A Survey of Quantum Learning Theory

by: Srinivasan Arunachalam, Ronald de Wolf

Yoni Asulin

December 2018

Outline

Intro and motivation

- Machine learning has become a very hot topic, giving outstanding results in multiple areas of CS such as image recognition, natural language processing and many more

Intro and motivation

- Machine learning has become a very hot topic, giving outstanding results in multiple areas of CS such as image recognition, natural language processing and many more
- we also know how powerful quantum computing might be on certain tasks

Intro and motivation

- Machine learning has become a very hot topic, giving outstanding results in multiple areas of CS such as image recognition, natural language processing and many more
- we also know how powerful quantum computing might be on certain tasks
- for example, classical hard problems such as Factoring and Discrete Log, can be solved efficiently using quantum computing (Shor's algorithm)

Intro and motivation

- Machine learning has become a very hot topic, giving outstanding results in multiple areas of CS such as image recognition, natural language processing and many more
- we also know how powerful quantum computing might be on certain tasks
- for example, classical hard problems such as Factoring and Discrete Log, can be solved efficiently using quantum computing (Shor's algorithm)

the question is:

Can we exploit the power of Quantum Computing to learn more efficiently? (in terms of sample complexity and runtime)

Quick recap of QC notation

- state vector of a single qubit:

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

- superposition: $|\alpha|^2 + |\beta|^2 = 1$
- single qubit lives in a 2-dimensional Hilbert space $\mathcal{H}$
- a system of n qubits live in a 2^n Hilbert space $\mathcal{H}^{\otimes n}$, and the state vector of the system is the tensor product of all the state vectors of the individual qubits

density operator

- an alternative formalism to state vector formalism

density operator

- an alternative formalism to state vector formalism
- very useful in composite systems (where it can be in some mixed state)

density operator

- an alternative formalism to state vector formalism
- very useful in composite systems (where it can be in some mixed state)

Definition: density operator

given an ensemble of states $\mathcal{E} = \{|\psi_i\rangle, p_i\}_{i=1}^{m}$, its density operator is defined as:

$$\rho = \sum_{i=1}^{m} p_i |\psi_i\rangle\langle\psi_i|$$

density operator

- an alternative formalism to state vector formalism
- very useful in composite systems (where it can be in some mixed state)

Definition: density operator

given an ensemble of states $\mathcal{E} = \{\left|\psi_i\right\rangle, p_i\}_{i=1}^{m}$, its density operator is defined as:

$$\rho = \sum_{i=1}^{m} p_i \left|\psi_i\right\rangle\!\left\langle\psi_i\right|$$

Lemma

*An operator ρ is a **density operator** (with respect to some ensemble of states) if and only if:*

density operator

- an alternative formalism to state vector formalism
- very useful in composite systems (where it can be in some mixed state)

Definition: density operator

given an ensemble of states $\mathcal{E} = \{|\psi_i\rangle, p_i\}_{i=1}^{m}$, its density operator is defined as:

$$\rho = \sum_{i=1}^{m} p_i |\psi_i\rangle\langle\psi_i|$$

Lemma

An operator ρ is a ***density operator*** *(with respect to some ensemble of states) if and only if:*

1. *ρ is a positive operator*

density operator

- an alternative formalism to state vector formalism
- very useful in composite systems (where it can be in some mixed state)

Definition: density operator

given an ensemble of states $\mathcal{E} = \{|\psi_i\rangle, p_i\}_{i=1}^{m}$, its density operator is defined as:

$$\rho = \sum_{i=1}^{m} p_i |\psi_i\rangle\langle\psi_i|$$

Lemma

*An operator ρ is a **density operator** (with respect to some ensemble of states) if and only if:*

1. *ρ is a positive operator*
2. $tr(\rho) = 1$

Measurement formalism

- closed quantum systems, i.e, isolated which don't interact with the rest of the world, evolve according to unitary evolution

Measurement formalism

- closed quantum systems, i.e, isolated which don't interact with the rest of the world, evolve according to unitary evolution
- in reality, the experimentalist and their experimental equipment observes the system to find out what is going on inside the system, **an interaction which makes the system no longer closed**

Measurement formalism

- closed quantum systems, i.e, isolated which don't interact with the rest of the world, evolve according to unitary evolution
- in reality, the experimentalist and their experimental equipment observes the system to find out what is going on inside the system, **an interaction which makes the system no longer closed**
- in such a scenario, the system will not necessarily evolve acording to a unitary evolution

Measurement formalism

- closed quantum systems, i.e, isolated which don't interact with the rest of the world, evolve according to unitary evolution
- in reality, the experimentalist and their experimental equipment observes the system to find out what is going on inside the system, **an interaction which makes the system no longer closed**
- in such a scenario, the system will not necessarily evolve acording to a unitary evolution
- we need a mechanism to explain what happens when this is the case.

Measurement formalism

- closed quantum systems, i.e, isolated which don't interact with the rest of the world, evolve according to unitary evolution
- in reality, the experimentalist and their experimental equipment observes the system to find out what is going on inside the system, **an interaction which makes the system no longer closed**
- in such a scenario, the system will not necessarily evolve acording to a unitary evolution
- we need a mechanism to explain what happens when this is the case.
- example: when the experiment is not repeatable (measuring a photon destroys it...) - must apply general measurement

Measurement formalism

- closed quantum systems, i.e, isolated which don't interact with the rest of the world, evolve according to unitary evolution
- in reality, the experimentalist and their experimental equipment observes the system to find out what is going on inside the system, **an interaction which makes the system no longer closed**
- in such a scenario, the system will not necessarily evolve acording to a unitary evolution
- we need a mechanism to explain what happens when this is the case.
- example: when the experiment is not repeatable (measuring a photon destroys it...) - must apply general measurement
- example: as we will see, the optimal way to distinguish a set of quantum states involves a (special) general measurement

General Measurments

Postulate: general measuement.

a general measurement is defined by **measurment operators** $\{M_m\}_m$, which satisfy:

General Measurments

Postulate: general measuement.

a general measurement is defined by **measurment operators** $\{M_m\}_m$, which satisfy:

1. Completeness equation: $\sum_m M_m^\dagger M_m = I$

General Measurments

Postulate: general measuement.

a general measurement is defined by **measurment operators** $\{M_m\}_m$, which satisfy:

1. Completeness equation: $\sum_m M_m^\dagger M_m = I$
2. $Pr(m) = \langle\psi|M_m^\dagger M_m|\psi\rangle$

General Measurments

Postulate: general measuement.

a general measurement is defined by **measurment operators** $\{M_m\}_m$, which satisfy:

1. Completeness equation: $\sum_m M_m^\dagger M_m = I$
2. $Pr(m) = \langle\psi|M_m^\dagger M_m|\psi\rangle$
3. after measurment, if we got m, the system collapses to the state:

$$|\psi'\rangle = \frac{M_m|\psi\rangle}{\sqrt{\langle\psi|M_m^\dagger M_m|\psi\rangle}}$$

POVM measurment

- POVM $=$ positive-operator valued measure

POVM measurment

- POVM = positive-operator valued measure
- a special case of general measurments

POVM measurment

- POVM $=$ positive-operator valued measure
- a special case of general measurments
- very useful in applications where we don't care about the post-measurment state, and only care about the outcome statistics

POVM measurment

- POVM = positive-operator valued measure
- a special case of general measurments
- very useful in applications where we don't care about the post-measurment state, and only care about the outcome statistics
- for example, in an experiment where the system is measured only once

POVM measurment

- POVM = positive-operator valued measure
- a special case of general measurments
- very useful in applications where we don't care about the post-measurment state, and only care about the outcome statistics
- for example, in an experiment where the system is measured only once
- more appropriate for open systems, such in labs, where noise is present and a completely closed (isolated) system is not possible.

POVM measurment

- POVM $=$ positive-operator valued measure
- a special case of general measurments
- very useful in applications where we don't care about the post-measurment state, and only care about the outcome statistics
- for example, in an experiment where the system is measured only once
- more appropriate for open systems, such in labs, where noise is present and a completely closed (isolated) system is not possible.

Definition: positive operator

an operator A is **positive** if for every vector $|v\rangle$ it holds that $(|v\rangle, A|v\rangle) \geq 0$

POVM measurment

Definition:POVM

a set of measurment operators $\{E_i\}_{i=1}^{m}$, such that:

POVM measurment

Definition:POVM

a set of measurment operators $\{E_i\}_{i=1}^{m}$, such that:

1. Completeness equation: $\sum_{i=1}^{m} E_i = I$

POVM measurment

Definition:POVM

a set of measurment operators $\{E_i\}_{i=1}^{m}$, such that:

1. Completeness equation: $\sum_{i=1}^{m} E_i = I$
2. E_i is a positive operator, for every $i \in [m]$

POVM measurment

Definition:POVM

a set of measurment operators $\{E_i\}_{i=1}^m$, such that:

1. Completeness equation: $\sum_{i=1}^m E_i = I$
2. E_i is a positive operator, for every $i \in [m]$
3. if the system (prior to measurement) is in state ψ, then upom measurment:

$$Pr(i) = \langle\psi|E_i|\psi\rangle$$

POVM measurment

Definition:POVM

a set of measurment operators $\{E_i\}_{i=1}^m$, such that:

1. Completeness equation: $\sum_{i=1}^m E_i = I$
2. E_i is a positive operator, for every $i \in [m]$
3. if the system (prior to measurement) is in state ψ, then upom measurment:
$$Pr(i) = \langle\psi|E_i|\psi\rangle$$
4. after measuement, the system will be in state:
$$\frac{E_i\,|\psi\rangle}{\sqrt{\langle\psi|E_i|\psi\rangle}}$$

PGM - motivation

- given an ensemble of unknown pure states $\mathcal{E} = \{p_i, |\psi_i\rangle\}_{i\in[m]}$.

PGM - motivation

- given an ensemble of unknown pure states $\mathcal{E} = \{p_i, |\psi_i\rangle\}_{i\in[m]}$.
- Suppose Alice picks at random a state $|\psi_?\rangle \in \mathcal{E}$ (according to the apriori pobabilities), and sends it to Bob.

PGM - motivation

- given an ensemble of unknown pure states $\mathcal{E} = \{p_{i,} |\psi_i\rangle\}_{i\in[m]}$.
- Suppose Alice picks at random a state $|\psi_?\rangle \in \mathcal{E}$ (according to the apriori pobabilities), and sends it to Bob.
- The goal for Bob is to to identify the index i of the state Alice gave him.

PGM - motivation

- given an ensemble of unknown pure states $\mathcal{E} = \{p_i, |\psi_i\rangle\}_{i\in[m]}$.
- Suppose Alice picks at random a state $|\psi_?\rangle \in \mathcal{E}$ (according to the apriori pobabilities), and sends it to Bob.
- The goal for Bob is to to identify the index i of the state Alice gave him.
- Bob does so by defining the appropriate measurment operators and use them to get the result (has to choose cleverly)

PGM - motivation

- given an ensemble of unknown pure states $\mathcal{E} = \{p_i, |\psi_i\rangle\}_{i\in[m]}$.
- Suppose Alice picks at random a state $|\psi_?\rangle \in \mathcal{E}$ (according to the apriori pobabilities), and sends it to Bob.
- The goal for Bob is to to identify the index i of the state Alice gave him.
- Bob does so by defining the appropriate measurment operators and use them to get the result (has to choose cleverly)
- A fundamental property of quantum mechanics is that **non-orthogonal pure quantum states may not be distinguished perfectly** (Bob will fail some of the times)

Pretty Good Measurement (PGM)

motivational problem:

Let $\mathcal{E} = \{|\psi_i\rangle, p_i\}_{i \in [m]}$ be and ensemble of m d-dimensional pure states $|\psi_i\rangle$ with their apriori probabilities p_i:
Given an unknown state $|\psi_?\rangle$, picked at random from $\mathcal{E}$, what is the optimal probability of identifying $|\psi_?\rangle$?

Pretty Good Measurement (PGM)

motivational problem:

Let $\mathcal{E} = \{\left|\psi_i\right\rangle, p_i\}_{i\in[m]}$ be and ensemble of m d-dimensional pure states $\left|\psi_i\right\rangle$ with their apriori probabilities p_i:

Given an unknown state $\left|\psi_?\right\rangle$, picked at random from $\mathcal{E}$, what is the optimal probability of identifying $\left|\psi_?\right\rangle$?

- i.e, we want:

$$P^{opt}(\mathcal{E}) = max_{\mathcal{M}} \sum_{i=1}^{m} p_i \left\langle\psi\right|E_i\left|\psi\right\rangle$$

Pretty Good Measurement (PGM)

motivational problem:

Let $\mathcal{E} = \{|\psi_i\rangle, p_i\}_{i\in[m]}$ be and ensemble of m d-dimensional pure states $|\psi_i\rangle$ with their apriori probabilities p_i:
Given an unknown state $|\psi_?\rangle$, picked at random from $\mathcal{E}$, what is the optimal probability of identifying $|\psi_?\rangle$?

- i.e, we want:

$$P^{opt}(\mathcal{E}) = max_{\mathcal{M}} \sum_{i=1}^{m} p_i \langle\psi|E_i|\psi\rangle$$

- where the maximum is taken over all m-outcome POVMs $\mathcal{M}$.

Pretty Good Measurement (PGM)

- for the case of $m = 2$ (where $\mathcal{E}$ contains two states) there is an analytic expression for $P^{opt}(\mathcal{E})$.
- but for $m \geq 3$ the problem seems intractable.
- we therefore want **lower bounds** for P^{opt}
- Pretty Good Measurement (PGM) is a specific POVM (depending on $\mathcal{E}$), that does reasonably well against $\mathcal{E}$.

Pretty Good Measurement (PGM)

- For pure states, the PGM is defined by the set of measurement operators $E_i = |\mu_i\rangle\langle\mu_i|$, where:

$$|\mu_i\rangle = \sqrt{p_i}\rho^{-1/2} |\psi_i\rangle$$

.

Pretty Good Measurement (PGM)

- For pure states, the PGM is defined by the set of measurement operators $E_i = |\mu_i\rangle\langle\mu_i|$, where:

$$|\mu_i\rangle = \sqrt{p_i}\rho^{-1/2} |\psi_i\rangle$$

.

- where $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$ is the density operator for the ensemble $\mathcal{E}$.

Pretty Good Measurement (PGM)

- For pure states, the PGM is defined by the set of measurement operators $E_i = |\mu_i\rangle\langle\mu_i|$, where:

$$|\mu_i\rangle = \sqrt{p_i}\rho^{-1/2}\,|\psi_i\rangle$$

.

- where $\rho = \sum_i p_i\,|\psi_i\rangle\langle\psi_i|$ is the density operator for the ensemble $\mathcal{E}$.
- and $\rho^{-1/2}$ is the pseudo-inverse matrix.

Pretty Good Measurement (PGM)

- For pure states, the PGM is defined by the set of measurement operators $E_i = |\mu_i\rangle\langle\mu_i|$, where:

$$|\mu_i\rangle = \sqrt{p_i}\rho^{-1/2}\,|\psi_i\rangle$$

.

- where $\rho = \sum_i p_i\,|\psi_i\rangle\langle\psi_i|$ is the density operator for the ensemble $\mathcal{E}$.
- and $\rho^{-1/2}$ is the pseudo-inverse matrix.
  - in the context of these density matrices, which are diagonal (as they are sums of pure states created from an orthonormal basis), this simply corresponds to performing the inverse operation only on the diagonal elements

Pretty Good Measurement (PGM)

- For pure states, the PGM is defined by the set of measurement operators $E_i = |\mu_i\rangle\langle\mu_i|$, where:

$$|\mu_i\rangle = \sqrt{p_i}\rho^{-1/2}\,|\psi_i\rangle$$

.

- where $\rho = \sum_i p_i\,|\psi_i\rangle\langle\psi_i|$ is the density operator for the ensemble $\mathcal{E}$.
- and $\rho^{-1/2}$ is the pseudo-inverse matrix.
  - in the context of these density matrices, which are diagonal (as they are sums of pure states created from an orthonormal basis), this simply corresponds to performing the inverse operation only on the diagonal elements
- one can show that these operators give a valid measurement (completeness equation)

the main properties of PGM we will need

Lemma (without proof)

$$P^{opt}(\mathcal{E}) \geq P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$$

the main properties of PGM we will need

Lemma (without proof)

$$P^{opt}(\mathcal{E}) \geq P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$$

- *i.e, PGM is almost optimal for any ensemble* $\mathcal{E}$: $P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$

the main properties of PGM we will need

Lemma (without proof)

$$P^{opt}(\mathcal{E}) \geq P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$$

- *i.e, PGM is almost optimal for any ensemble $\mathcal{E}$:* $P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$

Lemma

Let G be the rescaled Gram matrix for the ensemble $\mathcal{E}$. i.e, $G_{ij} = \sqrt{p_i}\sqrt{p_j}\langle\psi_i|\psi_j\rangle$. Then the probability of success of the PGM is:

$$P^{pgm}(\mathcal{E}) = \sum_{i=1}^{m} p_i |\langle\psi_i|\mu_i\rangle|^2 = \sum_{i=1}^{m} (\sqrt{G})_{ii}^2$$

the main properties of PGM we will need

Lemma (without proof)

$$P^{opt}(\mathcal{E}) \geq P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$$

- *i.e, PGM is almost optimal for any ensemble $\mathcal{E}$: $P^{pgm}(\mathcal{E}) \geq P^{opt}(\mathcal{E})^2$*

Lemma

Let G be the rescaled Gram matrix for the ensemble $\mathcal{E}$. i.e, $G_{ij} = \sqrt{p_i}\sqrt{p_j}\langle\psi_i|\psi_j\rangle$. Then the probability of success of the PGM is:

$$P^{pgm}(\mathcal{E}) = \sum_{i=1}^{m} p_i|\langle\psi_i|\mu_i\rangle|^2 = \sum_{i=1}^{m}(\sqrt{G})^2_{ii}$$

- *the same states, renormalised to reflect their probabilities..*

- Sample space (Domain set): $\mathcal{X}$.

the statistical learning framework

- Sample space (Domain set): $\mathcal{X}$.
- Label set: $\mathcal{Y}$

the statistical learning framework

- Sample space (Domain set): $\mathcal{X}$.
- Label set: $\mathcal{Y}$
- Training Data: $S = \{(x_i, y_i)\} \in \mathcal{X} \times \mathcal{Y}$. a finite set of labeled examples.

the statistical learning framework

- Sample space (Domain set): $\mathcal{X}$.
- Label set: $\mathcal{Y}$
- Training Data: $S = \{(x_i, y_i)\} \in \mathcal{X} \times \mathcal{Y}$. a finite set of labeled examples.
- the learner's output:

the statistical learning framework

- Sample space (Domain set): $\mathcal{X}$.
- Label set: $\mathcal{Y}$
- Training Data: $S = \{(x_i, y_i)\} \in \mathcal{X} \times \mathcal{Y}$. a finite set of labeled examples.
- the learner's output:

the learner's output:

the goal of the learner is to come up with a **prediction rule** $h : \mathcal{X} \to \mathcal{Y}$ that can be used to lablel any fresh sampled example $x \in \mathcal{X}$.

some remarks

- the examples in $\mathcal{X}$ distribute according to some probability distribution $\mathcal{D}$,**which is unknown to the learner.**

some remarks

- the examples in $\mathcal{X}$ distribute according to some probability distribution $\mathcal{D}$,**which is unknown to the learner.**
- the learner is also not aware of the **true labeling function** $f : \mathcal{X} \to \mathcal{Y}$.

some remarks

- the examples in $\mathcal{X}$ distribute according to some probability distribution $\mathcal{D}$,**which is unknown to the learner.**
- the learner is also not aware of the **true labeling function** $f : \mathcal{X} \to \mathcal{Y}$.
- an equivalent way to describe this scenerio is that the learner has access to a **random example oracle** $PEX(c, \mathcal{D})$, which when invoked, draws a fresh sample $x \in \mathcal{X}$ (i.i.d, according to $\mathcal{D}$) and returns the labeled example $(x, f(x))$.

some remarks

- the examples in $\mathcal{X}$ distribute according to some probability distribution $\mathcal{D}$,**which is unknown to the learner.**
- the learner is also not aware of the **true labeling function** $f : \mathcal{X} \to \mathcal{Y}$.
- an equivalent way to describe this scenerio is that the learner has access to a **random example oracle** $PEX(c, \mathcal{D})$, which when invoked, draws a fresh sample $x \in \mathcal{X}$ (i.i.d, according to $\mathcal{D}$) and returns the labeled example $(x, f(x))$.
- the **sample complexity** for a class $\mathcal{H}$ is the number of examples that are required to guarantee a probably approximately correct solution.

some remarks

- the examples in $\mathcal{X}$ distribute according to some probability distribution $\mathcal{D}$,**which is unknown to the learner.**
- the learner is also not aware of the **true labeling function** $f : \mathcal{X} \to \mathcal{Y}$.
- an equivalent way to describe this scenerio is that the learner has access to a **random example oracle** $PEX(c, \mathcal{D})$, which when invoked, draws a fresh sample $x \in \mathcal{X}$ (i.i.d, according to $\mathcal{D}$) and returns the labeled example $(x, f(x))$.
- the **sample complexity** for a class $\mathcal{H}$ is the number of examples that are required to guarantee a probably approximately correct solution.

Definition

We define the **error** of an hypothesis $h : \mathcal{X} \to \mathcal{Y}$ to be:

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)]$$

Shattering

Definition

Let $C = \{x_1, ..., x_m\} \subseteq \mathcal{X}$. The **restriction of $\mathcal{H}$ to C** is defined as all function from C to $\mathcal{Y}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \{h(x_1), ..., h(x_m) \mid h \in \mathcal{H}\}$$

Shattering

Definition

Let $C = \{x_1, ..., x_m\} \subseteq \mathcal{X}$. The **restriction of $\mathcal{H}$ to C** is defined as all function from C to $\mathcal{Y}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \{h(x_1), ..., h(x_m) \mid h \in \mathcal{H}\}$$

Definition

(Shattering) A hypothesis class $\mathcal{H}$ **shatters** a finite set $C \subseteq \mathcal{X}$ if

$$|\mathcal{H}_C| = 2^{|C|}$$

i.e, the restriction of $\mathcal{H}$ to C is the set of **all functions** from C to $\mathcal{Y}$.

trivial example - Threshold functions

Threshold functions

Let $a \in \mathbb{R}$. define $h_a : \mathbb{R} \to \{0,1\}$ to be $h_a(x) = \mathbb{1}_{[x<a]}$.
Define the class of threshold functions: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$

trivial example - Threshold functions

Threshold functions

Let $a \in \mathbb{R}$. define $h_a : \mathbb{R} \rightarrow \{0, 1\}$ to be $h_a(x) = \mathbb{1}_{[x<a]}$.
Define the class of threshold functions: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$

- for every singleton $x_0 \in \mathbb{R}$, $\mathcal{H}$ shatters the set $C = \{x_0\}$

trivial example - Threshold functions

Threshold functions

Let $a \in \mathbb{R}$. define $h_a : \mathbb{R} \to \{0, 1\}$ to be $h_a(x) = \mathbb{1}_{[x<a]}$.
Define the class of threshold functions: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$

- for every singleton $x_0 \in \mathbb{R}$, $\mathcal{H}$ shatters the set $C = \{x_0\}$
- but, for every $x_1 < x_2$, $\mathcal{H}$ does not shatter $C = \{x_1, x_2\}$ (why?)

VC dimension

Definition

The VC-dimension of an hypothesis class $\mathcal{H}$, denoted $VCdim(\mathcal{H})$ is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

VC dimension

Definition

The VC-dimension of an hypothesis class $\mathcal{H}$, denoted $VCdim(\mathcal{H})$ is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

- it turns out that VC dimension characterizes PAC learnability:

VC dimension

Definition

The VC-dimension of an hypothesis class $\mathcal{H}$, denoted $VCdim(\mathcal{H})$ is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

- it turns out that VC dimension characterizes PAC learnability:

Theorem

A class $\mathcal{H}$ is PAC-learnable if and only if $VCdim(\mathcal{H}) < \infty$

back to threshold functions

- remainder: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$

back to threshold functions

- remainder: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$
- we saw that for every singleton $x \in \mathbb{R}$, $\mathcal{H}$ shatters C.
  $\implies \quad VCdim(\mathcal{H}) \leq 1$

back to threshold functions

- remainder: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$
- we saw that for every singleton $x \in \mathbb{R}$, $\mathcal{H}$ shatters C.
  $\implies \quad VCdim(\mathcal{H}) \leq 1$
- and for every $x_1 < x_2$, $\mathcal{H}$ does not shatter $C = \{x_1, x_2\}$

back to threshold functions

- remainder: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$
- we saw that for every singleton $x \in \mathbb{R}$, $\mathcal{H}$ shatters C.
  $\implies \quad VCdim(\mathcal{H}) \leq 1$
- and for every $x_1 < x_2$, $\mathcal{H}$ does not shatter $C = \{x_1, x_2\}$
- $\implies \quad VCdim(\mathcal{H}) = 1$ and thus $\mathcal{H}$ is PAC learnable

reminder from IML

Theorem

Let $\mathcal{H}$ be an hypothesis class with VCdim($\mathcal{H}$) $= d + 1$. Then $\Theta(\frac{d}{\varepsilon} + \frac{log(1/\delta)}{\varepsilon})$ examples are necessary and sufficient for an (ε, δ)-PAC learner for $\mathcal{H}$.

Quantum PAC learning

the model

The learner has access to a quantum example oracle $QPEX(c, D)$ that produces an example:

$$\sum_{x \in \mathcal{X}} \sqrt{D(x)}\, |x, f(x)\rangle$$

Quantum PAC learning

the model

The learner has access to a quantum example oracle $QPEX(c, D)$ that produces an example:

$$\sum_{x \in \mathcal{X}} \sqrt{D(x)}\, |x, f(x)\rangle$$

- the quantum PAC learner is given access to **several copies** of the quantum example

Quantum PAC learning

the model

The learner has access to a quantum example oracle $QPEX(c, D)$ that produces an example:

$$\sum_{x \in \mathcal{X}} \sqrt{D(x)}\,|x, f(x)\rangle$$

- the quantum PAC learner is given access to **several copies** of the quantum example
- then he performs a POVM measuement, such that each outcome is associated with an hypothesis in $\mathcal{H}$.

Quantum PAC learning

the model

The learner has access to a quantum example oracle $QPEX(c, D)$ that produces an example:

$$\sum_{x \in \mathcal{X}} \sqrt{D(x)}\, |x, f(x)\rangle$$

- the quantum PAC learner is given access to **several copies** of the quantum example
- then he performs a POVM measuement, such that each outcome is associated with an hypothesis in $\mathcal{H}$.
- then, the learner needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.

sample complexity of Quantum PAC Learning

- the sample complexity of the learner is defined as the maximum number of invocations of the oracle, over all distributions $\mathcal{D}$and over the internal randomness of the learner

sample complexity of Quantum PAC Learning

- the sample complexity of the learner is defined as the maximum number of invocations of the oracle, over all distributions $\mathcal{D}$and over the internal randomness of the learner

Definition

the (ε, δ)-quantum PAC sample complexity of a hypothesis class $\mathcal{H}$ is the minimum sample complexity over all (ε, δ)-quantum PAC learners for $\mathcal{H}$.

sample complexity of Quantum PAC Learning

Theorem

Let $\mathcal{H}$ be an hypothesis class with VCdim($\mathcal{H}$) $= d + 1$. Then, for every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1/20)$, then $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log \frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$.

sample complexity of Quantum PAC Learning

Theorem

Let $\mathcal{H}$ be an hypothesis class with $VCdim(\mathcal{H}) = d + 1$. Then, for every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1/20)$, then $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log \frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$.

- i.e, quantum examples <u>are not more powerful</u> than classical examples in the PAC model.

sample complexity of Quantum PAC Learning

Theorem

Let $\mathcal{H}$ be an hypothesis class with $VCdim(\mathcal{H}) = d + 1$. Then, for every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1/20)$, then $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log \frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$.

- i.e, quantum examples are not more powerful than classical examples in the PAC model.
- (however, we will show later that for some particular cases quantum examples can be more powerful)

sample complexity of Quantum PAC Learning

Theorem

Let $\mathcal{H}$ be an hypothesis class with $VCdim(\mathcal{H}) = d + 1$. Then, for every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1/20)$, then $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log \frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$.

- i.e, quantum examples <u>are not more powerful</u> than classical examples in the PAC model.
- (however, we will show later that for some particular cases quantum examples can be more powerful)
- we will use PGM and linear error correcting codes to show the $\Omega(\frac{d}{\varepsilon})$ bound.

Linear error-correcting codes

- A linear code C encoding d bits of information into a k bit code space is specified by a $d \times k$ **generator matrix** G whose entries are all elements of $\mathbb{Z}_2$. G encodes a d-bits message into a k-bit codeword Gx.

Linear error-correcting codes

- A linear code C encoding d bits of information into a k bit code space is specified by a $d \times k$ **generator matrix** G whose entries are all elements of $\mathbb{Z}_2$. G encodes a d-bits message into a k-bit codeword Gx.
- in order that all messages be uniquely encoded we require that the columns of G be linearly independent.

Linear error-correcting codes

- A linear code C encoding d bits of information into a k bit code space is specified by a $d \times k$ **generator matrix** G whose entries are all elements of $\mathbb{Z}_2$. G encodes a d-bits message into a k-bit codeword Gx.
- in order that all messages be uniquely encoded we require that the columns of G be linearly independent.
- example: $[6, 2]$ code

Linear error-correcting codes

- A linear code C encoding d bits of information into a k bit code space is specified by a $d \times k$ **generator matrix** G whose entries are all elements of $\mathbb{Z}_2$. G encodes a d-bits message into a k-bit codeword Gx.
- in order that all messages be uniquely encoded we require that the columns of G be linearly independent.
- example: $[6, 2]$ code

Hamming distance

Suppose x and y are words of n bits each. The **Hamming distance** between x and y is defined to be the number of places at which x and y differ: $d(x, y) = |\{i : x_i \neq y_i\}|$
Given a set C of n-bit codewords, we define its distance to be:

$$d(C) = \min_{x \neq y \in C} d(x, y)$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.
- we want to show that $T \geq \Omega(\frac{d}{\varepsilon})$.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.
- we want to show that $T \geq \Omega(\frac{d}{\varepsilon})$.
- let $S = \{s_0, s_1, ...s_d\} \subseteq \{0,1\}^n$ be a maximal set shattered by $\mathcal{H}$ $(VCdim(\mathcal{H}) = d + 1)$.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.
- we want to show that $T \geq \Omega(\frac{d}{\varepsilon})$.
- let $S = \{s_0, s_1, ...s_d\} \subseteq \{0,1\}^n$ be a maximal set shattered by $\mathcal{H}$ $(VCdim(\mathcal{H}) = d + 1)$.
- we can define a distribution D on S as follows:

$$D(s_i) = \begin{cases} 1 - 20\varepsilon, & i = 0 \\ 20\varepsilon/d, & 1 \leq i \leq d \end{cases}$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.
- we want to show that $T \geq \Omega(\frac{d}{\varepsilon})$.
- let $S = \{s_0, s_1, ...s_d\} \subseteq \{0,1\}^n$ be a maximal set shattered by $\mathcal{H}$ ($VCdim(\mathcal{H}) = d + 1$).
- we can define a distribution D on S as follows:

$$D(s_i) = \begin{cases} 1 - 20\varepsilon, & i = 0 \\ 20\varepsilon/d, & 1 \leq i \leq d \end{cases}$$

- we will use $[d, k, r]$ linear error-correcting code for $k \geq d/4$ and distance $r \geq d/8$ and generator matrix $M \in \mathbb{F}_2^{d \times k}$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.
- we want to show that $T \geq \Omega(\frac{d}{\varepsilon})$.
- let $S = \{s_0, s_1, ...s_d\} \subseteq \{0,1\}^n$ be a maximal set shattered by $\mathcal{H}$ ($VCdim(\mathcal{H}) = d+1$).
- we can define a distribution D on S as follows:

$$D(s_i) = \begin{cases} 1 - 20\varepsilon, & i = 0 \\ 20\varepsilon/d, & 1 \leq i \leq d \end{cases}$$

- we will use $[d, k, r]$ linear error-correcting code for $k \geq d/4$ and distance $r \geq d/8$ and generator matrix $M \in \mathbb{F}_2^{d \times k}$
- the 2^k codewords in this linear code are $\{Mz \mid z \in \{0,1\}^k\}$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- reminder: the quantum PAC learner is given access to T copies of the quantum example and needs to output an hypothesis $h \in \mathcal{H}$ that is ε-close to f.
- we want to show that $T \geq \Omega(\frac{d}{\varepsilon})$.
- let $S = \{s_0, s_1, ...s_d\} \subseteq \{0,1\}^n$ be a maximal set shattered by $\mathcal{H}$ ($VCdim(\mathcal{H}) = d + 1$).
- we can define a distribution D on S as follows:

$$D(s_i) = \begin{cases} 1 - 20\varepsilon, & i = 0 \\ 20\varepsilon/d, & 1 \leq i \leq d \end{cases}$$

- we will use $[d, k, r]$ linear error-correcting code for $k \geq d/4$ and distance $r \geq d/8$ and generator matrix $M \in \mathbb{F}_2^{d \times k}$
- the 2^k codewords in this linear code are $\{Mz \mid z \in \{0,1\}^k\}$
- Hamming distance for this set is $d_H(Mz, My) \geq d/8$ for every $z \neq y$.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- for each $z \in \{0,1\}^k$ we define the hypothesis on the shattered set S, $h_z : S \to \{0,1\}$ to be:

$$h_z(s_i) = \begin{cases} 0, & i = 0 \\ (Mz)_i, & 1 \leq i \leq d \end{cases}$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- for each $z \in \{0,1\}^k$ we define the hypothesis on the shattered set S, $h_z : S \to \{0,1\}$ to be:

$$h_z(s_i) = \begin{cases} 0, & i = 0 \\ (Mz)_i, & 1 \leq i \leq d \end{cases}$$

- why such functions exist?

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- for each $z \in \{0,1\}^k$ we define the hypothesis on the shattered set S, $h_z : S \to \{0,1\}$ to be:

$$h_z(s_i) = \begin{cases} 0, & i = 0 \\ (Mz)_i, & 1 \leq i \leq d \end{cases}$$

- why such functions exist?
- observation: since $r \geq d/8$, it follows that for every $z \neq y$ in $\{0,1\}^k$:

$$\mathbb{P}_{s \sim_D S}[h_z(s) \neq h_y(s)] \geq 5\varepsilon/2$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound

- for each $z \in \{0,1\}^k$ we define the hypothesis on the shattered set S, $h_z : S \to \{0,1\}$ to be:

$$h_z(s_i) = \begin{cases} 0, & i = 0 \\ (Mz)_i, & 1 \leq i \leq d \end{cases}$$

- why such functions exist?
- observation: since $r \geq d/8$, it follows that for every $z \neq y$ in $\{0,1\}^k$:

$$\mathbb{P}_{s \sim_D S}[h_z(s) \neq h_y(s)] \geq 5\varepsilon/2$$

- $\implies$ with probability$\geq 1 - \delta$, an (ε, δ)- PAC quantum learner trying to ε-approximate an hypothesis $h \in \{h_z \mid z \in \{0,1\}^k\}$ will **exactly** identify the hypothesis.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- for each $z \in \{0,1\}^k$ we define the state

$$|\psi_z\rangle = \sum_{i=0}^{d} \sqrt{D(s_i)}\,|s_i, h_z(s_i)\rangle$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- for each $z \in \{0,1\}^k$ we define the state

$$|\psi_z\rangle = \sum_{i=0}^{d} \sqrt{D(s_i)}\,|s_i, h_z(s_i)\rangle$$

- now we take the ensemble $\mathcal{E} = \{|\psi_z\rangle^{\otimes T}, 2^{-k}\}$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- for each $z \in \{0,1\}^k$ we define the state

$$|\psi_z\rangle = \sum_{i=0}^{d} \sqrt{D(s_i)}\,|s_i, h_z(s_i)\rangle$$

- now we take the ensemble $\mathcal{E} = \{|\psi_z\rangle^{\otimes T}, 2^{-k}\}$
- let us look at the form of G_{zy} (the (z,y)-th entry of G):

$$G(z,y) = \frac{1}{2^k}\left(1 - \frac{20\varepsilon}{d}|M(z \oplus y)|\right)^T$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- for each $z \in \{0,1\}^k$ we define the state

$$|\psi_z\rangle = \sum_{i=0}^{d} \sqrt{D(s_i)}\,|s_i, h_z(s_i)\rangle$$

- now we take the ensemble $\mathcal{E} = \{|\psi_z\rangle^{\otimes T}, 2^{-k}\}$
- let us look at the form of G_{zy} (the (z,y)-th entry of G):

$$G(z,y) = \frac{1}{2^k}\left(1 - \frac{20\varepsilon}{d}|M(z \oplus y)|\right)^T$$

- note that it is only a function of $z \oplus y$.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

Theorem

for $m \geq 10$, *let* $f : \{0,1\}^m \to \mathbb{R}$ *be defined as* $f(w) = \left(1 - \beta\frac{|w|}{m}\right)^T$, *for some* $\beta \in (0,1]$ *and* $T \in \left[1, m/\left(e^3\beta\right)\right]$. *For* $k \leq m$, *let* $M \in \mathbb{F}_2^{m\times k}$ *be a matrix with rank* k. *Suppose a matrix* $A \in \mathbb{R}^{2^k \times 2^k}$ *is defined as* $A(z,y) = (f \circ M)(z \oplus y)$, *for* $z, y \in \{0,1\}^k$. *Then for all* $z \in \{0,1\}^k$ *:*

$$\sqrt{A}(z,z) \leq e^{O\left(T^2\beta^2/m + \sqrt{Tm\beta}\right)}$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

Theorem

for $m \geq 10$, *let* $f : \{0,1\}^m \to \mathbb{R}$ *be defined as* $f(w) = \left(1 - \beta \frac{|w|}{m}\right)^T$, *for some* $\beta \in (0,1]$ *and* $T \in \left[1, m/\left(e^3\beta\right)\right]$. *For* $k \leq m$, *let* $M \in \mathbb{F}_2^{m \times k}$ *be a matrix with rank* k. *Suppose a matrix* $A \in \mathbb{R}^{2^k \times 2^k}$ *is defined as* $A(z,y) = (f \circ M)(z \oplus y)$, *for* $z, y \in \{0,1\}^k$. *Then for all* $z \in \{0,1\}^k$ *:*

$$\sqrt{A}(z,z) \leq e^{O\left(T^2\beta^2/m + \sqrt{Tm\beta}\right)}$$

- using the properties we have seen for PGM:

$$P^{pgm}(\mathcal{E}) = \sum_{z \in \{0,1\}^k} \sqrt{G}(z,z)^2 \leq e^{O\left(T^2\varepsilon^2/d + \sqrt{Td\varepsilon} - d - T\varepsilon\right)}$$

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- The existance of an (ε, δ)-learner implies that $P^{opt}(\mathcal{E}) \geq 1 - \delta$. Since $P^{opt}(\mathcal{E})^2 \leq P^{pgm}(\mathcal{E})$,this quantity is $\Omega\left(1\right)$,which implies that $T \geq \Omega\left(d/\varepsilon\right)$.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- The existance of an (ε, δ)-learner implies that $P^{opt}(\mathcal{E}) \geq 1 - \delta$. Since $P^{opt}(\mathcal{E})^2 \leq P^{pgm}(\mathcal{E})$,this quantity is $\Omega\left(1\right)$,which implies that $T \geq \Omega\left(d/\varepsilon\right)$.
- so, we saw that $T \geq \Omega\left(d/\varepsilon\right)$ and overall $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log\frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$ (with $VCdim = d + 1$).

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- The existance of an (ε, δ)-learner implies that $P^{opt}(\mathcal{E}) \geq 1 - \delta$. Since $P^{opt}(\mathcal{E})^2 \leq P^{pgm}(\mathcal{E})$,this quantity is $\Omega\left(1\right)$,which implies that $T \geq \Omega\left(d/\varepsilon\right)$.
- so, we saw that $T \geq \Omega\left(d/\varepsilon\right)$ and overall $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log \frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$ (with $VCdim = d + 1$).
- which is exactly the sample complexity of classical PAC learning.

proof of the $\Omega(\frac{d}{\varepsilon})$ bound - PGM identification

- The existance of an (ε, δ)-learner implies that $P^{opt}(\mathcal{E}) \geq 1 - \delta$. Since $P^{opt}(\mathcal{E})^2 \leq P^{pgm}(\mathcal{E})$,this quantity is $\Omega\left(1\right)$,which implies that $T \geq \Omega\left(d/\varepsilon\right)$.
- so, we saw that $T \geq \Omega\left(d/\varepsilon\right)$ and overall $\Omega(\frac{d}{\varepsilon} + \frac{1}{\varepsilon} log \frac{1}{\delta})$ are necessary for an (ε, δ)-quantum PAC learner for $\mathcal{H}$ (with $VCdim = d + 1$).
- which is exactly the sample complexity of classical PAC learning.
- so we conclude that generally quantum examples **are not more powerful** than classical examples in the PAC model.

Is all hope lost?

- No. several positive results has been demonstrated in the special case of a **uniform distribution** (over the example set $\mathcal{X}$).

Is all hope lost?

- No. several positive results has been demonstrated in the special case of a **uniform distribution** (over the example set $\mathcal{X}$).
- these example utilize the properies of quantum fourier transform. among these are:

Is all hope lost?

- No. several positive results has been demonstrated in the special case of a **uniform distribution** (over the example set $\mathcal{X}$).
- these example utilize the properies of quantum fourier transform. among these are:
- if we look at time complxity, under the uniform distribution, some problems can be learned much more efficiently than we know how to do classically.

Is all hope lost?

- No. several positive results has been demonstrated in the special case of a **uniform distribution** (over the example set $\mathcal{X}$).
- these example utilize the properies of quantum fourier transform. among these are:
- if we look at time complxity, under the uniform distribution, some problems can be learned much more efficiently than we know how to do classically.
- for example: learning linear functions of the form $f(x) = ax\ mod 2$ over $\mathbb{F}_2$. by Fourier sampling we can perfectly recover a with $O(1)$ quantum sample complexity and $O(n)$ time complexity.

Is all hope lost?

- No. several positive results has been demonstrated in the special case of a **uniform distribution** (over the example set $\mathcal{X}$).
- these example utilize the properies of quantum fourier transform. among these are:
- if we look at time complxity, under the uniform distribution, some problems can be learned much more efficiently than we know how to do classically.
- for example: learning linear functions of the form $f(x) = ax \; mod 2$ over $\mathbb{F}_2$. by Fourier sampling we can perfectly recover a with $O(1)$ quantum sample complexity and $O(n)$ time complexity.
- on the other hand classical learners need $\Omega(n)$ examples to learn f.

Is all hope lost?

- No. several positive results has been demonstrated in the special case of a **uniform distribution** (over the example set $\mathcal{X}$).
- these example utilize the properies of quantum fourier transform. among these are:
- if we look at time complxity, under the uniform distribution, some problems can be learned much more efficiently than we know how to do classically.
- for example: learning linear functions of the form $f(x) = ax \ mod 2$ over $\mathbb{F}_2$. by Fourier sampling we can perfectly recover a with $O(1)$ quantum sample complexity and $O(n)$ time complexity.
- on the other hand classical learners need $\Omega(n)$ examples to learn f.
- other examples are learning k-juntas functions and DNF form.


Thank you