

A Survey of Quantum Learning Theory

Asim and Bhavishya

April 13, 2019

Table of contents

- 1 Introduction
- 2 QC & ML
 - QC Tools
 - How can QC help ML?
 - Learning Theory
- 3 Comparing Quantum & Classical
 - Framework
 - Using Quantum Data
- 4 Learnability of Quantum States
- 5 Conclusions

Introduction

- In recent years Machine Learning(ML) and Quantum computation(QC) have emerged as really successful technologies.
- A Lot of ML is based on *Heuristics*(i.e. not very mathematically rigorous) but this talk will focus on the theoretical side called “Quantum Learning Theory” (QML) and will compare the classical and quantum techniques.
- The learner will be quantum in QML, the data may be quantum

	Classical learner	Quantum learner
Classical data	Classical ML	QML
Quantum data	?	QML

Tools from QC

- Grover Search: Provides sqrt speedup while searching unstructured databases
- Fourier Sampling: Exponentially faster than classical Fourier Sampling.
- Pretty good measurement: Given an ensemble of m d -dimensional pure quantum states $\mathcal{E} = \{(p_i, |\psi_i\rangle)\}$, now we are given a state $|\psi_j\rangle$ and we have to identify j , For all POVMs $\mathcal{M} = \{M_i\}$ the average success prob is $\sum_{i=1}^m p_i \langle \psi_i | M_i | \psi_i \rangle$. Then $P^{opt} = \max_{\mathcal{M}} P_{\mathcal{M}}(\mathcal{E})$ is the optimal average success prob. Now PGM is defined as a POVM that does *reasonably well against* \mathcal{E}





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

How can QC help ML?

- **Main Idea:** Inputs to ML problems are mostly high-dimensional vectors (imagine pixels for images, frequencies for audio etc.) i.e. $v \in \mathbb{R}^d$. We can represent them using $\log(d)$ qubits,
$$\Rightarrow |v\rangle = \frac{\sum_{j=1}^d v_j |j\rangle}{||v||}.$$
- Then we can apply Quantum algorithms to learn from this efficient representation.
- **Caveat:** Quantum Machine Learning algorithms provide an exponential speedup only under certain assumptions like existence of quantum RAM, robustly invertible matrices etc. which make them difficult to implement in practice.

Learning Models

- **Concept:** A function $c : \{0,1\}^n \rightarrow \{0,1\}$.
- **Goal:** Assume $x \in \{0,1\}^n$ as a n “feature” vector, then our goal is to learn c (i.e. come up with a hypothesis) from small number of examples: $(x, c(x))$

	grey	brown	teeth	huge	$c(x)$
	1	0	1	0	1
	0	1	1	1	0
	0	1	1	0	1
	0	0	1	0	0

Output hypothesis could be: $(x_1 \text{ OR } x_2) \text{ AND } \neg x_4$

Figure: Taken from Ronald de Wolf's talk

Formal definitions

- **Concept:** some function $c : \{0,1\}^n \rightarrow \{0,1\}$
- **Concept class \mathcal{C} :** set of all concepts(eg: DNFs with small number of terms)
- Different types of learning models
 - ① Exact learning: $\forall c \in \mathcal{C}$, given access to MQ(c) oracle: w.p. $\geq 2/3$, output h s.t. $h(x) = c(x) \forall x$
 - ② PAC learning: $\forall c \in \mathcal{C}$ and distribution D , given access to PEX(c, D) oracle: w.p $\geq 1 - \delta$, outputs h s.t. $\Pr_{x \sim D} [h(x) \neq c(x)] \leq \varepsilon$
 - ③ Agnostic learning: \forall distributions D on $\{0,1\}^{n+1}$, given access to AEX(D) oracle: w.p. $\geq 1 - \delta$, outputs $h \in \mathcal{C}$ s.t. $\text{err}_D(h) \leq \text{opt}_D(\mathcal{C}) + \varepsilon$

Framework for measuring complexity of Learning

- **Sample Complexity:** Number of examples used
- **Time Complexity:** Number of time-steps used

A good learner has small time and sample complexity, Next we compare Quantum and Classical algorithms on the basis of their sample and time complexity.

VC dimension determines Sample Complexity

- It characterizes the power of hypothesis class. Eg: Linear classifiers < Polynomial classifiers.
- Measured by ability to *shatter* n points. Shatter means to classify n points in all possible labels.
- The max of n is called the VC dimension of the hypothesis class.
- Hanneke'16 showed that for every concept class \mathcal{C} there exists an (ϵ, δ) -PAC-learner using $O\left(\frac{VC}{\epsilon} + \frac{\log(1/\delta)}{\epsilon}\right)$ examples

Using Quantum Data

- We can put the classical data in a quantum superposition $\sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle$
- Under uniform D it is $\frac{1}{\sqrt{2^n}} \sum_{x \in \{0,1\}^n} |x, c(x)\rangle$
- Hadamard transformation changes this into $\sum_{s \in \{0,1\}^n} \hat{c}(s) |s\rangle$ where $\hat{c}(s) = \frac{1}{\sqrt{2^n}} \sum_x c(x) (-1)^{s \cdot x}$ are the Fourier coefficients of c . This allows us to sample s from distribution $\hat{c}(s)^2$.
- Eg: If c is linear mod 2 ($c(x) = s \cdot x$ for some s) then distribution peaks at s . Thus we can learn c from one example. (Think Simon's algorithm)
- But in general PAC learning where D can be any distribution quantum examples are not significantly better than classical examples [Arunachalam & dW'17]

Proof sketch of lower bound

- Let $S = \{s_0, s_1, \dots, s_d\}$ be shattered by \mathcal{C} . Here d is VC-dimension. Define distribution D s.t. $Pr[s_0] = 1 - 8\varepsilon$ and $Pr[\{s_1, \dots, s_d\}] = 8\varepsilon/d$
- Assume ε -error learner takes T examples, and produces hypothesis h that agrees with c more than $7/8^{th}$ of time. This reduces to an approximate state identification problem.
- Quantum learner cannot do much better than PGM And analysis of PGM lower bounds, $T \geq \Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$
- Note which is same as classical case

Similar results for Agnostic Learning

- Examples from unknown distribution D on (x, ℓ) . Predict ℓ from x , this allows us to model the case when target concept might not even exist.
- Best concept from \mathcal{C} has error $\text{opt} = \min_{c \in \mathcal{C}} \Pr_{(x, \ell) \sim D} [c(x) \neq \ell]$
- Therefor find $h \in \mathcal{C}$ with error $\leq \text{opt} + \varepsilon$
- Classical Sample complexity: $T = \Theta\left(\frac{VC}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$
- By methods similar to PAC lower bound it can be shown that Qunatum Sample complexity is same as the classical case.

Quantum improvements in time complexity

- Kearns & Vazirani'94 gave a concept class that is not efficiently PAC-learnable if factoring is hard
- But factoring is easy using Shor's algorithm. Therefore these classes can be learned efficiently[Servedio & Gortler'04]
- Servedio & Gortler'04: If classical one-way functions exist, then $\exists \mathcal{C}$ that is efficiently exactly learnable from membership queries by quantum but not by classical computers.

Proof Idea: Use pseudo-random function to generate instances of Simon's problem (special 2-to-1 functions). Simon's algorithm can solve this efficiently, but classical learner would have to distinguish random from pseudo-random

Learnability of Quantum States

- Why restrict ourselves to just learning classical objects?
- Can we learn a classical description of a quantum state which is close to it in some sense(full state tomography)?
- If the last is hard/expensive, Can we learn the behaviour of a quantum state rather than learning the whole quantum state(shadow state tomography)?
- Can we learn a measurement operator, given its behaviour on a set of quantum states?

Pretty Good Tomography: Preliminaries

A mixed State, ρ , over n qubits is described by a $2^n \times 2^n$ Hermitian.

Given a two-outcome measurement, E and $I - E$, where E is positive semi-definite, probability for the measurement to yield first outcome is $Tr(E\rho)$

γ -fat-shattering dimension: An extension of VC dimension from Boolean to real valued functions. For some set ε , let \mathcal{C} be a class of function $f : \varepsilon \rightarrow [0, 1]$. We say that the set $S = \{E_1, \dots, E_d\} \subseteq \varepsilon$ is γ -fat-shattered by \mathcal{C} if there exist $\alpha_1, \dots, \alpha_d \in [0, 1]$ such that for all $Z \subseteq [d]$ there is an $f \in \mathcal{C}$ satisfying:

- if $i \in Z$ then $f(E_i) \geq \alpha_i + \gamma$
- if $i \notin Z$ then $f(E_i) \leq \alpha_i - \gamma$

The γ -fat-shattering dimension of \mathcal{C} is the size of the largest S that is shattered by \mathcal{C} .

Pretty Good Tomography: Bounds

Theorem(Aaronson,2006): For every $\delta, \varepsilon, \gamma$, there exists a learner with the following property: For every distribution D on the set of two-outcome measurements, given $T = n \cdot \text{poly}(1/\varepsilon, 1/\gamma, \log(1/\delta))$ measurement results $(E_1, b_1), \dots, (E_T, b_T)$ where each E_i is drawn i.i.d. from D and b_i is a bit with $\Pr[b_i = 1] = \text{Tr}(E_i \rho)$, with probability $\geq 1 - \delta$ the learner produces the classical dewscription of a state σ such that

$$\Pr_{E \sim D} [| \text{Tr}(E\sigma) - \text{Tr}(E\rho) | > \gamma] \leq \varepsilon$$

Note the similarity of this theorem to that of PAC formulation of learning problem. However the "approximately correct" motivation from original PAC is now defined by two parameters $\varepsilon\gamma$, rather than only by single parameter. Also note that the Theorem is about Sample Complexity **NOT** about time complexity.

Pretty Good Tomography: Proof Sketch

γ -fat-shattering dimension: An extension of VC dimension from Boolean to real valued functions. For some set ε , let \mathcal{C} be a class of function $f : \varepsilon \rightarrow [0, 1]$. We say that the set $S = \{E_1, \dots, E_d\} \subseteq \varepsilon$ is γ -fat-shattered by \mathcal{C} if there exist $\alpha_1, \dots, \alpha_d \in [0, 1]$ such that for all $Z \subseteq [d]$ there is an $f \in \mathcal{C}$ satisfying:

- if $i \in Z$ then $f(E_i) \geq \alpha_i + \gamma$
- if $i \notin Z$ then $f(E_i) \leq \alpha_i - \gamma$

The γ -fat-shattering dimension of \mathcal{C} is the size of the largest S that is shattered by \mathcal{C} .

For the application of fat-shattering dimension to learning quantum states, consider

$$\varepsilon \text{ as set of all } n\text{-qubit measurement operators}$$

$$\mathcal{C} = \{f : \varepsilon \rightarrow [0, 1] \mid \exists \text{ } n \text{ qubit } \rho \text{ s.t. } \forall E \in \varepsilon, f(E) = \text{Tr}(E\rho)\}$$

- This indicates that for each string $z \in \{0, 1\}^d$, there exists a n qubit state ρ_z from which z_i can be recovered using measurement E_i .
- These states are called quantum random access codes as they encode the string z .
- Using known bounds on such code **Aaronson 2007** shows that $d = O(n \gamma^2)$.
- This is plugged into the results by **Anthony and Bartlett(2000)** and **Bartlett and Long(1998)** to obtain the required bounds.

Shadow State Tomography

- In shadow state tomography, instead of outputting a complete description of state ρ we need to output $\text{Tr}(E_i \rho)$ upto an additive error ϵ . E_i is from the set of known 2 outcome measurements E_1, \dots, E_m . Goal is to minimize the number of copies, K of state ρ required to do so.
- **Aaronson 2017** showed that we need $K = \text{poly}(n, \log(m), 1/\epsilon)$ copies of the state to do so. This was an exponential improvement in both n and m from the previous bound.

Learning Unknown Quantum Measurement

- In this problem, we need a bound on how many states are sufficient to learn an unknown Quantum Measurement.
- The bound turns out to be exponential in the number of qubits the measurement operator acts on.

Conclusions

- We can get quadratic speedup for some ML problems and exponential speedup under *very* strong assumptions.
- No improvements in Sample Complexity.
- Time complexity is exponentially improved for some concept classes like factoring.
- Various pragmatic issues still unsolved like how to put big classical data in superposition, quantum memory, appropriate dataset etc.
- Pretty good tomography can be done in $O(n)$ copies of state, where n is the dimensionality of state being learnt. Shadow state tomography can be done in $O(n \log(m))$ copies of state. To learn unknown measurement operators we need $\exp(n)$ different states.