ term. However, this would also remove either the $|00\rangle$ or $|11\rangle$ term, so we have a contradiction. □

A state of a composite system having this property is said to be entangled.

1.2 Superdense Coding

Suppose Alice is in possession of two classical bits of information she wishes to transmit to Bob, but is only allowed to send a single qubit to Bob.

Now, suppose that Alice and Bob initially share a pair of qubits in the entangled state from above

$$|\psi\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}$$

where Alice is initially holding the first qubit and Bob the second. She can then apply a particular gate to send a bit string. Below shows the corresponding gate and resulting state

Bit String	Applied gate	Resulting state
00	–	$\frac{\|00\rangle+\|11\rangle}{\sqrt{2}}$
01	Z	$\frac{\|00\rangle-\|11\rangle}{\sqrt{2}}$
10	X	$\frac{\|10\rangle+\|01\rangle}{\sqrt{2}}$
11	iY	$\frac{-\|10\rangle+\|01\rangle}{\sqrt{2}}$

Observe that these are the Bell states (see A.2). Furthermore, Bell states form an orthonormal basis and hence can be distinguished (as we've discussed in 1.1.6). Hence, Alice needs only to interact with the single qubit to transmit two classical bits of information to Bob.

Exercise 1.20. *(2.69) Verify that the Bell basis forms an orthonormal basis for the two qubit state space.*

Proof. Two qubit state space consists of states of the form $|\psi\rangle = a\,|00\rangle + b\,|01\rangle + c\,|10\rangle + d\,|11\rangle$. Evidently, $|00\rangle = \frac{\sqrt{2}}{2}\left[\frac{|00\rangle+|11\rangle}{\sqrt{2}} + \frac{|00\rangle-|11\rangle}{\sqrt{2}}\right]$, $|01\rangle = \frac{\sqrt{2}}{2}\left[\frac{|10\rangle+|01\rangle}{\sqrt{2}} - \frac{-|10\rangle+|01\rangle}{\sqrt{2}}\right]$ and similarly for the others. Hence, we span the same space.

Furthermore, $\langle\beta_{00}|\beta_{00}\rangle = \frac{\langle 00|+\langle 11|}{\sqrt{2}}\frac{|00\rangle+|11\rangle}{\sqrt{2}} = (\langle 00|00\rangle + \langle 11|11\rangle)/2 = 1$. Also, $\langle\beta_{00}|\beta_{01}\rangle = \frac{\langle 00|+\langle 11|}{\sqrt{2}}\frac{|00\rangle-|11\rangle}{\sqrt{2}} = (\langle 00|00\rangle - \langle 11|11\rangle)/2 = 0$. The other combinations follow similarly.

Therefore, we have an orthonormal basis. □

Exercise 1.21. *(2.70) Suppose E is any positive operator acting on Alices qubit. Show that $\langle\psi|\,E\otimes I\,|\psi\rangle$ takes the same value when $|\psi\rangle$ is any of the four Bell states. Suppose some malevolent third party ('Eve') intercepts Alices qubit on the way to Bob in the superdense coding protocol. Can Eve infer anything about which of the four possible bit strings 00, 01, 10, 11 Alice is trying to send? If so, how, or if not, why not?*

Proof.

$$\begin{aligned}
\langle 00| + \langle 11|\,(E\otimes I)\,|00\rangle + |11\rangle &= \langle 0|\,E\,|0\rangle + \langle 1|\,E\,|1\rangle \\
\langle 00| - \langle 11|\,(E\otimes I)\,|00\rangle - |11\rangle &= \langle 0|\,E\,|0\rangle + \langle 1|\,E\,|1\rangle \\
\langle 10| + \langle 01|\,(E\otimes I)\,|10\rangle + |01\rangle &= \langle 0|\,E\,|0\rangle + \langle 1|\,E\,|1\rangle \\
-\langle 10| + \langle 01|\,(E\otimes I) - |10\rangle + |01\rangle &= \langle 0|\,E\,|0\rangle + \langle 1|\,E\,|1\rangle
\end{aligned}$$

Hence, Eve can't infer anything. The states are only distinguishable if one can perform a measurement that acts on both qubits. □

1.3 The Density Operator

An alternative formulation of quantum mechanics is possible using a tool known as the density operator.

Suppose a quantum system is one of a number of states $|\psi\rangle$ with probability p_i. We call $\{p_i, |\psi_i\rangle\}$ an ensemble of pure states. The density operator is defined

$$\rho \coloneqq \sum_i p_i \, |\psi_i\rangle \, \langle\psi_i|$$

Evolution of the density operator (under a unitary transformation) can be derived readily

$$\sum_i p_i U \, |\psi_i\rangle \, \langle\psi_i| \, U^\dagger = U \rho U^\dagger$$

If we perform a measurement with operator M_m with initial state $|\psi_i\rangle$ then

$$\begin{aligned} p(m \mid i) &= \langle\psi_i| \, M_m^\dagger M_m \, |\psi_i\rangle \\ &= \operatorname{tr}\big(M_m^\dagger M_m \, |\psi_i\rangle \, \langle\psi_i|\big) \end{aligned}$$

using A.4.
Hence, summing this conditional probability across all initial states we have

$$\begin{aligned} p(m) &= \sum_i p_i \operatorname{tr}\big(M_m^\dagger M_m \, |\psi_i\rangle \, \langle\psi_i|\big) \\ &= \operatorname{tr}\big(M_m^\dagger M_m \rho\big) \end{aligned}$$

The state after obtaining measurement result m on initial state $|\psi_i\rangle$ is

$$|\psi_i^m\rangle = \frac{M_m \, |\psi_i\rangle}{\sqrt{\langle\psi_i| \, M_m^\dagger M_m \, |\psi_i\rangle}}$$

and so the density operator after result m is given by

$$\begin{aligned} \rho_m &= \sum_i p(i \mid m) \, |\psi_i^m\rangle \, \langle\psi_i^m| \\ &= \sum_i p(i \mid m) \frac{M_m \, |\psi_i\rangle \, \langle\psi_i| \, M_m^\dagger}{\langle\psi_i| \, M_m^\dagger M_m \, |\psi_i\rangle} \end{aligned}$$

Furthermore, from Bayes' rule we have that $p(i \mid m) = \frac{p(m|i)p(i)}{p(m)}$ so we can simplify

$$\begin{aligned}
\rho_m &= \sum_i \frac{p(m \mid i)p(i)}{p(m)} \frac{M_m |\psi_i\rangle \langle\psi_i| M_m^\dagger}{\langle\psi_i| M_m^\dagger M_m |\psi_i\rangle} \\
&= \sum_i \frac{p(i) \langle\psi_i| M_m^\dagger M_m |\psi_i\rangle}{\operatorname{tr}\Big(M_m^\dagger M_m \rho\Big)} \frac{M_m |\psi_i\rangle \langle\psi_i| M_m^\dagger}{\langle\psi_i| M_m^\dagger M_m |\psi_i\rangle} \\
&= \sum_i \frac{p(i) M_m |\psi_i\rangle \langle\psi_i| M_m^\dagger}{\operatorname{tr}\Big(M_m^\dagger M_m \rho\Big)} \\
&= \frac{M_m \rho M_m^\dagger}{\operatorname{tr}\Big(M_m^\dagger M_m \rho\Big)}
\end{aligned}$$

A quantum state whose state $|\psi\rangle$ is known exactly is said to be in a pure state. In this case the density operator is simply $\rho = |\psi\rangle \langle\psi|$. Otherwise, ρ is in a mixed state.

A pure state satisfies $\operatorname{tr}\big(\rho^2\big) = 1$ and a mixed state $\operatorname{tr}\big(\rho^2\big) < 1$.

Imagine that our record of the result m of a measurement was lost. We would have a quantum system in the state ρ_m with probability $p(m)$ without knowing the actual value of m. Hence, the system would be described as

$$\begin{aligned}
\rho &= \sum_m p(m) \rho_m \\
&= \sum_m M_m \rho M_m^\dagger
\end{aligned}$$

We may wish to move away from the interpretation of the density operator as a means of describing ensembles of quantum states.

Theorem 1.22. *An operator ρ is a density operator associated to some ensemble $\{p_i, |\psi_i\rangle\}$ if and only if it satisfies the conditions*

1. *ρ has trace equal to one*
2. *ρ is a positive operator*

Proof. We show one direction (see the text for the other). Let ρ be a density operator. Hence,

$$\begin{aligned}
\operatorname{tr}(\rho) &= \sum_i p_i \operatorname{tr}(|\psi_i\rangle \langle\psi_i|) \\
&= \sum_i p_i = 1
\end{aligned}$$

because $\operatorname{tr}(|\psi_i\rangle \langle\psi_i|) = \psi_{i,11}^2 + \psi_{i,22}^2 + \cdots \psi_{i,nn}^2 = 1$ by normalization.

Furthermore, suppose $|\varphi\rangle$ resides in the vector space

$$\begin{aligned}\langle\varphi|\,\rho\,|\varphi\rangle &= \sum_i p_i \,\langle\varphi|\psi_i\rangle\,\langle\psi_i|\varphi\rangle \\ &= \sum_i p_i |\,\langle\varphi|\psi_i\rangle\,|^2 \geq 0\end{aligned}$$

so we have positivity. □

The use of this theorem is that we can define a density operator to be a positive operator with trace one and hence reformulate the postulates of quantum mechanics without speaking of ensembles.

This reformulation shines when describing quantum systems whose state is not known and when describing subsystems a composite quantum system.

Exercise 1.23. *(2.71) Let ρ be a density operator. Show that $\operatorname{tr}\left(\rho^2\right) \leq 1$ with equality iff ρ is a pure state.*

Proof.

$$\begin{aligned}\rho^2 &= \sum_i p_i \,|\psi_i\rangle\,\langle\psi_i| \sum_{i'} p_{i'} \,|\psi_{i'}\rangle\,\langle\psi_{i'}| \\ &= \sum_i p_i^2 \,|\psi_i\rangle\,\langle\psi_i|\end{aligned}$$

by orthonormality. Hence,

$$\operatorname{tr}\rho^2 = \sum_i \sum_j p_i^2 \psi_{i,jj}^2$$

And $\sum_j \psi_{i,jj}^2 = 1$ by normalization

$$= \sum_i p_i^2$$

Now, we have that $\sum_i p_i = 1 \Rightarrow \sum_i p_i^2 = 1 \Leftrightarrow p_i = 1$. If $p_i = 1$, then there is only one index and hence we have a pure state. Otherwise, $\sum_i p_i^2 < 1$ and we have a mixed state. □

Remember that different ensembles of quantum states can give rise to a specific density matrix and hence one must avoid assuming that the eigenvectors and eigenvalues have special significance with regard to the represented ensemble of quantum states.

Nevertheless, there is value in discussing which ensembles give rise to the same density matrix (notably in quantum noise and error correction). Let $\left|\tilde{\psi}_i\right\rangle$ generate ρ i.e. $\rho \coloneqq \sum_i \left|\tilde{\psi}_i\right\rangle\left\langle\tilde{\psi}_i\right|$. Note that $\left|\tilde{\psi}_i\right\rangle = \sqrt{p_i}\,|\psi_i\rangle$ is clearly not necessarily normalized. Now, we have the following theorem.

Theorem 1.24. *The sets* $\left|\tilde{\psi}_i\right\rangle$ *and* $|\tilde{\varphi}_j\rangle$ *generate the same* ρ *if and only if*

$$\left|\tilde{\psi}_i\right\rangle = \sum_j u_{ij} \left|\tilde{\varphi}_j\right\rangle$$

where the matrix with matrix elements u_{ij} *is unitary.*

Proof. See the text. □

Exercise 1.25. *(2.72) Bloch Sphere for mixed states.*
(1) Show that an arbitrary density matrix for a mixed state qubit can be written as

$$\rho = \frac{I + r \cdot \sigma}{2}$$

where r *is a real 3-D vector such that* $\|r\| \leq 1$*. This vector is known as the Bloch vector for the state* ρ*.*

Proof. Let ρ be an arbitrary density matrix, and so $\rho = \sum_i p_i |\psi_i\rangle \langle\psi_i|$. □

(2) What is the Bloch vector representation for the state $\rho = I/2$*?*
(3) Show that a state ρ *is pure iff* $\|r\| = 1$*.*

Proof. □

(4) Show that for pure states the description of the Bloch vector we have given coincides with that in Section 1.2

Proof. □

Exercise 1.26. *(2.73) Let* ρ *be a density operator. A minimal ensemble for* ρ *is an ensemble* $\{p_i, |\psi_i\rangle\}$ *containing a number of elements equal to the rank of* ρ*.*

Proof. □

As mentioned above, density operators are powerful tools for describing subsystems of composite systems.

Suppose we have physical systems A and B, whose state is described by ρ^{AB}. The reduced density operator for system A is defined by

$$\rho^A \coloneqq \mathrm{tr}_B(\rho^{AB})$$

where tr_B is a map of operators known as the partial trace over system B which is defined by

$$\begin{aligned} \mathrm{tr}_B(|a_1\rangle \langle a_2| \otimes |b_1\rangle \langle b_2|) &= |a_1\rangle \langle a_2| \, \mathrm{tr}(|b_1\rangle \langle b_2|) \\ &= |a_1\rangle \langle a_2| \langle b_2|b_1\rangle \end{aligned}$$

Hence, consider the Bell state $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$ and the reduced density operator of its first qubit

$$\begin{aligned}
\rho &= \frac{|00\rangle + |11\rangle}{\sqrt{2}} \frac{\langle 00| + \langle 11|}{\sqrt{2}} \\
\rho^1 &= \frac{|0\rangle\langle 0|\langle 0|0\rangle + |1\rangle\langle 0|\langle 0|1\rangle + |0\rangle\langle 1|\langle 1|0\rangle + |1\rangle\langle 1|\langle 1|1\rangle}{2} \\
&= \frac{|0\rangle\langle 0| + |1\rangle\langle 1|}{2} \\
&= \frac{I}{2}
\end{aligned}$$

Oddly, $\mathrm{tr}\left(\frac{I^2}{4}\right) = 1/2 < 1$ so the first qubit is in a mixed state despite the system as a whole being in a pure state. This is another hallmark of quantum entanglement.

Exercise 1.27. *(2.74) Suppose a composite of systems A and B is in state $|a\rangle|b\rangle$, where $|a\rangle$ is a pure state of system A and $|b\rangle$ is a pure state of system B. Show that the reduced density operator of system A alone is a pure state.*

Proof.

$$\begin{aligned}
\rho &= |a\rangle|b\rangle\langle a|\langle b| \\
\rho^A &= |a\rangle\langle a|\langle b|b\rangle = |a\rangle\langle a|
\end{aligned}$$

where we were given that $|a\rangle$ is a pure state. □

Exercise 1.28. *(2.75) For each of the four Bell states, find the reduced density operator for each qubit*

Proof. First, $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$

$$\begin{aligned}
\rho &= \frac{|00\rangle + |11\rangle}{\sqrt{2}} \frac{\langle 00| + \langle 11|}{\sqrt{2}} \\
\rho^1 &= \frac{|0\rangle\langle 0|\langle 0|0\rangle + |1\rangle\langle 0|\langle 0|1\rangle + |0\rangle\langle 1|\langle 1|0\rangle + |1\rangle\langle 1|\langle 1|1\rangle}{2} \\
&= \frac{|0\rangle\langle 0| + |1\rangle\langle 1|}{2} = \frac{I}{2} \\
\rho^2 &= \frac{|0\rangle\langle 0|\langle 0|0\rangle + |1\rangle\langle 0|\langle 0|1\rangle + |0\rangle\langle 1|\langle 1|0\rangle + |1\rangle\langle 1|\langle 1|1\rangle}{2} \\
&= \frac{|0\rangle\langle 0| + |1\rangle\langle 1|}{2} = \frac{I}{2}
\end{aligned}$$

Next, $\frac{|00\rangle-|11\rangle}{\sqrt{2}}$

$$\rho = \frac{|00\rangle - |11\rangle}{\sqrt{2}} \frac{\langle 00| - \langle 11|}{\sqrt{2}}$$
$$\rho^1 = \frac{|0\rangle\langle 0|\langle 0|0\rangle - |1\rangle\langle 0|\langle 0|1\rangle - |0\rangle\langle 1|\langle 1|0\rangle + |1\rangle\langle 1|\langle 1|1\rangle}{2}$$
$$= \frac{|0\rangle\langle 0| + |1\rangle\langle 1|}{2} = \frac{I}{2}$$
$$\rho^2 = \frac{|0\rangle\langle 0|\langle 0|0\rangle - |1\rangle\langle 0|\langle 0|1\rangle - |0\rangle\langle 1|\langle 1|0\rangle + |1\rangle\langle 1|\langle 1|1\rangle}{2}$$
$$= \frac{|0\rangle\langle 0| + |1\rangle\langle 1|}{2} = \frac{I}{2}$$

The remaining two are similar. □

1.4 Quantum Teleportation

Quantum teleportation is a procedure for sending quantum information from Alice to Bob, given that Alice and Bob share an EPR pair, and have a classic communications channel. Recall that the need for Alice to communicate her result to Bob prevents faster than light communication.

The state to be teleported is $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. We can use the circuit shown in the figure below to perform this telerportation, with inputs $|\psi\rangle|\beta_{00}\rangle$ where $|\beta_{00}\rangle$ is the Bell state $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$. Hence,

$$|\psi\rangle|\beta_{00}\rangle = \frac{1}{\sqrt{2}}\Big[\alpha|0\rangle(|00\rangle + |11\rangle) + \beta|1\rangle(|00\rangle + |11\rangle)\Big]$$

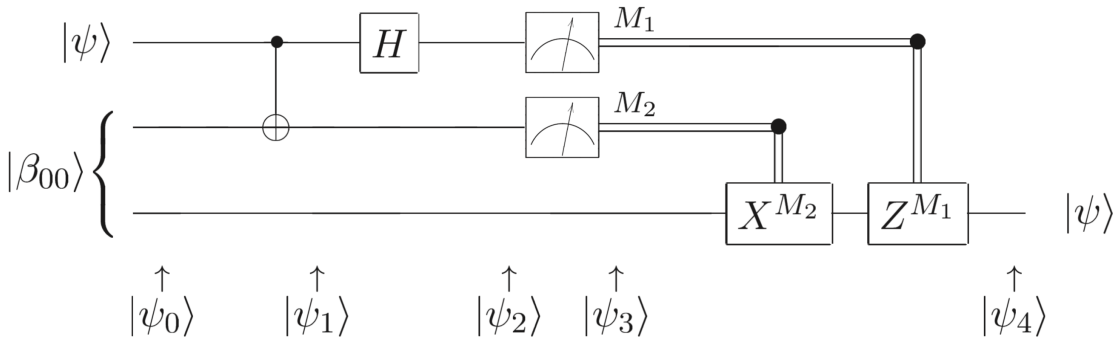


Figure 1: Two top lines are Alice's system and bottom is Bob's

Recall the controlled-`NOT` (`CNOT`) takes $|a\rangle|b\rangle$ to $|a\rangle|b \oplus a\rangle$. The other gates in the circuit are summarized in the diagram below.

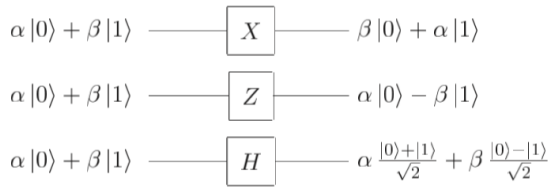


Figure 2: Basic gates

So, the first two qubits belong to Alice and the third to Bob. Alice sends her qubits through a `CNOT` obtaining

$$|\psi_1\rangle = \frac{1}{\sqrt{2}}\Big[\alpha|0\rangle(|00\rangle + |11\rangle) + \beta|1\rangle(|10\rangle + |01\rangle)\Big]$$

Then, Alice's first qubit is sent through a Hadamard gate which gives

$$\begin{aligned}|\psi_2\rangle &= \frac{1}{2}\Big[\alpha(|0\rangle + |1\rangle)(|00\rangle + |11\rangle) + \beta(|0\rangle - |1\rangle)(|10\rangle + |01\rangle)\Big] \\ &= \frac{1}{2}\Big[\,|00\rangle(\alpha|0\rangle + \beta|1\rangle) + |01\rangle(\alpha|1\rangle + \beta|0\rangle) + |10\rangle(\alpha|0\rangle - \beta|1\rangle) + |11\rangle(\alpha|1\rangle - \beta|0\rangle)\Big]\end{aligned}$$

Now, observe that each grouping in this expression has Alice's bits in a different state ($|00\rangle, |01\rangle, |10\rangle, |11\rangle$), each of which Alice may observe after a measurement. Curiously, each of these groupings has a unique corresponding state for Bob's qubits. Hence, we know the state of Bob's qubits given knowledge of the outcome of Alice's measurement.

Furthermore, note that each of the possible states of Bob's qubits after Alice's measurements can be readily transformed to $|\psi\rangle$. Consider the four cases

1. Alice measures 00. Hence, Bob's state is already $|\psi\rangle$.
2. Alice measures 01. Hence, Bob's state is $\alpha|1\rangle + \beta|0\rangle$. So, just apply the X gate.
3. Alice measures 10. Hence, Bob's state is $\alpha|0\rangle - \beta|1\rangle$. So, just apply the Z gate to flip the second sign.
4. Alice measures 11. Hence, Bob's state is $\alpha|1\rangle - \beta|0\rangle$. So, apply the X gate to flip the bits, then the Z gate to flip the second sign.

So that's what the notation in the circuit above means, we apply X or Z to Bob's qubits to recover $|\psi\rangle$ depending on the outcome of Alice's measurement.

Pretty cool, right? Let's look at this a level deeper using the density operator formalism we've developed. Each of the 4 cases have probability $\frac{1}{4}$ of occurring after the measurement. Hence, the density operator is given by

$$\begin{aligned}\rho =& \frac{1}{4}\Big[\,|00\rangle\langle 00|(\alpha|0\rangle + \beta|1\rangle)(\alpha^*\langle 0| + \beta^*\langle 1|) + |01\rangle\langle 01|(\alpha|1\rangle + \beta|0\rangle)(\alpha^*\langle 1| + \beta^*\langle 0|) \\ &+ |10\rangle\langle 10|(\alpha|0\rangle - \beta|1\rangle)(\alpha^*\langle 0| - \beta^*\langle 1|) + |11\rangle\langle 11|(\alpha|1\rangle - \beta|0\rangle)(\alpha^*\langle 1| - \beta^*\langle 0|)\Big]\end{aligned}$$

So, the reduced density operator of Bob's system is

$$\begin{aligned}
\rho^B =& \frac{1}{4}\Big[\langle 00|00\rangle (\alpha|0\rangle + \beta|1\rangle)(\alpha^*\langle 0| + \beta^*\langle 1|) + \langle 01|01\rangle (\alpha|1\rangle + \beta|0\rangle)(\alpha^*\langle 1| + \beta^*\langle 0|) \\
&+ \langle 10|10\rangle (\alpha|0\rangle - \beta|1\rangle)(\alpha^*\langle 0| - \beta^*\langle 1|) + \langle 11|11\rangle (\alpha|1\rangle - \beta|0\rangle)(\alpha^*\langle 1| - \beta^*\langle 0|)\Big] \\
=& \frac{1}{4}\Big[(\alpha|0\rangle + \beta|1\rangle)(\alpha^*\langle 0| + \beta^*\langle 1|) + (\alpha|1\rangle + \beta|0\rangle)(\alpha^*\langle 1| + \beta^*\langle 0|) \\
&+ (\alpha|0\rangle - \beta|1\rangle)(\alpha^*\langle 0| - \beta^*\langle 1|) + (\alpha|1\rangle - \beta|0\rangle)(\alpha^*\langle 1| - \beta^*\langle 0|)\Big] \\
=& \frac{1}{4}\Big[2(\alpha^*\alpha + \beta^*\beta)|0\rangle\langle 0| + 2(\alpha^*\alpha + \beta^*\beta)|1\rangle\langle 1|\Big] \\
=& \frac{I}{2}
\end{aligned}$$

by $|\alpha|^2 + |\beta|^2 = 1$ and completeness.

Hence, the state of Bob's system after Alice has performed the measurement (but before Bob has learned the measurement result) is $I/2$ which has no dependence upon the state $|\psi\rangle$ being teleported. Therefore, any measurements performed by Bob will contain no information about $|\psi\rangle$, so information being communicated is dependent on the classical communication channel, implying that the speed of light limit is obeyed.

1.5 The Schmidt Decomposition and purifications

Schmidt Decomposition theorem says given a pure state $|\psi\rangle$ in a composite system AB, then there are orthonormal states $|i_A\rangle$ and $|i_B\rangle$ in A and B, respectively, such that

$$|\psi\rangle = \sum_i \lambda_i |i_A\rangle |i_B\rangle$$

where λ_i is nonnegative real and $\sum_i \lambda_i^2 = 1$.

One readily seen implication is that the spectra of ρ^A and ρ^B are the same, given a pure state in composite system AB.

Exercise 1.29. *(2.76)*

Exercise 1.30. *(2.77)*

Exercise 1.31. *(2.78)*

A second technique is purification. Suppose we are given a state ρ^A of system A. We can then introduce another system R and define a pure state $|AR\rangle$ for the joint system AR such that $\rho^A = \mathrm{tr}_R(|AR\rangle\langle AR|)$. R is simply a reference and has no physical significance, the point is that we can associate pure states with mixed states.

Exercise 1.32. *(2.79)*

Exercise 1.33. *(2.80)*

Exercise 1.34. *(2.81)*

Exercise 1.35. *(2.82)*

1.6 EPR and the Bell Inequality

Imagine we perform the following measurement. Charlie prepares a quantum system of two qubits in the state

$$|\psi\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}}$$

He passes the first bit to Alice and second to Bob. They perform measurements of the following observables

$$\begin{aligned} Q &= Z_1 \\ R &= X_1 \\ S &= \frac{-Z_2 - X_2}{\sqrt{2}} \\ T &= \frac{Z_2 - X_2}{\sqrt{2}} \end{aligned}$$

Alice decides randomly to measure either Q or R once she receives the qubit and similarly Bob decides randomly whether to measure S or T. They perform these measurements at the same time.

Hence, there are 4 combinations of Alice-Bob measurements. We can calculate and show that

$$\begin{aligned} \langle QS\rangle &= \frac{1}{\sqrt{2}} \\ \langle RS\rangle &= \frac{1}{\sqrt{2}} \\ \langle RT\rangle &= \frac{1}{\sqrt{2}} \\ \langle QT\rangle &= \frac{1}{\sqrt{2}} \end{aligned}$$

Proof. □

And so $\langle QS\rangle + \langle RS\rangle + \langle RT\rangle - \langle QT\rangle = 2\sqrt{2}$. This violates Bell's inequality, derived in the text, which says that this value should never exceed 2.

Bell's inequality requires assuming that Q, R, S, T have definite values before the Alice-Bob measurements (realism). Additionally, we assumed that Alice performing the measurement does not influence the result of Bob's measurement (locality). Hence, at least one of these assumptions must be incorrect, since experimentation confirms this quantum picture.

1.7 No-cloning Theorem

It is impossible to copy an unknown quantum state.

Proof. Suppose we have a quantum machine with two slots labelled A and B. Slot A starts out with unknown state $|\psi\rangle$ which is two be copied to B. Assume that B starts out with some pure state $|s\rangle$.

Hence, the initial state of the machine is $|\psi\rangle\,|s\rangle$. So, some unitary evolution U now effects the copying procedure

$$U(|\psi\rangle\,|s\rangle) = |\psi\rangle\,|\psi\rangle$$

Suppose this works for two particular states $|\psi\rangle$ and $|\varphi\rangle$. Hence,

$$\begin{aligned} U(|\psi\rangle\,|s\rangle) &= |\psi\rangle\,|\psi\rangle \\ U(|\varphi\rangle\,|s\rangle) &= |\varphi\rangle\,|\varphi\rangle \end{aligned}$$

Hence, take the inner product of the two equations and

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

Hence, either $\langle\varphi|\psi\rangle$ is 0 or 1. Thus, either $|\psi\rangle = |\varphi\rangle$ (a contradiction to assuming they're distinct) or the two states are orthogonal.

Therefore, a cloning device can only clone states which are orthogonal to one another and so a general quantum cloning device is impossible. □

2 Quantum Circuits

Reference: Chapter 4 of [29]

2.1 Single Qubit Operations

A single qubit in the state $a\,|0\rangle + b\,|1\rangle$ can be visualized as a point (θ, φ) on the unit sphere, where $a = \cos(\theta/2), b = e^{i\varphi}sin(\theta/2)$. This is called the Bloch sphere representation and $(\cos\varphi\sin\theta, \sin\varphi\sin\theta, \cos\theta)$ is called the Bloch vector.

Exercise 2.1. *(4.1) Find the points on the Bloch sphere which correspond to the normalized eigenvectors of the different Paul matrices.*

Proof. Recall that, from Exercise 2.11, X has eigenvalues ±1 with respective eigevectors $\left\{\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}\right\}$. Similarly, Y has eigenvalues ±1 with respective eigenvectors $\left\{\frac{1}{\sqrt{2}}\begin{bmatrix}1\\i\end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-i\end{bmatrix}\right\}$. Finally, Z has eigenvalues ±1 with respective eigenvectors $\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\right\}$.

First, we solve for X.

$\left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\} \Leftrightarrow \{ \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \}$. First, for $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$, we have that $\cos(\theta/2) = \frac{1}{\sqrt{2}}$. Hence, $\theta = \pi/2$. Now, $e^{i\varphi} \sin(\theta/2) = e^{i\varphi}/\sqrt{2} = 1/\sqrt{2}$. Hence, $\varphi = 0$.

Similarly, for the second eigenvector, $\theta = \pi/2$ but $\varphi = -\pi$.

Therefore, for the first eigenvector,

$$\begin{aligned}(\cos\varphi \sin\theta, \sin\varphi \sin\theta, \cos\theta) &= (\cos(0)\sin(\pi/2), \sin(0)\sin(\pi/2), \cos(\pi/2)) \\ &= (1, 0, 0)\end{aligned}$$

And for the second we have,

$$\begin{aligned}&(\cos(\pi)\sin(\pi/2), \sin(\pi)\sin(\pi/2), \cos(\pi/2)) \\ &\qquad = (-1, 0, 0)\end{aligned}$$

Similarly, we find the Bloch vectors $(0, \pm 1, 0)$ for Y and $(0, 0, \pm 1)$ for Z. □

2.1.1 Action by Hadamard on the Bloch Sphere

Note that the above exercise also shows that, on the Bloch sphere, $|0\rangle = (0, 0, 1), |1\rangle = (0, 0, -1), |+\rangle = (1, 0, 0), |-\rangle = (-1, 0, 0)$. This can often aid intuition. For example, we know that Hadamard operator H is defined s.t. $|0\rangle \rightarrow^H |+\rangle$.

Hence, on the Bloch sphere, this transformation is equivalent to $(0, 0, 1) \rightarrow^H (1, 0, 0)$. So, we can define a series of rotations to emulate the action of H, by considering its action on a basis of the Bloch sphere. So we note the additional transformations, $H^2 = I \Rightarrow |+\rangle \rightarrow |0\rangle \Leftrightarrow (1, 0, 0) \rightarrow (0, 0, 1)$ and $H \begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} = \begin{bmatrix} 1 \\ -i \end{bmatrix}$ (up to a global phase) $\Leftrightarrow (0, 1, 0) \rightarrow (0, -1, 0)$.

Geometrically, we can convince ourselves that the following procedure suffices. For example, consider the effect of this procedure on $|0\rangle$:

(1) Begin with state $|0\rangle = (0, 0, 1)$

(2) Rotate by $-\pi/2$ about the $\hat{x}$ axis. Hence, we then have $(0, 1, 0)$.

(3) Rotate by $-\pi/2$ about the $\hat{z}$ axis. This gives $(1, 0, 0)$.

(4) Rotate by $-\pi/2$ about the $\hat{x}$ axis. This keeps us at $(1, 0, 0) = |+\rangle$

Similarly, using the same procedure

(1) $\begin{bmatrix} 1 \\ i \end{bmatrix} = (0, 1, 0)$.

(2) $(0, 0, -1)$.

(3) $(0, 0, -1)$.

(4) $(0, -1, 0)$.

The reader can verify the above for $|+\rangle$.

Exercise 2.2. *(4.2) Let $x \in \mathbb{R}$ and A be a matrix that satisfies $A^2 = I$. Show that*

$$\exp(iAx) = \cos(x)I + i\sin(x)A$$

Proof. From the power series definition of e^z, we have that

$$\begin{aligned}
\exp(iAx) &= \sum_{n=0}^{\infty} \frac{(iAx)^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{(iAx)^{2n}}{(2n)!} + \sum_{n=0}^{\infty} \frac{(iAx)^{2n+1}}{(2n+1)!} \\
\sum_{n=0}^{\infty} \frac{(iAx)^{2n}}{(2n)!} &= \sum_{n=0}^{\infty} \frac{i^{2n} A^{2n} x^{2n}}{(2n)!} \\
&= \sum_{n=0}^{\infty} \frac{(-1)^{2n} I x^{2n}}{(2n)!} \\
&= I \sum_{n=0}^{\infty} \frac{(-1)^{2n} x^{2n}}{(2n)!} = \cos(x) I \\
\sum_{n=0}^{\infty} \frac{(iAx)^{2n+1}}{(2n+1)!} &= \sum_{n=0}^{\infty} \frac{i^{2n+1} A^{2n+1} x^{2n+1}}{(2n+1)!} \\
&= \sum_{n=0}^{\infty} \frac{i(-1)^{2n+1} A x^{2n+1}}{(2n+1)!} \\
&= iA \sum_{n=0}^{\infty} \frac{(-1)^{2n+1} x^{2n+1}}{(2n+1)!} = i \sin(x) A
\end{aligned}$$

□

X, Y, Z give rise to three useful classes of unitary matrices when they are exponentiated, the rotation operators about $\hat{x}$, $\hat{y}$, and $\hat{z}$,

$$\begin{aligned}
R_x(\theta) &\equiv e^{-i\theta X/2} \\
R_y(\theta) &\equiv e^{-i\theta Y/2} \\
R_z(\theta) &\equiv e^{-i\theta Z/2}
\end{aligned}$$

We can use exercise 4.2 to write the above equations more conveniently.

Exercise 2.3. *(4.3) Show that, up to a global phase, the $\pi/8$ gate satisfies $T = R_z(\pi/4)$.*

Proof. Note that

$$T = \begin{bmatrix} 1 & 0 \\ 0 & \exp(i\pi/4) \end{bmatrix} = \exp(i\pi/8) \begin{bmatrix} \exp(-i\pi/8) & 0 \\ 0 & \exp(i\pi/8) \end{bmatrix}$$

Now, using the definition of R_z,

$$
\begin{aligned}
e^{-iZ\frac{\pi}{8}} &= \cos(-\pi/8)I + i\sin(-\pi/8)Z \\
&= \cos(\pi/8)I - i\sin(\pi/8)Z \\
&= \begin{bmatrix} \cos(\pi/8) - i\sin(\pi/8) & 0 \\ 0 & \cos(\pi/8) + i\sin(\pi/8) \end{bmatrix} \\
&= \begin{bmatrix} \exp(-i\pi/8) & 0 \\ 0 & \exp(i\pi/8) \end{bmatrix}
\end{aligned}
$$

□

Exercise 2.4. *(4.4) Express the Hadamard gate H as a product of R_x and R_z rotations and $e^{i\varphi}$ for some φ.*

Proof. In section 2.1.1 we discussed a procedure for expressing H as a product of rotations on the Bloch sphere, by considering its actions on a basis of the Bloch sphere. We showed that $R_x(-\pi/2)R_z(-\pi/2)R_x(-\pi/2)$ suffices. We can verify this result a second way by considering the respective rotation matrices.

We know that $H = \frac{1}{\sqrt{2}}(X + Z)$. Furthermore,

$$
\begin{aligned}
R_x(-\pi/2) &= \begin{bmatrix} \cos(\pi/4) & i\sin(\pi/4) \\ i\sin(\pi/4) & \cos(\pi/4) \end{bmatrix} \\
&= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} = \frac{1}{\sqrt{2}}(I + iX) \\
R_z(-\pi/2) &= \begin{bmatrix} \cos(\pi/4) + i\sin(\pi/4) & 0 \\ 0 & \cos(\pi/4) - i\sin(\pi/4) \end{bmatrix} \\
&= \frac{1}{\sqrt{2}} \begin{bmatrix} 1+i & 0 \\ 0 & 1-i \end{bmatrix} = \frac{1}{\sqrt{2}}(I + iZ)
\end{aligned}
$$

We'll use that

$$
\begin{aligned}
XZ &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\
&= \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = iY \\
ZX &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = -iY \\
\Rightarrow XZ + ZX &= 0 \\
\Rightarrow XZX + ZX^2 &= 0 \\
\Rightarrow XZX &= -Z
\end{aligned}
$$

Note that the above is simply showing that the anti-commutator of X and Z, $\{X, Z\} = 0$. This holds for any pair of distinct Pauli matrices (Exercise 2.41).

Hence,

$$\begin{aligned}
\frac{1}{2\sqrt{2}}(I+iX)(I+iZ)(I+iX) &= \frac{1}{2\sqrt{2}}[I+iX+iZ+i^2ZX+iX+i^2X^2+i^2XZ+i^3XZX] \\
&= \frac{1}{2\sqrt{2}}[I+iX+iZ-ZX+iX-I-XZ+i^3XZX] \\
&= \frac{1}{2\sqrt{2}}[i(X+Z)+iX-iXZX] \\
&= \frac{1}{2\sqrt{2}}[i(X+Z)+iX+iZ] \\
&= \frac{1}{\sqrt{2}}[i(X+Z)]
\end{aligned}$$

which gives the Hadamard transform with phase e^{i0}.

□

Exercise 2.5. *(4.5) Prove that $(\hat{n}\cdot\hat{\sigma})^2 = I$, and use this to verify the following equation*

$$R_n(\theta) \equiv \exp(-i\theta n\cdot\sigma/2) = \cos(\theta/2)I - i\sin(\theta/2)(n_xX + n_yY + n_zZ)$$

Proof. Evidently, $\hat{n}\cdot\hat{\sigma} = (n_xX + n_yY + n_zZ)$ so, recalling that distinct Pauli matrices anti-commute,

$$\begin{aligned}
(n_xX+n_yY+n_zZ)^2 &= n_x^2X^2 + n_xn_yXY + n_xn_zXZ + n_xn_yYX + n_y^2Y^2 + n_yn_zYZ + n_xn_zZX + n_yn_zZY + n_zZ \\
&= (n_x^2+n_y^2+n_z^2)I + n_xn_z(XZ+ZX) + n_yn_z(YZ+ZY) + n_xn_y(XY+YX) \\
&= (n_x^2+n_y^2+n_z^2)I = I
\end{aligned}$$

because $\hat{n}$ is a unit vector.

Therefore, using Exercise 4.2 (Nielsen & Chuang), if we let $A = \hat{n}\cdot\hat{\sigma}$, then the result follows directly. □

Exercise 2.6. *(4.7) Show that $XYX = Y$ and use this to prove that $XR_y(\theta)X = R_y(-\theta)$.*

Proof. From above, we have that distinct Pauli matrices anti-commute. Furthermore, the Pauli matrices are hermitian and unitary $\Rightarrow \sigma_i^2 = 0, i \in \{x, y, z\}$. Hence,

$$\begin{aligned}
XY + YX &= 0 \\
XYX + YX^2 &= 0 \\
XYX + Y &= 0 \\
XYX &= -Y
\end{aligned}$$

So,

$$\begin{aligned} XR_y(\theta)X &= X\big[\cos(\theta/2)I - i\sin(\theta/2)Y\big]X \\ &= \cos(\theta/2)X^2 - i\sin(\theta/2)XYX \\ &= \cos(\theta/2)I + i\sin(\theta/2)Y \\ &= \cos(-\theta/2)I - i\sin(-\theta/2)Y \\ &= R_y(-\theta) \end{aligned}$$

using that $\cos(-x) = \cos(x), \sin(-x) = -\sin(x)$. □

Lemma 2.7. *Suppose U is a unitary operation on a single qubit. Then there exist real numbers $\alpha, \beta, \gamma, \delta$ such that*

$$U = \begin{bmatrix} e^{i(\alpha-\beta/2-\delta/2)}\cos(\gamma/2) & -e^{i(\alpha-\beta/2+\delta/2)}\sin(\gamma/2) \\ e^{i(\alpha+\beta/2-\delta/2)}\sin(\gamma/2) & e^{i(\alpha+\beta/2+\delta/2)}\cos(\gamma/2) \end{bmatrix}$$

Theorem 2.8. *Suppose U is a unitary operation on a single qubit. Then there exist real numbers $\alpha, \beta, \gamma, \delta$ such that*

$$U = e^{i\alpha}R_z(\beta)R_y(\gamma)R_z(\delta).$$

Corollary 2.8.1. *Suppose U is a unitary gate on a single qubit. Then there exist unitary operators A, B, C on a single qubit such that $ABC = I$ and $U = e^{i\alpha}AXBXC$, where α is some overall phase factor.*

Exercise 2.9. *(4.12) Give A, B, C, and α for the Hadamard gate.*

Proof. Using Lemma 2.7 above we can solve, assuming $\gamma = \pi/2$,

$$H = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} e^{i(\alpha-\beta/2-\delta/2)} & -e^{i(\alpha-\beta/2+\delta/2)} \\ e^{i(\alpha+\beta/2-\delta/2)} & e^{i(\alpha+\beta/2+\delta/2)} \end{bmatrix}$$

$$\begin{aligned} \alpha - \beta/2 - \delta/2 &= 0 \\ \alpha - \beta/2 + \delta/2 &= \pi \\ \alpha + \beta/2 - \delta/2 &= 0 \\ (\alpha + \beta/2 + \delta/2) &= \pi \\ \Rightarrow \alpha = \pi/2, \beta = 0, \delta &= \pi \end{aligned}$$

So, the proof of Corollary 2.8.1 in Nielsen & Chuang tells us to set

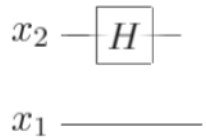
$$
\begin{aligned}
A &= R_z(\beta)R_y(\gamma/2) \\
&= R_z(0)R_y(\pi/4) \\
&= \begin{bmatrix} \cos(\pi/8) & -\sin(\pi/8) \\ \sin(\pi/8) & \cos(\pi/8) \end{bmatrix} \\
B &= R_y(-\gamma/2)R_z(-(\delta+\beta)/2) \\
&= R_y(-\pi/4)R_z(-\pi/2) \\
&= \begin{bmatrix} \cos(\pi/8) & \sin(\pi/8) \\ -\sin(\pi/8) & \cos(\pi/8) \end{bmatrix} \begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix} \\
C &= R_z((\delta-\beta)/2) \\
&= R_z(\pi/2) \\
&= \begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}
\end{aligned}
$$

and α remains set $\alpha = \pi/2$. □

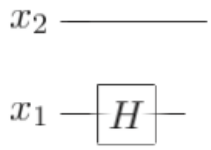
2.2 Controlled Operations

In terms of the computational basis, the action of the `CNOT` is given by $|c\rangle\,|t\rangle \to |c\rangle\,|t \oplus c\rangle$.

Exercise 2.10. *(4.16)*
What is the 4×4 unitary matrix for the circuit



in the computational basis? What is the unitary matrix for the circuit



in the computational basis?

Proof. For the first circuit, we consider action on the computational basis.

$$|x_1\rangle |x_2\rangle \to |x_1\rangle H |x_2\rangle = (I \otimes H) |x_1\rangle |x_2\rangle$$

Now, given that $H = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ w.r.t the computation basis, then

$$(I \otimes H) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Similarly, for the second circuit we have

$$(H \otimes I) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

□

Exercise 2.11. *(4.17) Construct a* `CNOT` *gate from one controlled-Z gate, that is, the gate whose action in the computational basis is specified by the unitary matrix*

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

and two Hadamard gates, specifying the control and target qubits.

Proof. Recall that, in terms of the computational basis, the action of the `CNOT` is given by $|c\rangle |t\rangle \to |c\rangle |t \oplus c\rangle$ and that $Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

We construct our algorithm by first making the observation that $H |+\rangle = |0\rangle$ and $H |-\rangle = |1\rangle$. Hence, beginning with state $|c\rangle |t\rangle$ we can initially apply H to $|t\rangle$. Now, using the control-Z gate with $|c\rangle$ as the control and $H |t\rangle$ as the target, we have two cases:

(1) If $|c = 1\rangle$, then the second qubit will swap either from $|+\rangle$ to $|-\rangle$ or vis versa. Therefore, we can apply another Hadamard to the second qubit and have $|t \oplus c\rangle$ at the second qubit, as expected. The first qubit is unaltered, as expected.

(2) If $|c = 1\rangle$, then the second qubit will remain unchanged. Hence, if we apply another Hadamard to the second qubit, then $|t\rangle$ is recovered since $H^2 = I$. So, we have the expected behavior.

In summary, we have the circuit, beginning with state $|c\rangle |t\rangle$:

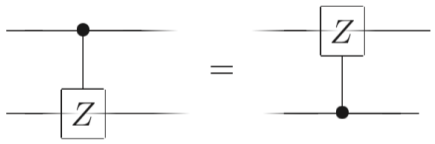
(1) Apply H to the second qubit

(2) Controlled-Z with the first qubit as the control and second as the target

(3) Apply H to the second qubit. □

From the next exercise, we'll see that it didn't actually matter whether we used the first or second qubit as control/target:

Exercise 2.12. *(4.18) Show that*



Proof. We simply prove the statement for the computational basis.

(1) $|0\rangle\,|0\rangle$: Both circuits give the identity transform since they are conditioned on a qubit which is $|0\rangle$, in either case.

(2) $|1\rangle\,|0\rangle$: The first circuit is conditioned on $|1\rangle$, so it applies Z to $|0\rangle$ which gives $|0\rangle$. Hence, we have $|1\rangle\,|0\rangle$. The second circuit is conditioned on $|0\rangle$, so we have the identity transform which gives $|1\rangle\,|0\rangle$, similarly.

(3) $|0\rangle\,|1\rangle$: By symmetry, we have the same outcome as in (2).

(4) $|1\rangle\,|1\rangle$: The first circuit is conditioned on the first $|1\rangle$, so it applies Z to the second qubit which gives $-|1\rangle\,|1\rangle$. Similarly, the second circuit gives $-|1\rangle\,|1\rangle$. □

Exercise 2.13. *(4.19) The* `CNOT` *gate is a simple permutation whose action on a density matrix ρ is to rearrange the elements in the matrix. Write out this action explicitly in the computational basis.*

Proof.

$$|00\rangle\langle 00| = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$|01\rangle\langle 01| = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$|10\rangle\langle 10| = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$|11\rangle\langle 11| = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Now,

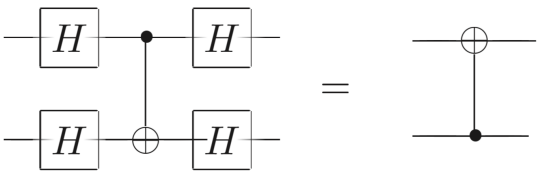
$$C_1(X)\,|00\rangle = |00\rangle$$
$$C_1(X)\,|01\rangle = |01\rangle$$
$$C_1(X)\,|10\rangle = |11\rangle$$
$$C_1(X)\,|11\rangle = |10\rangle$$

Hence, the permutation matrix acting on the computational basis as

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

satisfies this permutation. □

Exercise 2.14. *(4.20) Unlike ideal classical gates, ideal quantum gates do not have (as electrical engineers say) high-impedance inputs. In fact, the role of control and target are arbitrary they depend on what basis you think of a device as operating in. We have described how the* `CNOT` *behaves with respect to the computational basis, and in this description the state of the control qubit is not changed. However, if we work in a different basis then the control qubit does change: we will show that its phase is flipped depending on the state of the target qubit! Show that*



Introducing basis states $|\pm\rangle$ *use this circuit identity to show that the effect of a* `CNOT` *with the first qubit as control and the second qubit as target is as follows:*

$$|+\rangle\,|+\rangle \to |+\rangle\,|+\rangle$$
$$|-\rangle\,|+\rangle \to |-\rangle\,|+\rangle$$
$$|+\rangle\,|-\rangle \to |-\rangle\,|-\rangle$$
$$|-\rangle\,|-\rangle \to |+\rangle\,|-\rangle$$

Thus, with respect to this new basis, the state of the target qubit is not changed, while the state of the control qubit is flipped if the target starts as $|-\rangle$*, otherwise it is left alone. That is, in this basis, the target and control have essentially interchanged roles!*

Proof. Consider action on $|c\rangle\,|t\rangle$ by the circuit on the LHS. The action of this circuit is given by $(H \otimes H)C^1(X)\,|c\rangle\,|t\rangle\,(H \otimes H)$ using c as the control and t as the target for the

controlled operation. So, in Exercise 4.17, we showed that we can decompose $C^1(X)$ as $HC^1(Z)H$ using the same control and target as used for $C^1(X)$ originally, and with the H transforms acting on the target qubit. Hence, we can rewrite action by the LHS circuit as $(H \otimes H)(I \otimes H)C^1(Z) |c\rangle |t\rangle (I \otimes H)(H \otimes H) = (H \otimes I)C^1(Z) |c\rangle |t\rangle (H \otimes I)$.

Similarly, for the circuit on the RHS, action on $|c\rangle |t\rangle$ is given by $C^1(X) |c\rangle |t\rangle$ where in this case t is the control and c is the target. Hence, using the same result, we can rewrite this as $(H \otimes I)C^1(Z) |c\rangle |t\rangle (H \otimes I)$ with t as control and c as target. Finally, using Exercise 4.18, we can swap which qubits we regard as control/target in a controlled-Z operation. Hence, we have the action $(H \otimes I)C^1(Z) |c\rangle |t\rangle (H \otimes I)$ with c as control and t as target, as in the LHS.

Now, using that $H^2 = I$, we note that the identity given by the circuit is equivalent to $C^1(X)(H \otimes H) |c\rangle |t\rangle = C^1(X) |t\rangle |c\rangle (H \otimes H)$ (applying $H \otimes H$ to the end of both circuits). Hence, this directly gives the effect of `CNOT` on the basis $|\pm\rangle$.

□

Exercise 2.15. *(4.21) Verify that Figure 4.8 implements the $C^2(U)$ operation.*

Proof. □

Exercise 2.16.

2.3 Universal quantum gates

A set of gates is said to be universal for quantum computation if any unitary operation may be approximated to arbitrary accuracy by a quantum circuit only involving those gates. We can show

(1) An arbitrary unitary operator may be expressed exactly as a product of unitary operators that each acts non-trivially only on a subspace spanned by two computational basis states

(2) An arbitrary unitary operator may be expressed exactly using single qubit and `CNOT` gates.

(3) Any unitary operation can be approximated to arbitrary accuracy using Hadamard, phase, `CNOT`, and $\pi/8$ gates.

We are showing existence not efficiency.

3 Quantum Fourier Transform

Reference: Chapter 5 of [29]

3.1 Quantum Fourier Transform

The quantum fourier transform on an orthonormal basis $|0\rangle, \cdots, |N-1\rangle$ is defined to be a linear operator with the following action on the basis states,

$$|j\rangle \to \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{2\pi ijk/N} |k\rangle$$

Exercise 3.1. *(5.2) Explicitly compute the Fourier transform of the n qubit state* $|00\cdots0\rangle$.

Proof. $|00\cdots0\rangle$ corresponds to state $|0\rangle$ in the size $N = 2^n$ computational basis. Hence, using the formula above we have

$$\begin{aligned}|0\rangle &\to \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}|k\rangle \\ &= \frac{|0\rangle + |1\rangle + \cdots + |N-1\rangle}{\sqrt{N}}\end{aligned}$$

□

We can derive an alternative product representation of the quantum fourier transform. First, represent some state $|j\rangle$ using its binary representation $j = j_1 j_2 \cdots j_n$, $j_i \in \{0, 1\}$. Then,

$$|j_1, \cdots, j_n\rangle \to \frac{(|0\rangle + e^{2\pi i 0.j_n}|1\rangle)(|0\rangle + e^{2\pi i 0.j_{n-1}j_n}|1\rangle)\cdots(|0\rangle + e^{2\pi i 0.j_1 j_2 \cdots j_n}|1\rangle)}{2^{n/2}}$$

So, define the unitary transformation

$$R_k = \begin{bmatrix}1 & 0 \\ 0 & e^{2\pi i/2^k}\end{bmatrix} \tag{2}$$

Then, using the circuit below, we can see that this transformation is correctly implemented.

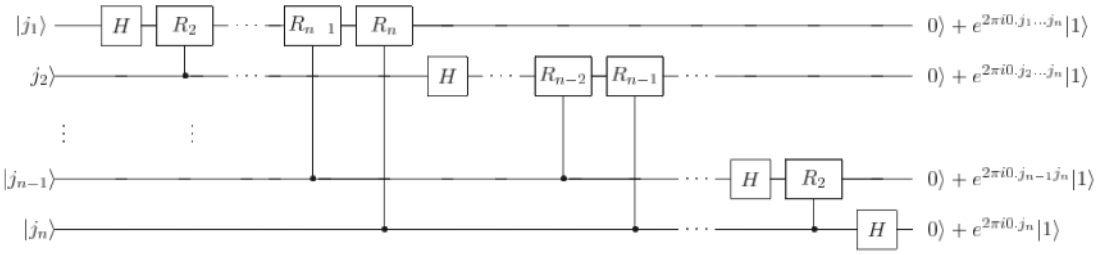


Figure 5.1. Efficient circuit for the quantum Fourier transform. This circuit is easily derived from the product representation (5.4) for the quantum Fourier transform. Not shown are swap gates at the end of the circuit which reverse the order of the qubits, or normalization factors of $1/\sqrt{2}$ in the output.

Furthermore, the gate complexity is $O(n^2)$ as opposed to $O(n2^n)$, classically.

3.2 Quantum Phase Estimation Algorithm

Suppose a unitary operator U has an eigenvector $|u\rangle$ with eigenvalue $e^{2\pi i\varphi}$, where the value of φ is unknown.

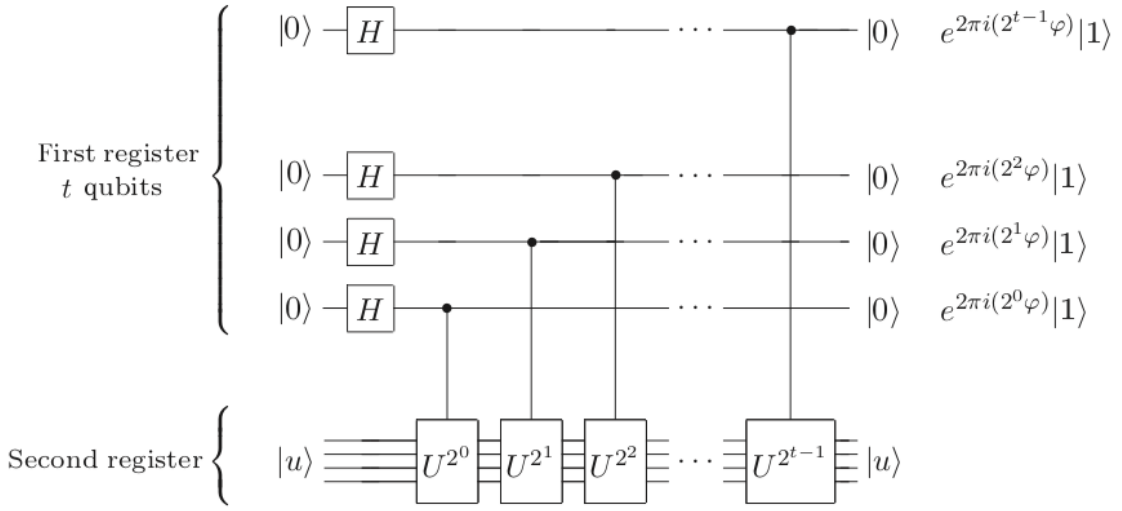


Figure 5.2. The first stage of the phase estimation procedure. Normalization factors of $1/\sqrt{2}$ have been omitted, on the right.

Then, observe that the circuit above gives the state

$$\frac{1}{2^t}(|0\rangle + e^{2\pi i 0.\varphi_t}|1\rangle)(|0\rangle + e^{2\pi i 0.\varphi_{t-1}\varphi_t}|1\rangle)\cdots(|0\rangle + e^{2\pi i 0.\varphi_1\varphi_2\cdots\varphi_t}|1\rangle)$$

Hence, we can apply the inverse QFT and get the state $|\varphi_1\cdots\varphi_t\rangle$, which is an approximation of φ.

Exercise 3.2. *(5.7) Additional insight into the circuit above may be obtained by showing, as you should now do, that the effect of the sequence of controlled-U operations like that in the figure is to take the state $|j\rangle|u\rangle$ to $|j\rangle U^j|u\rangle$. (Note that this does not depend on $|u\rangle$ being an eigenstate of U.)*

Proof. Consider an arbitrary j in its binary representation $j_0 j_1 \cdots j_{t-1}$ where $j_i \in \{0,1\}$. Hence, for each $|j_i\rangle$, the control-U acts on $|j_i\rangle|u\rangle$ such that $|j_i\rangle|u\rangle \mapsto |j_i\rangle U^{j_i 2^i}|u\rangle$. Therefore, the final state is given by

$$\begin{aligned}|j_0\rangle\cdots|j_{t-1}\rangle U^{j_0 2^0}\cdots U^{j_{t-1}2^{t-1}}|u\rangle &= |j\rangle U^{j_0 2^0}\cdots U^{j_{t-1}2^{t-1}}|u\rangle \\ &= |j\rangle U^{j_0 2^0 + j_{t-1}2^{t-1}}|u\rangle \\ &= |j\rangle U^j|u\rangle\end{aligned}$$

□

3.3 Order-Finding and Factoring

4 Quantum Search Algorithms

Reference: Chapter 6 of [29]

4.1 The quantum search algorithm

The Grover iteration may be broken up into four steps.
(1) Apply the oracle O
(2) Apply the Hadamard transform $H^{\otimes n}$
(3) Perform the conditional phase shift on the computer, with every computational basis state except $|0\rangle$ receiving a phase shift of -1,

$$|x\rangle \to -(-1)^{\delta x_0} |x\rangle$$

(4) Apply the Hadamard transform $H^{\otimes n}$

Exercise 4.1. *(6.1) Show that the Unitary operator corresponding to the phase shift in the Grover iteration is* $2|0\rangle\langle 0| - I$.

Proof. Consider arbitrary state $|x\rangle$. There are two cases:
(1) $|x\rangle = |0\rangle$. Hence,

$$\begin{aligned}(2|0\rangle\langle 0| - I)|0\rangle &= 2|0\rangle\langle 0|0\rangle - |0\rangle \\ &= |0\rangle\end{aligned}$$

as expected.
(2) $|x\rangle \neq |0\rangle$. Hence,

$$\begin{aligned}(2|0\rangle\langle 0| - I)|x\rangle &= 2|0\rangle\langle 0|x \neq 0\rangle - |x\rangle \\ &= 0 - |x\rangle = -|x\rangle\end{aligned}$$

as expected. □

Exercise 4.2. *(6.2) Show that the operation* $(2|\psi\rangle\langle\psi| - I)$ *(where* $|\psi\rangle$ *is the equally weighted superposition of states) applied to general state* $\sum_k \alpha_k |k\rangle$ *produces*

$$\sum_k [-\alpha_k + 2\langle\alpha\rangle]|k\rangle$$

where $\langle\alpha\rangle \equiv \sum_k \alpha_k/N$ *is the mean value of* α_k.

Proof.

$$
\begin{aligned}
(2\left|\psi\right\rangle\left\langle\psi\right| - I)\sum_k \alpha_k \left|k\right\rangle &= \Big(2\frac{1}{N^{1/2}}\left|x\right\rangle \frac{1}{N^{1/2}}\sum_{x'=0}^{N-1}\left\langle x'\right| - I\Big)\sum_k \alpha_k \left|k\right\rangle \\
&= 2\frac{1}{N}\sum_{x=0}^{N-1}\left|x\right\rangle\left\langle x'\right|\sum_k \alpha_k \left|k\right\rangle - \sum_k \alpha_k \left|k\right\rangle \\
&= \frac{2}{N}\sum_k \left|k\right\rangle \alpha_k - \sum_k \alpha_k \left|k\right\rangle \\
&= \sum_k [2\langle\alpha\rangle - \alpha_k]\left|k\right\rangle
\end{aligned}
$$

□

5 Entropy and Information

Reference: Chapter 11 of [29]

Exercise 5.1. *(11.8)*
Let X, Y be i.i.d random variables uniformly distributed over set $\{0,1\}$. Hence,

$$
\begin{aligned}
H(X) &= H(1/2) = 1 = H(Y) \\
H(X,Y) &= -4[1/4\log(1/4)] = 2 \\
H(X \mid Y) &= H(X,Y) - H(Y) \\
&= 2 - 1 = 1 \\
&= H(Y \mid X) \\
I(X:Y) &= H(X) - H(X \mid Y) \\
&= 0
\end{aligned}
$$

Now, let $Z = X \oplus Y$. Then,

$$
\begin{aligned}
H(Z) &= H(1/2) = 1 \\
H(X,Y,Z) &= -4[1/4\log(1/4)] = 2 \\
H(X,Y \mid Z) &= H(X,Y,Z) - H(Z) \\
&= 1 \\
I(X,Y:Z) &= H(X,Y) - H(X,Y \mid Z) \\
&= 2 - 1 = 1
\end{aligned}
$$

Furthermore,

$$\begin{aligned} H(X,Z) &= 2 \\ H(X \mid Z) &= H(X,Z) - H(Z) \\ &= 2 - 1 = 1 \\ I(X:Z) &= H(X) - H(X \mid Z) \\ &= 1 - 1 = 0 \end{aligned}$$

Similarly, $I(Y:Z) = 0$*).*

$\therefore I(X:Z) + I(Y:Z) < I(X,Y:Z)$ *in this case.*

Thinking through this problem intuitively, it just says that if we've specified both X, Y*, then we don't need to send a bit for* Z *through our channel since we can compute its value readily. However, if we send just* X *or* Y*, the XOR function provides uniform outcomes across* Z*.*

Exercise 5.2. *(11.9) Let r.v.* X_1 *be uniformly distributed across* $\{0,1\}$*. Furthermore, require that* $X_2 = Y_2 = Y_1 = X_1$ *(identically).*

In this case,

$$\begin{aligned} H(X_1 \mid Y_1) &= H(X_1, Y_1) - H(Y_1) \\ &= H(1/2) - H(1/2) = 0 \\ I(X_1 : Y_1) &= H(X_1) - H(X_1 \mid Y_1) \\ 1 - 0 &= 1 \\ &= I(X_2 \mid Y_2) \end{aligned}$$

However,

$$\begin{aligned} H(X_1, X_2 \mid Y_1, Y_2) &= H(X_1, X_2, Y_1, Y_2) - H(X_1, X_2) \\ &= H(1/2) - H(1/2) = 0 \\ I(X_1 : Y_1) &= H(X_1, X_2) - H(X_1, X_2 \mid Y_1, Y_2) \\ &= 1 - 0 = 0 \end{aligned}$$

$\therefore I(X_1 : Y_1) + I(X_2 : Y_2) > I(X_1, X_2 : Y_1, Y_2)$ *in this case.*

Intuitively, the random variables are distributed identically, so we always only need a single bit to communicate their distribution across a channel. Hence, there will always be a single bit of mutual information across the r.v.'s since their conditional entropy will be zero bits (we know everything we need to know given one variable's value) and their joint entropy will be a single bit.

Conclusion from the above two exercises: mutual information is neither sub-additive nor super-additive.

6 Quantum Information Theory

Reference: Chapter 12 of [29]

Theorem 6.1. *Holevo's Theorem*

Let $\{\rho_1, \rho_2, ..., \rho_n\}$ be a set of mixed states and let ρ_X be one of these states drawn according to the probability distribution $P = \{p_1, p_2, \cdots, p_n\}$.

Then, for any measurement described by POVM elements E_Y and performed on $\rho = \sum_X p_X \rho_X$, the amount of accessible information about the variable X knowing the outcome Y of the measurement is bounded from above as follows:

$$I(X:Y) \leq S(\rho) - \sum_i p_i S(\rho_i)$$

where $\rho = \sum_i p_i \rho_i$ and $S(\cdot)$ is the von Neumann entropy.

The quantity on the right hand side of this inequality is called the Holevo information or Holevo χ quantity:

$$\chi := S(\rho) - \sum_i p_i S(\rho_i)$$

7 Quantum Cryptography

Quantum cryptography or quantum key distribution (QKD) is a procedure that enables provably secure distribution of private information.

7.1 Private key cryptography

Private key cryptosystems which employ a one-time pad (OTP) are provably secure i.e. as long as the key strings are secret, Alice and Bob can guarantee that Eve's mutual information with their unencoded information can be made as small as desired regardless of Eve's eavesdropping strategy.

The major difficulty of private key cryptosystems is secure distribution of key bits. The above cryptosystem, using a OTP, requires key bits to be as long as the message and for key bits to not be reused.

Exercise 7.1. *(12.25) Consider a system with n users, any pair of which would like to be able to communicate privately. Using public key cryptography how many keys are required? Using private key cryptography how many keys are required?*

Proof. We'd need a unique OTP for each pair of users if we used private key cryptography. Hence, we'd need $\binom{n}{2}$ keys.

For public key, we'd need a private-public keypair for each user. Hence, we'd need $2n$ keys. □

7.2 Privacy amplification and information reconciliation

Suppose Alice and Bob share correlated random classical bit strings X and Y. Furthermore, suppose we have an upper bound on Eve's mutual information with X and Y. We can use "information reconciliation" and then "privacy amplification" to systematically increase the correlation between their key strings, while reducing Eve's mutual information about the result, to any desired level of security.

Information reconciliation simply entails conducting error-correction over a public channel, fixing errors between X and Y, until Alice and Bob obtain a shared string W. In the end, Eve will also have a string Z partially correlated to W. We can then use privacy amplification to "amplify" from W a subset of bits S whose correlation with Z are below the desired threshold.

8 Classical Learning Theory

9 Boolean Fourier Analysis

10 Randomized Linear Algebra

11 Algorithms for solving linear systems of equations

One such application of Phase Estimation (Section 3.2) is with respect to solving linear systems of equations. This is the so-called HHL algorithm [23].

The general problem statement of a linear system is if we are given matrix A and unit vector $\vec{b}$, then find $\vec{x}$ satisfying, $A\vec{x} = \vec{b}$.

However, assume that instead of solving for x itself, we instead solve for an expectation value $x^T Mx$ for some linear operator M. Hence, one can show that our algorithm has a runtime bound of $O(\log(N)\kappa^2)$, if we can further assume that the linear system is sparse and has a low condition number κ.

So, assume that A in our linear system is an $N \times N$ Hermitian matrix. Notice that this is an "unrestrictive" constraint on A because we can always take non-Hermitian matrix A' and linear system $A'\vec{x} = \vec{b}$ and instead solve $\begin{bmatrix} 0 & A' \\ A'^\dagger & 0 \end{bmatrix} \begin{bmatrix} 0 \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$. Hence, we we will assume that A is Hermitian from here on.

Recall that because A is hermitian $\Rightarrow$ we can perform quantum phase estimation using e^{-iAt} as the unitary transformation. This can be done efficiently if A is sparse.

So, we first prepare $|b\rangle$ (the representation of $\vec{b}$). We assume that this can be done efficiently or that $|b\rangle$ is supplied as an input.

Denote by $|\psi_j\rangle$ the eigenvectors of A with associated eigenvalues λ_j. Hence, we can express $|b\rangle$ as $|b\rangle = \sum_j \beta_j |\psi_j\rangle$. So, we initialize a first register to state $\sum_j \beta_j |\psi_j\rangle$ and second register to state $|0\rangle$. After applying phase estimation, we then have the joint state $\sum_j \beta_j |\psi_j\rangle \left|\widetilde{\lambda}_j\right\rangle$, where $\widetilde{\lambda}_j$ is an approximation of λ_j. We'll assume that this approximation is perfect from here on.

Next we add an ancilla qubit and perform a rotation conditional on the first register while now holds $|\lambda_j\rangle$. The rotation transforms the system to

$$\sum_j \beta_j |\psi_j\rangle |\lambda_j\rangle \left(\sqrt{1 - \frac{C^2}{\lambda_j^2}} |0\rangle + \frac{C}{\lambda_j} |1\rangle\right)$$

for some small constant $C \in \mathbb{R}$ that is $O(1/\kappa)$.

Hence, we can undo phase estimation to restore the second register to $|0\rangle$.

Now, if we measure the ancillary qubit in the computational basis, we'll evidently collapse the state to $|1\rangle$ with some probability. We'd then have

$$\sum_j \frac{C}{\lambda_j}\beta_j |\psi_j\rangle |\lambda_j\rangle |1\rangle = C(A^{-1}|b\rangle)$$

In particular, the probability of getting this result is

$$\begin{aligned}
p(-1) &= \left(\sum_j \beta_j \langle\psi_j| \langle\lambda_j| \left(\sqrt{1-\frac{C^2}{\lambda_j^2}}\langle 0| + \frac{C}{\lambda_j}\langle 1|\right)\right) |1\rangle\langle 1| \left(\sum_j \beta_j |\psi_j\rangle |\lambda_j\rangle \left(\sqrt{1-\frac{C^2}{\lambda_j^2}}|0\rangle + \frac{C}{\lambda_j}|1\rangle\right)\right) \\
&= \sum_j \beta_j \langle\psi_j| \langle\lambda_j| \left(\sqrt{1-\frac{C^2}{\lambda_j^2}}\langle 0| + \frac{C}{\lambda_j}\langle 1|\right) |1\rangle\langle 1| \beta_j |\psi_j\rangle |\lambda_j\rangle \left(\sqrt{1-\frac{C^2}{\lambda_j^2}}|0\rangle + \frac{C}{\lambda_j}|1\rangle\right) \\
&= \sum_j \beta_j \langle\psi_j| \langle\lambda_j| \frac{C}{\lambda_j}\langle 1|1\rangle\langle 1| \beta_j |\psi_j\rangle |\lambda_j\rangle \frac{C}{\lambda_j}|1\rangle \\
&= \sum_j \beta_j^2 \frac{C^2}{\lambda_j^2} \\
&= \|A^{-1}|b\rangle\|^2 C^2 = O(1/\kappa^4)
\end{aligned}$$

Finally, we can make a measurement M whose expectation value $\langle x| M |x\rangle$ corresponds to the feature of x we wish to evaluate.

12 Supervised learning with quantum feature Hilbert spaces

[20]

12.1 Prelude

We are given data from a training set T and a test set S of a subset $\Omega \subset \mathbb{R}^d$. We assume that S and T are drawn from the same input space X. Furthermore, there exists output space $Y = \{-1, +1\}$ and a distribution D on $X \times Y$.

Now, suppose we have a labelling $m : T \cup S \to Y$. Our goal is to use this information to find some approximation function $\tilde{f} : X \to Y$ that minimizes estimation error for function class F. In other words, let true risk for function f be defined as

$$R^{true}(f) = P_{X,Y\sim D}(f(X) \neq Y)$$

Then, estimation error is the difference in true risk between $\tilde{f}$ and optimal choice $f^* = \inf_{f\in F} R^{true}(f)$.

One classical method is using so-called Support Vector Machines (SVM), which construct a separating hyperplane such that the distance to the nearest training observation

(minimum margin) is maximized. Much of the popularity of SVMs can be attributed to its association with the "kernel trick" which maps the data to a higher dimensional space so that it is separable or approximately separable.

Here, we suppose that the data is given classically and we seek to show that, in some cases, we can obtain a quantum advantage by either generating the separating hyperplane in quantum feature space or simply estimating the kernel function.

12.2 Feature Map

Consider the feature vector kernel $K(x, z) = |\langle\Phi(x)|\Phi(z)\rangle|^2$

12.3 Quantum Variational Classification

12.4 Quantum Kernel Estimation

12.5 Non-Trivial Feature Map with Entanglement

12.6 Geometric Analysis of Candidate Feature Maps

12.7 Experimental Simulation of Candidate Feature Maps

13 Singular Value Transformation using Quantum-Inspired Length-Square Sampling Methods

As we've seen, most well-known QML algorithms convert input quantum states to a desired output state or value. Thus, they do not provide a routine to get necessary copies of these input states (a state preparation routine) and a strategy to extract information from an output state. Both are essential to making the algorithm useful.

- We can compute $A^+ |b\rangle = |x_{LS}\rangle$ in $\tilde{O}(log(N)(s^3\kappa^6)/\epsilon)$ time (query complexity)
- Uses a quantum algorithm based on phase estimation and Hamiltonian simulation
- Assumption: A is sparse with low condition number κ. Hamiltonian ($\hat{H}$) simulation is efficient when $\hat{H}$ is sparse. No low-rank assumptions are necessary.
- "Key" assumption: the quantum state $|b\rangle$ can be prepared efficiently.
- What happens if we assume low rank?
- In general, quantum machine learning algorithms convert quantum input states to the desired quantum output states.
- In practice, data is initially stored classically and the algorithm's output must be accessed classically as well.
- Today's focus: A practical way to make comparisons between classical and quantum algorithms is to analyze classical algorithms under ℓ^2 sampling conditions
- Tang: linear algebra problems in low-dimensional spaces (say constant or polylogarithmic) likely can be solved "efficiently" under these conditions

- Many of the initial practical applications of quantum machine learning were to problems of this type (e.g. Quantum Recommendation Systems - Kerendis, Prakash, 2016)
- How can we compare the speed of quantum algorithms with quantum input and quantum output to classical algorithms with classical input and classical output?
- Quantum machine learning algorithms can be exponentially faster than the best standard classical algorithms for similar tasks, but quantum algorithms get help through input state preparation.
- Want a practical classical model that helps its algorithms offer similar guarantees to quantum algorithms, while still ensuring that they can be run in nearly all circumstances one would run the quantum algorithm.
- Solution (Tang): compare quantum algorithms with quantum state preparation to classical algorithms with sample and query access to input.

13.0.1 Definitions

Definition 13.1. *We have "query access" to $x \in \mathbb{C}^n$ if, given $i \in [n]$, we can efficiently compute x_i. We say that $x \in \mathcal{Q}$.*

Definition 13.2. *We have sample **and** query access to $x \in \mathbb{C}^n$ if*

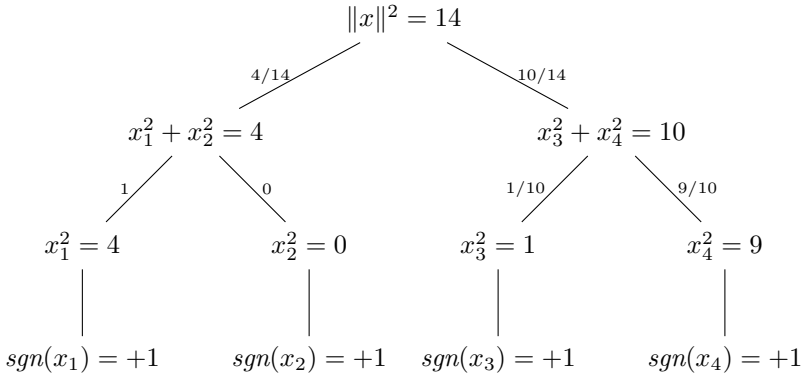
1. *We have query access to x i.e. $x \in \mathcal{Q}$ ($\Rightarrow \mathcal{SQ} \subset \mathcal{Q}$)*
2. *can produce independent random samples $i \in [n]$ where we sample i with probability $|x_i|^2/\|x\|^2$ and can query for $\|x\|$.*

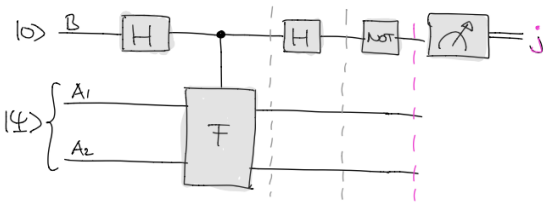
We say that $x \in \mathcal{SQ}$.

Definition 13.3. *For $A \in \mathbb{C}^{m \times n}$, $A \in \mathcal{SQ}$ (abuse) if*

1. *$A_i \in \mathcal{SQ}$ where A_i is the ith row of A*
2. *$\tilde{A} \in \mathcal{SQ}$ for $\tilde{A}$ the vector of row norms (so $\tilde{A}_i = \|A_i\|$).*

Example 13.4. *Say we have the vector $\vec{x} = (2, 0, 1, 3)$ and $\vec{x} \in \mathcal{SQ}$. Consider the following binary tree data structure.*





13.0.2 Low-Rank Estimation

- For $A \in \mathbb{C}^{m \times n}$, given $A \in \mathcal{SQ}$ and some threshold k, we can output a description of a low-rank approximation of A with $\text{poly}(k)$ queries.
- Specifically, we output two matrices $S, \hat{U} \in \mathcal{SQ}$ where $S \in \mathbb{C}^{\ell \times n}$, $\hat{U} \in \mathbb{C}^{\ell \times k}$ ($\ell = \text{poly}(k, \frac{1}{\epsilon})$), and this implicitly describes the low-rank approximation to A, $D := A(S^\dagger \hat{U})(S^\dagger \hat{U})^\dagger$ ($\Rightarrow$ rank $D \leq k$).
- This matrix satisfies the following low-rank guarantee with probability $\geq 1 - \delta$: for $\sigma := \sqrt{2/k}\|A\|_F$, and $A_\sigma := \sum_{\sigma_i \geq \sigma} \sigma_i u_i v_i^\dagger$ (using SVD),

$$\|A - D\|_F^2 \leq \|A - A_\sigma\|_F^2 + \epsilon^2 \|A\|_F^2$$

- Note the $\|A - A_\sigma\|_F^2$ term. This says that our guarantee is weak if A has no large singular values.
- Quantum analog: phase estimation

$$\begin{bmatrix} \cdots A \cdots \end{bmatrix} \begin{bmatrix} S^\dagger \end{bmatrix} \left[\hat{U}\right] \left[\hat{U}^\dagger\right] \left[\cdots S \cdots\right]$$

13.0.3 Trace Inner Product Estimation

- For $x, y \in \mathbb{C}^n$, if we are given that $x \in \mathcal{SQ}$ and $y \in \mathcal{Q}$, then we can estimate $\langle x, y \rangle$ with probability $\geq 1 - \delta$ and error $\epsilon \|x\| \|y\|$
- Quantum analog: SWAP test

Fact 13.5. *For $\{X_{i,j}\}$ i.i.d random variables with mean μ and variance σ^2, let*

$$Y := \underset{j \in [\log 1/\delta]}{\text{median}} \underset{i \in [1/\epsilon^2]}{\text{mean}} X_{i,j}$$

Then $|Y - \mu| \leq \epsilon\sigma$ with probability $\geq 1 - \delta$, using only $O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ samples.

- In words: We may create a mean estimator from $1/\epsilon^2$ samples of X. We compute the median of $\log 1/\delta$ such estimators
- Catoni (2012) shows that Chebyshev's inequality is the best guarantee one can provide when considering pure empirical mean estimators for an unknown distribution (and finite μ, σ)

- "Median of means" provides an exponential improvement in probability of success $(1-\delta)$ guarantee

Corollary 13.5.1. *For $x, y \in \mathbb{C}^n$, given $x \in \mathcal{SQ}$ and $y \in \mathcal{Q}$, we can estimate $\langle x, y \rangle$ to $\epsilon\|x\|\|y\|$ error with probability $\geq 1-\delta$ with query complexity $O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$*

Proof. Sample an **index** s from x. Then, define $Z := x_s y_s \frac{\|y\|^2}{|y_s|^2}$. Apply the Fact with $X_{i,j}$ being independent samples Z. □

13.0.4 Least-Square Sample Generation

- For $V \in \mathbb{C}^{n\times k}, w \in \mathbb{C}^k$, given $V^\dagger \in \mathcal{SQ}$ (*column*-wise sampling of V) and $w \in \mathcal{Q}$, we can simulate $Vw \in \mathcal{SQ}$ with poly(k) queries
- In words: if we can least-square sample the columns of matrix V and query the entries of vector w, then
  1. We can query entries of their multiplication (Vw)
  2. We can least-square sample from a distribution that emulates their multiplication
- Hence, as long as $k \ll n$, we can perform each using a number of steps polynomial in the number of columns of V.

Definition 13.6. *Rejection sampling*

Algorithm 13.7. *Input: Samples from distribution P*
Output: Samples from distribution Q

- *Sample s from P*
- *Compute $r_s = \frac{1}{N}\frac{Q(s)}{P(s)}$, for fixed constant N*
- *Output s with probability r_s and restart otherwise*

Fact 13.8. *Fact. If $r_i \leq 1, \forall i$, then the above procedure is well-defined and outputs a sample from Q in N iterations in expectation.*

Proposition 13.9. *For $V \in \mathbb{R}^{n\times k}$ and $w \in \mathbb{R}^k$, given $V^\dagger \in \mathcal{SQ}$ and $w \in \mathcal{Q}$, we can simulate $Vw \in \mathcal{SQ}$ with expected query complexity $\tilde{O}((\frac{1}{\epsilon^2}\log\frac{1}{\delta}))$*
We can compute entries $(Vw)_i$ with $O(k)$ queries.
We can sample using rejection sampling:

- *P is the distribution formed by sampling from $V_{(\cdot,j)}$.*
- *Q is the target Vw.*
- *Hence, compute r_s to be a constant factor of Q/P*

$$r_i = \frac{\|w^T V_{\cdot,i}\|^2}{\|w\|^2\|V_{\cdot,i}\|^2}$$

- Notice that we can compute these r_i's (in fact, despite that we cannot compute probabilities from the target distribution), and that the rejection sampling guarantee is satisfied (via Cauchy-Schwarz).

- Since the probability of success is $\|Vw\|^2/\|w\|^2$, it suffices to estimate the probability of success of this rejection sampling process to estimate this norm.
- Through a Chernoff bound, we see that the average of $O(\|w\|^2(\frac{1}{\epsilon^2}\log\frac{1}{\delta}))$ "coin flips" is in $[(1-\epsilon)\|Vw\|, (1+\epsilon)\|Vw\|]$ with probability $\geq 1-\delta$.

13.0.5 Application: Stochastic Regression

For a low-rank matrix $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^n$, given $b, A \in \mathcal{SQ}$, (approximately) simulate $A^+b \in \mathcal{SQ}$.

Algorithm 13.10.
- *Low-rank approximation (3) gives us $S, \hat{U} \in \mathcal{SQ}$.*
- *Applying thin-matrix vector (2), we get $\hat{V} \in \mathcal{SQ}$, where $\hat{V} := S^T\hat{U}$; we can show that the columns of $\hat{V}$ behave like the right singular vectors of A.*
- *Let $\hat{U}$ have columns $\{\hat{u}_i\}$. Hence, $\hat{V}$ has columns $\{S\hat{u}_i\}$. Write its ith column as $\hat{v}_i := S\hat{u}_i$.*
- *Low-rank approximation (3) also outputs the approximate singular values $\hat{\sigma}_i$ of A*

Now, we can write the approximate vector we wish to sample in terms of these approximations:

$$A^+b = (A^TA)^+A^Tb \approx \sum_{i=1}^{k} \frac{1}{\hat{\sigma}_i^2}\hat{v}_i\hat{v}_i^TA^Tb$$

- We approximate $\hat{v}_i^TA^Tb$ to additive error for all by noticing that $\hat{v}_i^TA^Tb = \mathrm{tr}\big(A^Tb\hat{v}_i^T\big)$ is an inner product of A^T and $b\hat{v}_i^T$.
- Thus, we can apply (1), since being given $A \in \mathcal{SQ}$ implies $A^T \in \mathcal{SQ}$ for A^T viewed as a long vector.
- Define the approximation of $\hat{v}_i^TA^Tb$ to be $\hat{\lambda}_i$. At this point we have (recalling that $\hat{v}_i := S\hat{u}_i$)

$$A^+b \approx \sum_{i=1}^{k} \frac{1}{\hat{\sigma}_i^2}\hat{v}_i\hat{\lambda}_i = S\sum_{i=1}^{k} \frac{1}{\hat{\sigma}_i^2}\hat{u}_i\hat{\lambda}_i$$

- Finally, using (2) to provide sample access to each $S\hat{u}_i$, we are done ! $\tilde{O}(\kappa^{16}k^6\|A\|_F^6/\epsilon^6)$ complexity.

13.0.6 Definitions and Assumptions

Let $b \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m\times n}$ s.t. $\|A\| \leq 1$ where $\|\cdot\|$ signifies the operator norm (or spectral norm). Furthermore, require that $\mathrm{rank}(A) = k$ and $\|A^+\| \leq \kappa$ where A^+ is the pseudoinverse of A. Hence, observe that $\|A\| \leq 1$ is equivalent to A having maximum singular value 1[4]. Similarly, A^+ has inverted singular values from A and so $\|A^+\|$ is equal to the reciprocal of the minimum nonzero singular value. Therefore, the condition number of A is given by $\|A\|\|A^+\| \leq \kappa$.

[4]To see this, simply consider Spectral Theorem applied to Hermitian matrix $A^\dagger A$

So, define x to be the least-squares solution to the linear system $Ax = b$ i.e. $x = A^+b$. Then, in terms of these definitions, we define two primary goals:

1. Query a vector $\tilde{x}$ s.t. $\|\tilde{x} - x\| \leq \epsilon\|x\|$
2. Sample from a distribution that approximates $\frac{|x_j|^2}{\|x\|^2}$ within total variation distance (Theorem 14.9) 2ϵ.

In order to do this, we simply assume that we have length-square sampling access to A. In other words, we are able to sample row indices of A from the distribution $\frac{\|A_{(i,\cdot)}\|^2}{\|A\|_F^2}$

13.0.7 Sequence of Approximations

First, we'll summarize the sequence of approximations that we'll perform using length-squared sampling techniques. We'll describe these steps in depth in the following sections.

Of course, we know that the least squares solution of the linear system is given by the orthogonal projection

$$(A^\dagger A)^+ A^\dagger = A^+ b$$

So, we first approximate $A^\dagger A$ by $R^\dagger R$ where $R \in \mathbb{C}^{r\times n}$, $r \ll m$ is constructed from length-square sampling r rows of A. Now, denote the spectral decomposition

$$A^\dagger A \approx R^\dagger R = \sum_{l=1}^{k} \frac{1}{\sigma_l^2} \left|v^{(l)}\right\rangle \left\langle v^{(l)}\right|$$

where of course σ_i and $\left|v^{(i)}\right\rangle \in \mathbb{C}^n$ are the singular values and right singular vectors of R, respectively.

We see that computing these right singular vectors of R can still be computationally prohibitive given the dimension n. Hence, we can use length-square sampling again, this time on the columns of R to give a matrix $C \in \mathbb{C}^{r\times c}$, $c \ll n$. Now, the left singular vectors of C which we denote as $\left|w^{(i)}\right\rangle \in \mathbb{C}^r$ can be efficiently computed via standard SVD methods. So,

$$RR^\dagger \approx CC^\dagger = \sum_{l=1}^{k} \frac{1}{\sigma_l^2} \left|w^{(l)}\right\rangle \left\langle w^{(l)}\right|$$

We can then show that ()

$$\left|\tilde{v}^{(i)}\right\rangle := R^\dagger \left|w^{(l)}\right\rangle / \tilde{\sigma}_l \tag{3}$$

provides a good approximation of $\left|v^{(i)}\right\rangle$. Note that $\tilde{\sigma}_l$ are the singular values of C which then approximate the singular values of R which similarly approximate the singular values of A. This follows from $A^\dagger A \approx R^\dagger R$ and $RR^\dagger \approx CC^\dagger$ by the Hoffman–Wielandt inequality detailed in Lemma 2.7 of [21] and stated without proof below.

Lemma 13.11. *Hoffman–Wielandt inequality*

If P, Q are two real, symmetric $n \times n$ matrices and $\lambda_1, \cdots \lambda_n$ denote eigenvalues in non-decreasing order, then

$$\sum_{t=1}^{n} (\lambda_t(P) - \lambda_t(Q))^2 \leq \|P - Q\|_F^2$$

At this point, it seems like we haven't made much progress since computing $R^\dagger \left|w^{(l)}\right\rangle$ is still expensive. However, it turns out that all we need to enable query access to $\tilde{x}$ is the ability to efficiently estimate the trace inner product $\operatorname{tr}\left(U^\dagger V\right)$ where U and V are operators such that U can be the length-square sampled and V can be queried. To see this, we write our solution, $\tilde{x}$, in terms of the approximations thus far

$$\begin{aligned}
\tilde{x} &\approx A^+ \left|b\right\rangle \\
&\approx (R^\dagger R)^+ A^\dagger \left|b\right\rangle \\
&\approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|\tilde{v}^{(l)}\right\rangle \left\langle \tilde{v}^{(l)}\right| A^\dagger \left|b\right\rangle
\end{aligned}$$

Hence, define $U := A$, $V := \left|b\right\rangle \left\langle \tilde{v}^{(l)}\right|$ in which case

$$\begin{aligned}
\operatorname{tr}\left(U^\dagger V\right) &= \operatorname{tr}\left(A^\dagger \left|b\right\rangle \left\langle \tilde{v}^{(l)}\right|\right) \\
&= \operatorname{tr}\left(\left\langle \tilde{v}^{(l)}\right| A^\dagger \left|b\right\rangle\right) \\
&= \left\langle \tilde{v}^{(l)}\right| A^\dagger \left|b\right\rangle
\end{aligned}$$

since $\left\langle \tilde{v}^{(l)}\right| A^\dagger \left|b\right\rangle$ is a scalar. Therefore, say that

$$\tilde{\lambda}_l \approx \operatorname{tr}\left(A^\dagger \left|b\right\rangle \left\langle \tilde{v}^{(l)}\right|\right)$$

and assume that we can compute and memoize these scalars $\tilde{\lambda}_i$ efficiently. In which case,

$$\tilde{x} \approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|\tilde{v}^{(l)}\right\rangle \tilde{\lambda}_l$$

Recalling the definition of $\left|\tilde{v}^{(i)}\right\rangle$ (3),

$$\begin{aligned}
&= \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} R^\dagger \left|w^{(l)}\right\rangle \tilde{\lambda}_l \\
&= R^\dagger \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} \left|w^{(l)}\right\rangle \tilde{\lambda}_l
\end{aligned}$$

and so defining $z := \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} \left|w^{(l)}\right\rangle \tilde{\lambda}_l$,

$$= R^\dagger z$$

We see that we can compute z efficiently (and memoize it for future queries) because it is a k-linear combination of left singular vectors in $\mathbb{C}^r$. So, say that we wish to query an element $\tilde{x}_j$. We can simply query column $R_{\cdot,j} \in \mathbb{C}^r$ (or equivalently row $R^\dagger_{j,\cdot}$) and compute $R_{\cdot,j} \cdot z$. Hence, we've achieved our first goal.

In order to achieve our second goal, enabling sample access to a distribution that approximates $\frac{|x_j|^2}{\|x\|^2}$, we require one more trick: rejection sampling which we detail in Section ().

All in all, we've performed the chain of approximations,

$$\begin{aligned}
|x\rangle = A^+ |b\rangle &= (A^\dagger A)^+ A^\dagger |b\rangle \\
&\approx (R^\dagger R)^+ A^\dagger |b\rangle = \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|v^{(l)}\right\rangle \left\langle v^{(l)}\right| A^\dagger |b\rangle \\
&\approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|\tilde{v}^{(l)}\right\rangle \left\langle \tilde{v}^{(l)}\right| A^\dagger |b\rangle \\
&\approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|\tilde{v}^{(l)}\right\rangle \tilde{\lambda}_l = R^\dagger \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} \left|w^{(l)}\right\rangle \tilde{\lambda}_l = R^\dagger z
\end{aligned}$$

Now that we've sketched the steps of this process, we detail each approximation and show that we can achieve the claimed correctness and complexity bounds.

13.0.8 Computing Approximate Singular Vectors

As described above, we begin by length-square sampling the original matrix $A \in \mathbb{C}^{m \times n}$. Suppose we want to draw s rows in s i.i.d. trials. Then, pick row index i of A with probability

$$p_i = \frac{\|A_{(i,\cdot)}\|^2}{\|A\|_F^2} \tag{4}$$

and output random row

$$\begin{aligned}
Y &= \frac{1}{\sqrt{sp_i}} \left\langle A_{(i,\cdot)}\right| \\
&= \frac{1}{\sqrt{s}} \frac{\|A\|_F}{\|A_{(i,\cdot)}\|} \left\langle A_{(i,\cdot)}\right|
\end{aligned}$$

which is just a scaling of the ith row of A[5]. In other words,

[5]The reason that we scale by s is so that the expectations of $A^\dagger A$ and $R^\dagger R$ coincide in the theorem that follows. The reason that we scale by p_i is so that the norms of all rows are equivalent—a fact which we'll utilize when we sample R column-wise.

$$\Pr\left(Y = \frac{1}{\sqrt{sp_i}} \langle A_{(i,\cdot)}|\right) = p_i$$

After sampling s rows, we implicitly define matrix R to be the concatenation of the outputted random rows. Therefore,

$$R = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} \in \mathbb{C}^{s \times n} \tag{5}$$

Note that $\langle Y_k|$ denotes the random row outputted by the procedure on the kth i.i.d. draw.

Lemma 13.12. *Let $X = Y^\dagger Y - E[Y^\dagger Y]$ which evidently satisfies $E[X] = 0$. Then,*

$$E[X^2] \preceq E[(Y^\dagger Y)^2] \tag{6}$$

$$= A^\dagger A \|A\|_F^2 \frac{1}{s^2} \tag{7}$$

and so

$$\|E[X^2]\|_2 \leq \frac{1}{s^2} \|A\|_2^2 \|A\|_F^2 \tag{8}$$

Furthermore,

$$\|X\|_2 = \frac{1}{s} \|A\|_F^2 \tag{9}$$

Proof. First, observe that $E[X^2]$ is the element-wise variance of $Y^\dagger Y$ and $E[(Y^\dagger Y)^2]$ is the corresponding second moment. Hence, the relation $E[X^2] \leq E[(Y^\dagger Y)^2]$ holds element-wise which implies that the matrix relation $E[X^2] \preceq E[(Y^\dagger Y)^2]$ holds as well.

Furthermore,

$$\begin{aligned}
E[(Y^\dagger Y)^2] &= \frac{1}{s^2} \sum_{i=1}^{m} \frac{p_i}{p_i^2} \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| \\
&= \frac{1}{s^2} \sum_{i=1}^{m} \frac{\|A\|_F^2}{\left\langle A_{(i,\cdot)}|A_{(i,\cdot)}\right\rangle} \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| \qquad \text{(using (4))} \\
&= \frac{1}{s^2} \sum_{i=1}^{m} \|A\|_F^2 \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| = \frac{\|A\|_F^2}{s^2} \sum_{i=1}^{m} \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| \\
&= A^\dagger A \|A\|_F^2 \frac{1}{s^2}
\end{aligned}$$

Recall that $E[R^\dagger R] = \frac{1}{s}A^\dagger A$. So, observe that

$$\begin{aligned}
\|X\|_2 &= \|Y^\dagger Y - E[Y^\dagger Y]\|_2 \\
&= \frac{1}{s}\left\| \frac{1}{p_i} \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| - A^\dagger A \right\|_2 \\
&\leq \frac{1}{s} \max\left\{ \left\| \frac{1}{p_i} \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| \right\|_2, \left\| A^\dagger A \right\|_2 \right\} \\
&\leq \frac{1}{s} \max\left\{ \frac{\|A\|_F^2}{\| \left|A_{(i,\cdot)}\right\rangle \|_2^2} \left\| \left|A_{(i,\cdot)}\right\rangle \left\langle A_{(i,\cdot)}\right| \right\|_2, \|A^\dagger A\|_F^2 \right\}
\end{aligned}$$

using $\|A^\dagger A\|_2 = \|A\|_2^2 \leq \|A\|_F^2$ and plugging in (4).

$$= \frac{1}{s}\|A\|_F^2$$

□

Proposition 13.13. *If $t > 0, t \in \mathbb{C}$ satisfies $\|tX\|_2 \leq 1$ for all possible values of X, then*

$$\begin{aligned}
\|E[e^{\pm tX}]\|_2 &\leq 1 + \frac{t^2}{s^2}\|A\|_2^2\|A\|_F^2 \\
&\leq e^{t^2\|A\|_2^2\|A\|_F^2/s^2}
\end{aligned} \tag{10}$$

Proof. First, from (14.13) we know that $E[e^{tX}] \preceq E[I + X + X^2] = I + E[X^2]$ since $E[X] = 0$. Hence, we have the proposition by (8). □

Theorem 13.14. *Let $A \in \mathbb{C}^{m\times n}$ and $R \in \mathbb{C}^{r\times n}$ be constructed by the length-square sampling and scaling so that $E[R^\dagger R] = E[A^\dagger A]$ (requirements that are met by R defined in (5)). Then, for all $\epsilon \in [0, \|A\|/\|A\|_F]$[6], we have*

$$\Pr\left(\|R^\dagger R - A^\dagger A\| \geq \|A\|\|A\|_F\right) \leq 2ne^{\frac{-\epsilon^2 s}{4}}$$

Hence, for $s \geq (4\ln\frac{2n}{\eta})/\epsilon^2$, with probability at least $(1-\eta)$ we have

$$\|R^\dagger R - A^\dagger A\| \leq \epsilon\|A\|\|A\|_F$$

Proof. From our definition above, we have that

[6] If $\epsilon \geq \|A\|/\|A\|_F$, then we can simply use $\hat{0}$ to approximate $A^\dagger A$

$$R^\dagger R = \begin{bmatrix} Y_1^\dagger & Y_2^\dagger & \dots & Y_s^\dagger \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}$$
$$= \sum_{k=1}^{s} |Y_k\rangle \langle Y_k|$$

Let i_k give the index of the row sampled from A on the kth draw. Hence,

$$= \frac{1}{s} \sum_{k=1}^{s} \frac{1}{p_{i_k}} \left|A_{(i_k,\cdot)}\right\rangle \left\langle A_{(i_k,\cdot)}\right|$$

Furthermore,

$$E[R^\dagger R] = \frac{1}{s} \sum_{k=1}^{s} \sum_{i_k=1}^{m} \frac{p_{i_k}}{p_{i_k}} \left|A_{(i_k,\cdot)}\right\rangle \left\langle A_{(i_k,\cdot)}\right|$$
$$= \frac{1}{s} \sum_{k=1}^{s} A^\dagger A$$
$$= A^\dagger A$$

Note that similarly $E[Y^\dagger Y] = \frac{1}{s} A^\dagger A$.

So, we can define $X_i = |Y_k\rangle \langle Y_k| - E\Big[\, |Y_k\rangle \langle Y_k| \,\Big]$ which is evidently an i.i.d. copy of X as we've defined previously. Hence,

$$\sum_{i=1}^{s} X_i = R^\dagger R - E[R^\dagger R]$$
$$= R^\dagger R - A^\dagger A$$

Now, we can first apply Theorem 11 with $a = \epsilon \|A\|_2 \|A\|_F$,

$$\Pr\left(\left\| \Big(\sum_{i=1}^{s} X_i \Big) \right\|_2 \geq \epsilon \|A\|_2 \|A\|_F \right)$$
$$\leq n e^{-t\epsilon \|A\|_2 \|A\|_F} (\|E[e^{tX}]\|_2^s + \|E[e^{-tX}]\|_2^s)$$

for any $t > 0$. Hence, we can apply Proposition 13.13 which then gives us

$$\leq 2n e^{-t\epsilon \|A\|_2 \|A\|_F} e^{t^2 \|A\|_2^2 \|A\|_F^2 / s^2}$$

for $t \leq s/\|A\|_F^2$. Hence, we can set $t = \frac{\epsilon s}{2\|A\|_F\|A\|_2}$ (which is indeed less than $s/\|A\|_F^2$) and finally,

$$\leq 2ne^{-\epsilon^2 s/4}$$

Therefore, if we require that $s \geq (4 \ln \frac{2n}{\eta})/\epsilon^2$ we then have

$$\Pr\left(\left\|\left(\sum_{i=1}^{s} X_i\right)\right\|_2 \geq \epsilon\|A\|_2\|A\|_F\right) \leq 2ne^{-\frac{\epsilon^2}{4}\frac{4\ln\frac{2n}{\eta}}{\epsilon^2}}$$
$$= 2ne^{-\ln\frac{2n}{\eta}} = \eta$$

□

13.1 Conclusions

- Claim (Tang): For machine learning problems, $\mathcal{SQ}$ assumptions are more reasonable than state preparation assumptions.
- We discussed pseudo-inverse which inverts singular values, but in principle we could have applied any function to the singular values
- Gilyen et. al (2018) show that many quantum machine learning algorithms indeed apply polynomial functions to singular values
- Our discussion suggests that exponential quantum speedups are tightly related to problems where high-rank matrices play a crucial role (e.g. Hamiltonian simulation or QFT)

14 Optimal Quantum Sample Complexity

This paper[5] provides an instructive example of how one can use quantum information theory to discuss the ability of a quantum learning algorithm to learn from a distribution of quantum states. Perhaps, this is unsurprising given the surface-level connection with quantum state discrimination which has been closely studied in quantum information theory.

14.1 Definitions

14.1.1 Quantum Learning Models: PAC Setting

A quantum example oracle $QPEX(c, D)$ acts on $|0\rangle^{\otimes n}|0\rangle$ and produces a quantum example $\sum_{x\in\{0,1\}^n} D(x)|x, c(x)\rangle$.

A quantum learner is given access to some copies of the state generated by $QPEX(c, D)$ and performs a POVM where each outcome is associated with a hypothesis.

A learning algorithm A is an (ϵ, δ)-PAC quantum learner for C if for every $c \in C$ and distribution D, given access to the $QPEX(c, D)$ oracle, A outputs an h such that $err_D(h, c) \leq \epsilon$, with probability at least $1 - \delta$.

The sample complexity of A is the maximum number invocations of the $QPEX(c, D)$ oracle, maximized over all $c \in C$, distributions D, and the learners internal randomness. The (ϵ, δ)-PAC quantum sample complexity of a concept class C is the minimum sample complexity over all (ϵ, δ)-PAC quantum learners for C.

14.1.2 Quantum Learning Models: Agnostic Setting

For a joint distribution $D : \{0,1\}^{n+1} \to [0,1]$ over the set of examples, the learner has access to an $QAEX(D)$ oracle which acts on $|0\rangle^{\otimes n} |0\rangle$ and produces a quantum example $\sum_{(x,b)\in\{0,1\}^{n+1}} D(x,b) |x,b\rangle$.

A learning algorithm A is an (ϵ, δ)-agnostic quantum learner for C if for every distribution D, given access to the $QAEX(D)$ oracle, A outputs an $h \in C$ such that $err_D(h) \leq opt_D(h) + \epsilon$ with probability at least $1 - \delta$.

The sample complexity of A is the maximum number invocations of the $QAEX(D)$ oracle over all distributions D and over the learners internal randomness. The (ϵ, δ)-agnostic quantum sample complexity of a concept class C is the minimum sample complexity over all (ϵ, δ)-agnostic quantum learners for C.

14.2 Goals

We seek to show that quantum examples are not actually more powerful than classical labeled examples in the PAC model and in the agnostic model when the underlying data distribution is arbitrary. We emphasize this point on the data distribution because we plan to later detail advantages that can be reaped if the example distribution is e.g. uniform (hint: think about the Bernstein-Vazirani algorithm and perceptron).

In the classical case, the sample complexity of concept class $\mathcal{C}$ with VC dimension d in the PAC setting is

$$\Theta(\frac{d}{\epsilon} + \frac{\log(1/\delta)}{\epsilon})$$

where ϵ is the approximation coefficient and δ is the probability of success, as usual.

In the agnostic case, the optimal sample complexity of such agnostic learners is tightly determined by the VC dimension of $\mathcal{C}$

$$\Theta(\frac{d}{\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2})$$

The authors indeed show that, using the quantum learning models above, the bounds are the same in the quantum case. This requires a state identification argument which uses Fourier Analysis to analyze the performance of a Pretty Good Measurement14.A. However, we can get close by instead using simple concepts from quantum information theory. In my view, this approach has a strong aesthetic and is likely transferrable to similar problems.

Of course, we know that the upper bounds in sample complexity are the same as the classical case since we can always implement a classical algorithm on a quantum computer. Hence, we seek lower bounds instead.

14.3 Information Theoretic Lower Bounds on Sample Complexity

14.3.1 VC-independent lower bounds

Lemma 14.1. *Let $\mathcal{C}$ be a non-trivial concept class. For every $\delta \in (0, 1/2), \epsilon \in (0, 1/4)$, a (ϵ, δ)-PAC quantum learner for $\mathcal{C}$ has sample complexity $\Omega(\frac{1}{\epsilon} \log \frac{1}{\delta})$*

Proof. Since $\mathcal{C}$ is non-trivial, we may assume there are two concepts $c_1, c_2 \in C$ defined on two inputs $\{x_1, x_2\}$ as follows $c_1(x_1) = c_2(x_1) = 0$ and $c_1(x_2) = 0, c_2(x_2) = 1$.

Consider the distribution $D(x_1) = 1\epsilon$ and $D(x_2) = \epsilon$. For $i \in \{1, 2\}$, the state of the algorithm after T queries to $QPEX(c_i, D)$ is

$$|\psi_i\rangle = (\sqrt{1-\epsilon}\,|x_1, 0\rangle + \sqrt{\epsilon}\,|x_2, c_i(x_2)\rangle)^{\otimes T}$$

Therefore, $\langle\psi_1|\psi_2\rangle = (1-\epsilon)^T$. Since the success probability of an (ϵ, δ)-PAC quantum learner is $\geq 1 - \delta$, Corollary 14.7.1 implies $|\psi_1\rangle\,|\psi_2\rangle \leq 2\sqrt{\delta(1-\delta)}$.

$\therefore T = \Omega(\frac{1}{\epsilon} \log \frac{1}{\delta})$ □

Lemma 14.2. *Let $\mathcal{C}$ be a non-trivial concept class. For every $\delta \in (0, 1/2), \epsilon \in (0, 1/4)$, a (ϵ, δ)-agnostic quantum learner for $\mathcal{C}$ has sample complexity $\Omega(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$*

Proof. Since $\mathcal{C}$ is non-trivial, we may assume there are two concepts $c_1, c_2 \in C$ defined on two inputs $\{x_1, x_2\}$ such that $c_1(x) \neq c_2(x)$.

Consider the two distributions $D_\pm$ s.t.

$$D_\pm(x, c_1(x)) = (1 \pm \epsilon)/2$$
$$D_\pm(x, c_2(x)) = (1 \mp \epsilon)/2$$

$D_+(x_1) = 1\epsilon$ and $D(x_2) = \epsilon$. The state of the algorithm after T queries to $QAEX(D_\pm)$ is

$$|\psi_\pm\rangle = ((\sqrt{(1 \pm \epsilon)/2})\,|x, c_1(x)\rangle + (\sqrt{(1 \mp \epsilon)/2})\,|x, c_2(x)\rangle)^{\otimes T}$$

Therefore, $\langle\psi_+|\psi_-\rangle = (1 - \epsilon^2)^{T/2}$. Since the success probability of an (ϵ, δ)-agnostic quantum learner is $\geq 1 - \delta$, Corollary 14.7.1 implies $\langle\psi_+|\psi_-\rangle \leq 2\sqrt{\delta(1-\delta)}$.

$\therefore T = \Omega(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ □

14.3.2 Classical PAC Learning

Theorem 14.3. *Let C be a concept class with $VC - dim(\mathcal{C}) = d + 1$. Then for every $\delta \in (0, 1/2)$ and $\epsilon \in (0, 1/4)$, every (ϵ, δ)-PAC learner for $\mathcal{C}$ has sample complexity $\Omega(\frac{d}{\epsilon} + \frac{\log(1/\delta)}{\epsilon})$.*

Proof. Consider an (ϵ, δ)-PAC learner for $\mathcal{C}$ that uses T examples. The d-independent part of the lower bound, T = (log(1/)/), even holds for quantum examples and was proven in Lemma 10.

Hence it remains to prove T = (d/). It suffices to show this for a specific distribution D, defined as follows. Let S = s0,s1,...,sd 0,1n be some (d + 1)-element set shattered by C . Define D(s0) = 14 and D(si) = 4/d for all i [d]. Because S is shattered by C , for each string a 0, 1d , there exists a concept ca C such that ca(s0) = 0 and ca(si) = ai for all i [d]. We define two correlated random variables A and B corresponding to the concept and to the examples, respectively. Let A be a random variable that is uniformly distributed over 0, 1d ; if A = a, let B = B1 . . . BT be T i.i.d. examples from ca according to D. We give the following three-step analysis of these random variables: 1. I(A:B)(1)(1H(1/4))dH()=(d). Proof. Let random variable h(B) 0, 1d be the hypothesis that the learner produces (given the examples in B) restricted to the elements s1,...,sd. Note that the error of the hypothesis □

14.A Pretty Good Measurement (PGM)

Given a density matrix ensemble $\mathcal{E} = \{p_i, \sigma_i\}$ and a quantum state ρ we are promised that ρ is in state σ_i with probability p_i. In the general case we have $i \in [m]$ and of course $\sum_{i=1}^{m} p_i = 1$. Our goal is then to successfully identify which of the σ_i that our state ρ is actually in. This is known as Quantum Hypothesis Testing.

In some sense, thinking back to Holevo's Theorem (Theorem 6.1), this is related to Bob attempting to access information transported from Alice that is given from the distribution above.

Hence, we perform a maximization with respect to both the probabilities on each state as well as with respect to any randomness that our approach employs. We then must choose a Quantum POVM $\{E_i\}$ that carries put a measurement and maximizes our probability of getting the state right.

So say we pick a POVM. Hence, we know that

$$\Pr(Success) = \sum_i p_i \operatorname{Tr}(\sigma_i E_i)$$

So this is the quantity that we seek to maximize.

We've shown above that the trace distance provides the solution for $m = 2$. As it turns out for $m > 2$ this is not an easy problem. However, PGM provides a sound approximation:

Intuitively, it might seem reasonable to simply choose

$$E_i = p_i \sigma_i$$

Unfortunately, then, $\sum_i E_i \neq I$. Well, one case we may think about is if we can guarantee σ_i is a the sum of pure states from an orthonormal basis. In which case, let $S = \sum_i \sigma_i$ and we choose

$$E_i = S^{1/2} \sigma_i S^{1/2}$$

Then we have

$$\operatorname{tr}(E_i \sigma_i) = \operatorname{tr}(\sigma_i) = 1$$

by orthonormality. Inspired by this, define the PGM POVM to be

$$E_i = S^{-1/2} p_i \sigma_i S^{-1/2}$$

for our original problem. Positive semidefiniteness is clear, so it remains to show that we have completeness

$$\begin{aligned}\sum_i E_i &= \sum_i S^{-1/2} p_i \sigma_i S^{-1/2} \\ &= S^{-1/2} \sum_i p_i \sigma_i S^{-1/2} \\ &= S^{-1/2} S S^{-1/2} = I\end{aligned}$$

Theorem 14.4. *Let $\Pr_{opt}(\mathcal{E})$ by the optimal success probability for our quantum hypothesis testing problem. Define $\Pr_{PGM}(\mathcal{E})$ to be the average success probability using the PGM POVM. Then,*

$$Pr_{opt}(\mathcal{E})^2 \leq Pr_{PGM}(\mathcal{E}) \leq Pr_{opt}(\mathcal{E})$$

14.B Trace Distance

Definition 14.5. *Trace Distance*

The trace distance $T(\cdot, \cdot)$ is a metric on the space of density operators and gives a measure of distinguishability between states. In particular, let ρ, σ be density operators,

$$\begin{aligned}T(\rho, \sigma) &= \frac{1}{2} \operatorname{Tr}\left[\sqrt{(\rho - \sigma)^2}\right] \\ &= \frac{1}{2} \sum_i |\lambda_i|\end{aligned}$$

where λ_i are the eigenvalues of Hermitian $\rho - \sigma$.

Hence, it is simply the trace norm of the positivization of the difference of matrices.

Lemma 14.6. *For any states ρ, σ one may write $\rho - \sigma = Q - S$ where Q and S are positive operators with support on orthogonal vector spaces (Exercise 9.7 [29])*

Proof. ρ, σ are p.s.d operators. Hence, $\rho - \sigma$ is Hermitian, so we can write $\rho - \sigma = \sum_i \lambda_i |u_i\rangle \langle u_i|$ where $\{u_i\}$ is an orthonormal basis of the Hilbert space, by spectral theorem. Now, we can decompose the eigenbasis into positive and negative components. Then, $\rho - \sigma = \sum_i \lambda_i^+ |u_i\rangle \langle u_i| + \sum_j \lambda_j^- |u_j\rangle \langle u_j|$ where $\lambda_i^+ > 0, \lambda_j^- < 0$. Since, each component partitions the vector space (other than at the additive identity) by the orthogonality condition, this is a direct sum. □

Lemma 14.7. *The maximum probability of distinguishing between two states with an optimal measurement is given by*

$$1/2[1 + T(\rho_1, \rho_2)]$$

Proof. Say that we have the ensemble $\{(p_1, \rho_1), (p_2, \rho_2)\}$. We seek to define a POVM $\{E_1, E_2\}$ where E_1 indicates ρ_1 and similarly for ρ_2. hence the probability of success is given by

$$\begin{aligned}
\Pr_{\max} &= \max_{E_1, E_2} p_1 \operatorname{Tr}[E_1\rho_1] + p_2 \operatorname{Tr}[E_2\rho_2] \\
&= \max_{E_1} p_1 \operatorname{Tr}[E_1\rho_1] + p_2 \operatorname{Tr}[(I - E_1)\rho_2] && \text{(completeness of POVM)} \\
&= \max_{E_1} p_2 \operatorname{Tr}[\rho_2] + \operatorname{Tr}[E_1(p_1\rho_1 - p_2\rho_2)] && \text{(linearity of trace)} \\
&= \max_{E_1} \Big(p_2 + \operatorname{Tr}[E_1(p_1\rho_1 - p_2\rho_2)]\Big) && (\operatorname{Tr}(\rho) = 1 \text{ for any density operator})
\end{aligned}$$

Therefore, the optimal projection E_1 is onto the positive eigenspace of $(p_1\rho_1 - p_2\rho_2)$. In which case, recalling that ρ_i are p.s.d, $E_1(p_1\rho_1 - p_2\rho_2)] = p_1\rho_1$ and so

$$= \Big(p_2 + \operatorname{Tr}[p_1\rho_1]\Big)$$

We did implicitly assume that $E_1 + E_2 = I$ where we could've had an additional indeterminate E_3. However, the proof would still follow in any case ([29]). □

Corollary 14.7.1. *Consider attempting to distinguish two pure states $|\psi_0\rangle, |\psi_1\rangle$. Then, we will distinguish correctly with probability at most $1/2[1 + \sqrt{1 - |\langle\psi_0|\psi_1\rangle|^2}]$.*

Equivalently, if we can distinguish between the two states w.p. $1-\delta$, then $|\langle\psi_0|\psi_1\rangle| \leq 2\sqrt{\delta(1-\delta)}$.

Proof. Applying the previous Lemma, we have maximum probability

$$1/2[1 + T(|\psi_0\rangle\langle\psi_0|, |\psi_1\rangle\langle\psi_1|)]$$

So, write $|\psi_1\rangle = \cos(\theta)|\psi_0\rangle + e^{i\varphi}\sin(\theta)|\psi_0^\perp\rangle$. Hence,

$$|\psi_1\rangle\langle\psi_1| = \cos^2(\theta)|\psi_0\rangle\langle\psi_0| + \sin^2(\theta)|\psi_0^\perp\rangle\langle\psi_0^\perp|$$
$$+e^{i\varphi}\cos(\theta)\sin(\theta)|\psi_0^\perp\rangle\langle\psi_0| + e^{-i\varphi}\cos(\theta)\sin(\theta)|\psi_0\rangle\langle\psi_0^\perp|$$

Since trace is basis-independent, we can write $|\psi_0\rangle\langle\psi_0| - |\psi_1\rangle\langle\psi_1|$ in the above used orthogonal basis. This gives us characteristic polynomial

$$\begin{aligned}
0 &= (1 - \cos^2(\theta) - \lambda)(-\sin^2(\theta) - \lambda) - \cos^2(\theta)\sin^2(\theta) \\
&= (\sin^2(\theta) - \lambda)(-\sin^2(\theta) - \lambda) - \cos^2(\theta)\sin^2(\theta) \\
&= -\sin^4(\theta) + \lambda^2 - \cos^2(\theta)\sin^2(\theta) \\
&= -\sin^2(\theta) + \lambda^2
\end{aligned}$$

So, $\lambda = \pm|\sin(\theta)|$. Therefore, since the trace distance is the absolute sum of the eigenvalues of this difference,

$$T(|\psi_0\rangle\langle\psi_0|, |\psi_1\rangle\langle\psi_1|) = 2|\sin(\theta)|sam$$

and indeed

$$|\langle\psi_0|\psi_1\rangle|^2 = |\cos(\theta)|^2 \Rightarrow \sqrt{1 - |\langle\psi_0|\psi_1\rangle|^2} \qquad = |\sin(\theta)|$$

as desired. □

Lemma 14.8. *Let A, B, C by symmetric $d \times d$ matrices satisfying $A \succeq 0$ and $B \preceq C$. Hence,* $\mathrm{Tr}(AB) \leq \mathrm{Tr}(AC)$

Proof. Write A in its spectral decomposition $A = \sum \lambda_i |i\rangle\langle i|$, invoking Spectral Theorem (A.5). Hence,

$$\begin{aligned}
\mathrm{Tr}(AB) &= \mathrm{Tr}\Big(\sum \lambda_i |i\rangle\langle i| B\Big) \\
&= \sum \lambda_i \mathrm{Tr}(|i\rangle\langle i| B) && \text{(linearity of trace)} \\
&= \sum \lambda_i \mathrm{Tr}(\langle i| B |i\rangle) && \text{(cyclic property of trace)} \\
&\leq \sum \lambda_i \mathrm{Tr}(\langle i| C |i\rangle) \\
&= \sum \lambda_i \mathrm{Tr}(|i\rangle\langle i| C) = \mathrm{Tr}\Big(\sum \lambda_i |i\rangle\langle i| C\Big) = \mathrm{Tr}(AC)
\end{aligned}$$

□

Corollary 14.8.1. *If $A, B \succeq 0$, then* $\mathrm{Tr}(AB) \leq \|B\|_2 \mathrm{Tr}(A)$

Proof. Note that the singular values of B coincide with the eigenvalues of B since $B^\dagger B = B^2$ and $B \succeq 0 \Rightarrow \lambda_i(B) \geq 0, \forall i$. So, let $C = \|B\|_2 I$ which then trivially satisfies $\lambda_i(C) = \lambda_{\max}(B), \forall i$ since C is the diagonal matrix with diagonal values all equal to $\lambda_{\max}(B)$. Therefore, $B \preceq C$. So, we can simply apply 14.8 above,

$$\begin{aligned}
\mathrm{Tr}(AB) &\leq \mathrm{Tr}(AC) \\
&= \mathrm{Tr}\left(A\|B\|_2 I\right) \\
&= \|B\|_2 \mathrm{Tr}(A)
\end{aligned}$$

□

Definition 14.9. *Total Variation Distance.*

Let P and Q be distinct probability measures on a σ-algebra $\mathcal{F}$ of subsets of the sample space Ω. Then, the total variation distance is given by

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

Lemma 14.10. *Hoeffding–Chernoff Inequality*

Let $X_1, X_2, \cdots, X_s$ be i.i.d real random variables. For any positive, real numbers a, t we have that, from Markov's inequality,

$$\begin{aligned}\Pr\left(\sum_{i=1}^{s} X_i \geq a\right) &\leq e^{-ta} E\left[\prod_{i=1}^{s} e^{tX_i}\right] \\ &= e^{-ta} \prod_{i=1}^{s} E\left[e^{tX_i}\right]\end{aligned}$$

by independence. □

Theorem 14.11. *Hoeffding–Chernoff Inequality for matrix-valued random variables [21]*

Let X be a random variable taking values which are real symmetric $d \times d$ matrices. Suppose $X_1, X_2, \cdots, X_s$ are i.i.d. draws of X. For any positive real numbers a, t, we have

$$\Pr\left(\lambda_{\max}\Big(\sum_{i=1}^{s} X_i\Big) \geq a\right) \leq d e^{-ta} \|E[e^{tX}]\|_2^s \tag{11}$$

$$\Pr\left(\Big\|\Big(\sum_{i=1}^{s} X_i\Big)\Big\|_2 \geq a\right) \leq d e^{-ta} (\|E[e^{tX}]\|_2^s + \|E[e^{-tX}]\|_2^s) \tag{12}$$

where $\lambda_{\max}$ is the largest eigenvalue.

Proof. First, we can show that (11) ⇒ (12). By definition of the 2-norm of a matrix,

$$\|\sum_i X_i\|_2 = \max\left(\lambda_{\max}\Big(\sum_i X_i\Big), \lambda_{\max}\Big(\sum_i (-X_i)\Big)\right)$$

since it is the square root of the maximum eigenvalue of $(\sum_i X_i^T)\sum_i X_i = (\sum_i X_i)\sum_i X_i$ and hence, equivalently, the maximum absolute value of an eigenvalue of X_i. Therefore, we can simply apply (11) to both X_i and $-X_i$ and we get (12).

So, we can focus our attention on (12). Let $S = \sum_i^s X_i$. Hence,

$$\lambda_{\max}(S) \geq a \Leftrightarrow \lambda_{\max}(tS) \geq ta$$

Furthermore, by considering the power series definition of the exponential,

$$\begin{aligned}&\Leftrightarrow \lambda_{\max}(e^{tS}) \geq e^{ta} \\ &\Rightarrow \mathrm{Tr}\big(e^{tS}\big) \geq e^{ta}\end{aligned}$$

since the trace is the sum of the matrix's eigenvalues. Since $\mathrm{Tr}\big(e^{tS}\big) \geq 0$, we can apply Markov's inequality

$$\Pr\big(\mathrm{Tr}\big(e^{tS}\big) \geq e^{ta}\big) \leq \frac{E[\mathrm{Tr}\big(e^{tS}\big)]}{e^{ta}}$$

Now, we use the following lemma

Lemma 14.12. *Golden-Thompson Inequality*
If A and B are Hermitian matrices, then

$$\mathrm{Tr}\big(e^{A+B}\big) \leq \mathrm{Tr}\big(e^{A}e^{B}\big)$$

□

Hence, we can let $A = t(\sum_i^{s-1} X_i)$ and $B = tX_s$. Then,

$$E_X\Big[\mathrm{Tr}\big(e^{tS}\big)\Big] \leq E_X\Big[\mathrm{Tr}\Big(e^{t\left(\sum_i^{s-1} X_i\right)}e^{tX_s}\Big)\Big]$$

Since the expectation operator commutes with the summation of the trace by linearity of trace,

$$\begin{aligned}
&= \mathrm{Tr}\left(E_X\left[e^{t\left(\sum_i^{s-1} X_i\right)}e^{tX_s}\right]\right) \\
&= \mathrm{Tr}\left(E_{X_1,X_2,\cdots,X_{s-1}}\left[e^{t\left(\sum_i^{s-1} X_i\right)}\right]E_{X_s}\left[e^{tX_s}\right]\right) \qquad \text{(by independence)}
\end{aligned}$$

Now, we can apply Corollary (14.8.1), which gives

$$\begin{aligned}
&\leq \mathrm{Tr}\left(E_{X_1,X_2,\cdots,X_{s-1}}\left[e^{t\left(\sum_i^{s-1} X_i\right)}\right]\right)\left\|E_{X_s}\left[e^{tX_s}\right]\right\|_2 \\
&= \mathrm{Tr}\left(E_X\left[e^{t\left(\sum_i^{s-1} X_i\right)}\right]\right)\left\|E_X\left[e^{tX}\right]\right\|_2 \\
&= E_X\left[\mathrm{Tr}\left(e^{t\left(\sum_i^{s-1} X_i\right)}\right)\right]\left\|E_X\left[e^{tX}\right]\right\|_2
\end{aligned}$$

So we can repeat this process iteratively, peeling an X_i each time from the left term. For clarity, the next step gives,

$$\begin{aligned}
E_X\left[\mathrm{Tr}\left(e^{t\left(\sum_i^{s-1} X_i\right)}\right)\right] &\leq E_X\left[\mathrm{Tr}\Big(e^{t\left(\sum_i^{s-2} X_i\right)}e^{tX_{s-1}}\Big)\right] \\
&\leq E_X\left[\mathrm{Tr}\left(e^{t\left(\sum_i^{s-2} X_i\right)}\right)\right]\left\|E_X\left[e^{tX}\right]\right\|_2 \\
&\qquad \text{(applying (14.8.1) again)}
\end{aligned}$$

Therefore, after peeling all terms but the last X_i, we have

$$E_X\Big[\mathrm{Tr}\big(e^{tS}\big)\Big] \leq E_X\left[\mathrm{Tr}\left(e^{tX}\right)\right]\left\|E_X\left[e^{tX}\right]\right\|_2^{s-1}$$

Hence, since the trace is the sum of eigenvalues, $\mathrm{Tr}\left(e^{tX}\right) \leq d\lambda_{\max}(e^{tX})$ i.e. the worst case of all d eigenvalues being the max

$$\leq d\left\|E_X\left[e^{tX}\right]\right\|_2^s$$

as desired. □

Lemma 14.13. *If $B \in \mathbb{C}^{d\times d}$ is a hermitian matrix for which $\|B\|_2 \leq 1$, then $e^B \leq I + B + B^2$*

Proof. We know that $e^{\lambda_i} \leq 1 + \lambda_i + \lambda_i^2$, $|\lambda_i|^2 \leq 1$. Hence,

$$e^{\lambda_i} |v_i\rangle \langle v_i| \leq (1 + \lambda_i + \lambda_i^2) |v_i\rangle \langle v_i|$$

where $|v_i\rangle$ is the corresponding eigenvector. This then implies

$$\begin{aligned} e^B &= \sum_{i=1}^{d} e^{\lambda_i} |v_i\rangle \langle v_i| \preceq \sum_{i=1}^{d} (1 + \lambda_i + \lambda_i^2) |v_i\rangle \langle v_i| \\ &= I + B + B^2 \end{aligned}$$

□

15 Online Learning of Quantum States

15.1 Goals

We will prove that

Theorem 15.1. *Let $E_1, E_2, \cdots$ be a sequence of two-outcome measurements on an n-qubit state presented to the learner, and $l_1, l_2, \cdots$ be the corresponding loss functions revealed in successive iterations in the regret minimization model. Suppose l_t is convex and L-lipschitz; in particular for every $x \in \mathbb{R}$, there is a subderivative $l_t'(x)$ such that $|l_t'(x)| \leq L$. Then there is an explicit learning strategy that guarantees regret $R_T = O(L\sqrt{Tn})$ for all T. This is so even assuming the measurement E_t and loss function l_t are chosen adaptively, in response to the learner's previous behavior.*

Specifically, the algorithm applies to L_1 loss and L_2 loss, and achieves regret $O(\sqrt{Tn})$ for both.

15.2 Online Learning and Regret

Appendices

A Quantum Mechanics

Definition A.1. *Pauli Matrices*

$$\sigma_x = X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
$$\sigma_y = Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$
$$\sigma_z = Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Definition A.2. *Bell States*

$$\frac{|00\rangle + |11\rangle}{\sqrt{2}}$$
$$\frac{|00\rangle - |11\rangle}{\sqrt{2}}$$
$$\frac{|10\rangle + |01\rangle}{\sqrt{2}}$$
$$\frac{|01\rangle - |10\rangle}{\sqrt{2}}$$

Definition A.3. *Positive Operators*

Let A be a bounded[7] linear operator on complex Hilbert space $\mathcal{H}$. The following conditions are equivalent to A being positive

1. *$A = S^\dagger S$ for some bounded operator S on $\mathcal{H}$*
2. *A is hermitian and $\langle x| A |x\rangle \geq 0$ for every $|x\rangle \in \mathcal{H}$*
3. *the spectrum of A is non-negative*

Definition A.4. *Trace of an Operator*

Let $\{|i\rangle\}$ be an orthonormal basis for A and so

$$\begin{aligned} \operatorname{tr}(A) &= \sum_i A_{ii} \\ &= \sum_i \langle i| A |i\rangle \end{aligned}$$

Hence, if we extend $|\psi\rangle$ to the orthonormal basis $\{|i\rangle\}$ which includes $|\psi\rangle$ as the first element (for example via the Gram-Schmidt procedure) then

$$\begin{aligned} \operatorname{tr}(A |\psi\rangle \langle\psi|) &= \sum_i \langle i| A |\psi\rangle \langle\psi|i\rangle \\ &= \langle\psi| A |\psi\rangle \end{aligned}$$

by orthonormality.

[7] $\|Av\| \leq M\|v\|$ for some $M > 0$ and all $v \in \mathcal{H}$

Theorem A.5. *Spectral Theorem*

Suppose A is a compact[8] *hermitian operator (compactness ensures A has eigenvectors) on complex Hilbert space $\mathcal{H}$. Hence, there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of A. Each eigenvalue is in $\mathbb{R}$.*

References

[1] Scott Aaronson. "Read the fine print". In: *Nature Physics* 11.4 (2015), p. 291.

[2] Scott Aaronson. "The learnability of quantum states". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 463. 2088. The Royal Society. 2007, pp. 3089–3114.

[3] Andris Ambainis. "Variable time amplitude amplification and quantum algorithms for linear algebra problems". In: *STACS'12 (29th Symposium on Theoretical Aspects of Computer Science)*. Vol. 14. LIPIcs. 2012, pp. 636–647.

[4] Srinivasan Arunachalam and Ronald de Wolf. "Guest column: a survey of quantum learning theory". In: *ACM SIGACT News* 48.2 (2017), pp. 41–67.

[5] Srinivasan Arunachalam and Ronald de Wolf. "Optimal quantum sample complexity of learning algorithms". In: *arXiv preprint arXiv:1607.00932* (2016).

[6] Jacob Biamonte et al. "Quantum machine learning". In: *Nature* 549.7671 (2017), p. 195.

[7] Gilles Brassard et al. "Quantum amplitude amplification and estimation". In: *Contemporary Mathematics* 305 (2002), pp. 53–74.

[8] Yudong Cao et al. "Quantum Circuit Design for Solving Linear Systems of Equations". In: 110 (Oct. 2011).

[9] Andrew M Childs, Robin Kothari, and Rolando D Somma. "Quantum linear systems algorithm with exponentially improved dependence on precision". In: *arXiv preprint arXiv:1511.02306* (2015).

[10] Carlo Ciliberto et al. "Quantum machine learning: a classical perspective". In: *Proc. R. Soc. A* 474.2209 (2018), p. 20170551.

[11] Don Coppersmith and Shmuel Winograd. "Matrix multiplication via arithmetic progressions". In: *Proceedings of the nineteenth annual ACM symposium on Theory of computing*. ACM. 1987, pp. 1–6.

[8] the image under A acting on any bounded subset of $\mathcal{H}$ is a compact subset of $\mathcal{H}$

[12] Daoyi Dong et al. “Quantum reinforcement learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38.5 (2008), pp. 1207–1220.

[13] Vedran Dunjko and Hans J Briegel. “Machine learning & artificial intelligence in the quantum domain: a review of recent progress”. In: *Reports on Progress in Physics* (2018). URL: http://iopscience.iop.org/10.1088/1361-6633/aab406.

[14] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. “A quantum approximate optimization algorithm”. In: *arXiv preprint arXiv:1411.4028* (2014).

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, 2001.

[16] András Gilyén, Seth Lloyd, and Ewin Tang. “Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension”. In: *arXiv preprint arXiv:1811.04909* (2018).

[17] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. “Quantum random access memory”. In: *Physical review letters* 100.16 (2008), p. 160501.

[18] Gene H Golub and Charles F Van Loan. *Matrix computations*. Vol. 3. JHU Press, 2012.

[19] Lov K Grover. “A fast quantum mechanical algorithm for database search”. In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. ACM. 1996, pp. 212–219.

[20] Vojtech Havlicek et al. “Supervised learning with quantum enhanced feature spaces”. In: *arXiv preprint arXiv:1804.11326* (2018).

[21] Ravindran Kannan and Santosh Vempala. “Randomized algorithms in numerical linear algebra”. In: *Acta Numerica* 26 (2017), pp. 95–135.

[22] Iordanis Kerenidis and Anupam Prakash. “Quantum recommendation systems”. In: *arXiv preprint arXiv:1603.08675* (2016).

[23] Seth Lloyd. “Quantum algorithm for solving linear systems of equations”. In: *APS March Meeting Abstracts*. 2010.

[24] Seth Lloyd, Silvano Garnerone, and Paolo Zanardi. “Quantum algorithms for topological and geometric analysis of data”. In: *Nature communications* 7 (2016), p. 10138.

[25] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. “Quantum algorithms for supervised and unsupervised machine learning”. In: *arXiv preprint arXiv:1307.0411* (2013).

[26] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. "Quantum principal component analysis". In: *Nature Physics* 10.9 (2014), p. 631.

[27] Seth Lloyd and Christian Weedbrook. "Quantum generative adversarial learning". In: *arXiv preprint arXiv:1804.09139* (2018).

[28] Kosuke Mitarai et al. "Quantum Circuit Learning: Framework for Machine Learning with Quantum Enhanced Feature Space." In: ().

[29] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

[30] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. "Quantum support vector machine for big data classification". In: *Physical review letters* 113.13 (2014), p. 130503.

[31] Cynthia Rudin. "Support Vector Machines". In: *Duke Course Notes* ().

[32] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. "Prediction by linear regression on a quantum computer". In: *Physical Review A* 94.2 (2016), p. 022342.

[33] Changpeng Shao. "Reconsider HHL algorithm and its related quantum machine learning algorithms". In: *arXiv preprint arXiv:1803.01486* (2018).

[34] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge, 1998.

[35] Ewin Tang. "Quantum-inspired classical algorithms for principal component analysis and supervised clustering". In: *arXiv preprint arXiv:1811.00414* (2018).

[36] Nathan Wiebe, Daniel Braun, and Seth Lloyd. "Quantum algorithm for data fitting". In: *Physical review letters* 109.5 (2012), p. 050505.

[37] Leonard Wossnig, Zhikuan Zhao, and Anupam Prakash. "Quantum linear system algorithm for dense matrices". In: *Physical review letters* 120.5 (2018), p. 050502.

[38] VU Thi Xuan. "CR04 Report: Solving linear equations on a quantum computer". In: (2016).

[39] Zhikuan Zhao et al. "A note on state preparation for quantum machine learning". In: *arXiv preprint arXiv:1804.00281* (2018).

[40] Zhikuan Zhao et al. "Quantum algorithms for training Gaussian Processes". In: *arXiv preprint arXiv:1803.10520* (2018).

[41] Yarui Zheng et al. "Solving systems of linear equations with a superconducting quantum processor". In: *Physical review letters* 118.21 (2017), p. 210504.