Optimal Quantum Sample Complexity of Learning Algorithms

Srinivasan Arunachalam
(Joint work with Ronald de Wolf)
Machine learning

Classical machine learning

Grand goal: enable AI systems to improve themselves
Practical goal: learn “something” from given data
Recent success: deep learning is extremely good at image recognition, natural language processing, even the game of Go
Why the recent interest? Flood of available data, increasing computational power, growing progress in algorithms

Quantum machine learning
What can quantum computing do for machine learning?
The learner will be quantum, the data may be quantum
Some examples are known of reduction in time complexity:
- clustering (Aּımeur et al. ’06)
- principal component analysis (Lloyd et al. ’13)
- perceptron learning (Wiebe et al. ’16)
- recommendation systems (Kerenidis & Prakash ’16)
Classical machine learning

- **Grand goal**: enable AI systems to improve themselves
Machine learning

Classical machine learning

- **Grand goal**: enable AI systems to improve themselves
- **Practical goal**: learn “something” from given data

Recent success: deep learning is extremely good at image recognition, natural language processing, even the game of Go

Why the recent interest? Flood of available data, increasing computational power, growing progress in algorithms

Quantum machine learning

What can quantum computing do for machine learning?

The learner will be quantum, the data may be quantum

Some examples are known of reduction in time complexity:

- clustering (ÁEmeur et al. ’06)
- principal component analysis (Lloyd et al. ’13)
- perceptron learning (Wiebe et al. ’16)
- recommendation systems (Kerenidis & Prakash ’16)
Machine learning

Classical machine learning

- **Grand goal**: enable AI systems to improve themselves
- **Practical goal**: learn “something” from given data
- **Recent success**: deep learning is extremely good at image recognition, natural language processing, even the game of Go

Quantum machine learning

What can quantum computing do for machine learning?

The learner will be quantum, the data may be quantum.

Some examples are known of reduction in time complexity:

- clustering (A. ¨ ımeur et al. ’06)
- principal component analysis (Lloyd et al. ’13)
- perceptron learning (Wiebe et al. ’16)
- recommendation systems (Kerenidis & Prakash ’16)
Machine learning

Classical machine learning
- **Grand goal**: enable AI systems to improve themselves
- **Practical goal**: learn “something” from given data
- **Recent success**: deep learning is extremely good at image recognition, natural language processing, even the game of Go
- **Why the recent interest?** Flood of available data, increasing computational power, growing progress in algorithms
### Classical machine learning

- **Grand goal**: enable AI systems to improve themselves
- **Practical goal**: learn “something” from given data
- **Recent success**: deep learning is extremely good at image recognition, natural language processing, even the game of Go
- **Why the recent interest?** Flood of available data, increasing computational power, growing progress in algorithms

### Quantum machine learning

- **What can quantum computing do for machine learning?**
Classical machine learning

- **Grand goal**: enable AI systems to improve themselves
- **Practical goal**: learn “something” from given data
- **Recent success**: deep learning is extremely good at image recognition, natural language processing, even the game of Go
- **Why the recent interest?** Flood of available data, increasing computational power, growing progress in algorithms

Quantum machine learning

- **What can quantum computing do for machine learning?**
- **The learner will be quantum, the data may be quantum**
Machine learning

Classical machine learning

- **Grand goal**: enable AI systems to improve themselves
- **Practical goal**: learn “something” from given data
- **Recent success**: deep learning is extremely good at image recognition, natural language processing, even the game of Go
- **Why the recent interest?** Flood of available data, increasing computational power, growing progress in algorithms

Quantum machine learning

- What can **quantum computing** do for machine learning?
- The learner will be quantum, the data may be quantum
- Some examples are known of reduction in time complexity:
  - clustering (Aïmeur et al. ’06)
  - principal component analysis (Lloyd et al. ’13)
  - perceptron learning (Wiebe et al. ’16)
  - recommendation systems (Kerenidis & Prakash ’16)
Probably Approximately Correct (PAC) learning

Basic definitions

Concept class $C$: collection of Boolean functions on $n$ bits (Known)

Target concept $c$: some function $c \in C$ (Unknown)

Distribution $D$: $\{0, 1\}^n \to [0, 1]$ (Unknown)

Labeled example for $c \in C$: $(x, c(x))$ where $x \sim D$
Basic definitions

- **Concept class \( C \):** collection of Boolean functions on \( n \) bits (Known)
### Basic definitions

- **Concept class \( C \):** collection of Boolean functions on \( n \) bits (Known)
- **Target concept \( c \):** some function \( c \in C \) (Unknown)
Basic definitions

- **Concept class** $C$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in C$ (Unknown)
- **Distribution** $D: \{0, 1\}^n \rightarrow [0, 1]$ (Unknown)
**Basic definitions**

- **Concept class \( \mathcal{C} \):** collection of Boolean functions on \( n \) bits (Known)
- **Target concept \( c \):** some function \( c \in \mathcal{C} \) (Unknown)
- **Distribution \( D \):** \( \{0, 1\}^n \rightarrow [0, 1] \) (Unknown)
- **Labeled example for \( c \in \mathcal{C} \):** \((x, c(x))\) where \( x \sim D \)
### Basic definitions

- **Concept class \( C \):** collection of Boolean functions on \( n \) bits (Known)
- **Target concept \( c \):** some function \( c \in C \). (Unknown)
- **Distribution \( D \):** \( \{0, 1\}^n \to [0, 1] \). (Unknown)
- **Labeled example for \( c \in C \):** \((x, c(x))\) where \( x \sim D \)

\[
\begin{array}{c}
C \\
\downarrow \\
C \\
\text{target concept}
\end{array}
\]
Probably Approximately Correct (PAC) learning

Basic definitions

- **Concept class** $\mathcal{C}$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in \mathcal{C}$. (Unknown)
- **Distribution** $D$: $\{0, 1\}^n \rightarrow [0, 1]$. (Unknown)
- **Labeled example** for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$
Probably Approximately Correct (PAC) learning

### Basic definitions

- **Concept class** $\mathcal{C}$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in \mathcal{C}$. (Unknown)
- **Distribution** $D : \{0, 1\}^n \rightarrow [0, 1]$. (Unknown)
- **Labeled example for** $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$
## Basic definitions

- **Concept class** $C$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in C$. (Unknown)
- **Distribution** $D: \{0, 1\}^n \rightarrow [0, 1]$. (Unknown)
- **Labeled example** for $c \in C$: $(x, c(x))$ where $x \sim D$

### Diagram

- $C$
- $\downarrow$
- $C$
- $\text{target concept}$

<table>
<thead>
<tr>
<th>$x_1 \sim D$</th>
<th>$\rightarrow$</th>
<th>$(x_1, c(x_1))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2 \sim D$</td>
<td>$\rightarrow$</td>
<td>$(x_2, c(x_2))$</td>
</tr>
<tr>
<td>\vdots</td>
<td></td>
<td>\vdots</td>
</tr>
<tr>
<td>$x_T \sim D$</td>
<td>$\rightarrow$</td>
<td>$(x_T, c(x_T))$</td>
</tr>
</tbody>
</table>

**Output:** Hypothesis $h$

- $h$ is probably approximately correct!
Probably Approximately Correct (PAC) learning

Basic definitions

- **Concept class** $\mathcal{C}$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in \mathcal{C}$. (Unknown)
- **Distribution** $D: \{0, 1\}^n \rightarrow [0, 1]$. (Unknown)
- **Labeled example** for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$.

Formally: A theory of the learnable (L.G. Valiant’84)
Basic definitions

- **Concept class** $\mathcal{C}$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in \mathcal{C}$. (Unknown)
- **Distribution** $D : \{0, 1\}^n \to [0, 1]$. (Unknown)
- **Labeled example** for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$.

Formally: A theory of the learnable (L.G. Valiant’84)

- Using i.i.d. labeled examples, learner for $\mathcal{C}$ should output hypothesis $h$ that is *Probably Approximately Correct*
Basic definitions

- **Concept class** $\mathcal{C}$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in \mathcal{C}$. (Unknown)
- **Distribution** $D : \{0, 1\}^n \rightarrow [0, 1]$. (Unknown)
- **Labeled example** for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$.

Formally: A theory of the learnable (L.G. Valiant’84)

- Using i.i.d. labeled examples, learner for $\mathcal{C}$ should output hypothesis $h$ that is *Probably Approximately Correct*
- **Error of** $h$ w.r.t. target $c$: $err_D(c, h) = \Pr_{x \sim D}[c(x) \neq h(x)]$
Probably Approximately Correct (PAC) learning

Basic definitions

- **Concept class** $\mathcal{C}$: collection of Boolean functions on $n$ bits (Known)
- **Target concept** $c$: some function $c \in \mathcal{C}$. (Unknown)
- **Distribution** $D : \{0, 1\}^n \rightarrow [0, 1]$. (Unknown)
- **Labeled example** for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$.

Formally: A theory of the learnable (L.G. Valiant’84)

- Using i.i.d. labeled examples, learner for $\mathcal{C}$ should output hypothesis $h$ that is *Probably Approximately Correct*
- **Error of $h$ w.r.t. target $c$**: $\text{err}_D(c, h) = \Pr_{x \sim D}[c(x) \neq h(x)]$
- An algorithm $(\varepsilon, \delta)$-PAC-learns $\mathcal{C}$ if:

$$\forall c \in \mathcal{C} \ \forall D : \ Pr[\underbrace{\text{err}_D(c, h) \leq \varepsilon}_{\text{Approximately Correct}}] \geq 1 - \delta$$
Complexity of learning

Recap

- Concept: some function $c : \{0, 1\}^n \rightarrow \{0, 1\}$
  - Concept class $C$: set of concepts
- An algorithm $(\varepsilon, \delta)$-PAC-learns $C$ if:

$$\forall c \in C \quad \forall D : \quad \Pr[\text{err}_D(c, h) \leq \varepsilon] \geq 1 - \delta$$

  - Approximately Correct
  - Probably

- How to measure the efficiency of the learning algorithm?

Sample complexity: number of labeled examples used by learner

Time complexity: number of time-steps used by learner

This talk: focus on sample complexity

No need for complexity-theoretic assumptions
No need to worry about the format of hypothesis $h$
Recap

- Concept: some function $c : \{0, 1\}^n \rightarrow \{0, 1\}$
  - Concept class $C$: set of concepts
- An algorithm $(\varepsilon, \delta)$-PAC-learns $C$ if:

  \[
  \forall c \in C \quad \forall D : \quad \Pr[\ err_D(c, h) \leq \varepsilon \ ] \geq 1 - \delta
  \]

  Approximately Correct
  Probably

- How to measure the efficiency of the learning algorithm?
  - Sample complexity: number of labeled examples used by learner
Complexity of learning

Recap

- Concept: some function \( c : \{0, 1\}^n \rightarrow \{0, 1\} \)
  - Concept class \( C \): set of concepts
- An algorithm \((\varepsilon, \delta)\)-PAC-learns \( C \) if:

\[
\forall c \in C \quad \forall D : \quad \Pr[\text{err}_D(c, h) \leq \varepsilon] \geq 1 - \delta
\]

How to measure the efficiency of the learning algorithm?

- **Sample complexity**: number of labeled examples used by learner
- **Time complexity**: number of time-steps used by learner
Complexity of learning

Recap

- Concept: some function $c : \{0, 1\}^n \rightarrow \{0, 1\}$
  - Concept class $C$: set of concepts
- An algorithm $(\varepsilon, \delta)$-PAC-learns $C$ if:
  \[ \forall c \in C \quad \forall D : \quad \Pr\left[ err_D(c, h) \leq \varepsilon \right] \geq 1 - \delta \]

- How to measure the efficiency of the learning algorithm?
  - Sample complexity: number of labeled examples used by learner
  - Time complexity: number of time-steps used by learner
- This talk: focus on sample complexity
Complexity of learning

Recap

- Concept: some function $c : \{0, 1\}^n \rightarrow \{0, 1\}$
  - Concept class $C$: set of concepts
- An algorithm $(\varepsilon, \delta)$-PAC-learns $C$ if:
  $$\forall c \in C \quad \forall D : \quad \Pr[\text{err}_D(c, h) \leq \varepsilon] \geq 1 - \delta$$

- Approximately Correct
- Probably

How to measure the efficiency of the learning algorithm?
- Sample complexity: number of labeled examples used by learner
- Time complexity: number of time-steps used by learner

This talk: focus on sample complexity
- No need for complexity-theoretic assumptions
Complexity of learning

Recap

- Concept: some function $c : \{0, 1\}^n \rightarrow \{0, 1\}$
  - Concept class $C$: set of concepts
- An algorithm $(\varepsilon, \delta)$-PAC-learns $C$ if:

\[
\forall c \in C \forall D : \Pr[\text{err}_D(c, h) \leq \varepsilon] \geq 1 - \delta
\]

Approximately Correct

How to measure the efficiency of the learning algorithm?

- **Sample complexity**: number of labeled examples used by learner
- **Time complexity**: number of time-steps used by learner

**This talk**: focus on *sample complexity*

- No need for complexity-theoretic assumptions
- No need to worry about the format of hypothesis $h$
VC dimension of $C \subseteq \{c : \{0, 1\}^n \rightarrow \{0, 1\}\}$
Vapnik and Chervonenkis (VC) dimension

VC dimension of $C \subseteq \{ c : \{0, 1\}^n \rightarrow \{0, 1\} \}$

Let $M$ be the $|C| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of concept $c : \{0, 1\}^n \rightarrow \{0, 1\}$
Let $M$ be the $|C| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of concept $c : \{0, 1\}^n \rightarrow \{0, 1\}$

VC-dim$(C)$: largest $d$ s.t. the $|C| \times d$ rectangle in $M$ contains $\{0, 1\}^d$
Vapnik and Chervonenkis (VC) dimension

VC dimension of $C \subseteq \{ c : \{0, 1\}^n \rightarrow \{0, 1\} \}$

Let $M$ be the $|C| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of concept $c : \{0, 1\}^n \rightarrow \{0, 1\}$

$VC$-$dim(C)$: largest $d$ s.t. the $|C| \times d$ rectangle in $M$ contains $\{0, 1\}^d$

These $d$ column indices are shattered by $C$
VC dimension of $\mathcal{C} \subseteq \{c : \{0, 1\}^n \rightarrow \{0, 1\}\}$

$M$ is the $|\mathcal{C}| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of $c$

$\text{VC-dim}(\mathcal{C})$: largest $d$ s.t. the $|\mathcal{C}| \times d$ rectangle in $M$ contains $\{0, 1\}^d$

These $d$ column indices are shattered by $\mathcal{C}$

Table: $\text{VC-dim}(\mathcal{C}) = 2$

<table>
<thead>
<tr>
<th>Concepts $c_i$</th>
<th>Truth table</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>0 1 0 1</td>
</tr>
<tr>
<td>$c_2$</td>
<td>0 1 1 0</td>
</tr>
<tr>
<td>$c_3$</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>$c_4$</td>
<td>1 0 1 0</td>
</tr>
<tr>
<td>$c_5$</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>$c_6$</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>$c_7$</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>$c_8$</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>$c_9$</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>
Vapnik and Chervonenkis (VC) dimension

VC dimension of $\mathcal{C} \subseteq \{c : \{0, 1\}^n \rightarrow \{0, 1\}\}$

$M$ is the $|\mathcal{C}| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of $c$

$\text{VC-dim}(\mathcal{C})$: largest $d$ s.t. the $|\mathcal{C}| \times d$ rectangle in $M$ contains $\{0, 1\}^d$

These $d$ column indices are shattered by $\mathcal{C}$

### Table: $\text{VC-dim}(\mathcal{C}) = 2$

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Truth table</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>0 1 0 1</td>
</tr>
<tr>
<td>$c_2$</td>
<td>0 1 1 0</td>
</tr>
<tr>
<td>$c_3$</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>$c_4$</td>
<td>1 0 1 0</td>
</tr>
<tr>
<td>$c_5$</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>$c_6$</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>$c_7$</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>$c_8$</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>$c_9$</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

### Table: $\text{VC-dim}(\mathcal{C}) = 3$

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Truth table</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>0 1 1 1 0</td>
</tr>
<tr>
<td>$c_2$</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>$c_3$</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>$c_4$</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>$c_5$</td>
<td>1 0 1 0</td>
</tr>
<tr>
<td>$c_6$</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>$c_7$</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>$c_8$</td>
<td>0 1 0 1</td>
</tr>
<tr>
<td>$c_9$</td>
<td>0 1 0 0</td>
</tr>
</tbody>
</table>
VC dimension of $C$

$M$ is the $|C| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of $c$

$\text{VC-dim}(C)$: largest $d$ s.t. the $|C| \times d$ rectangle in $M$ contains $\{0, 1\}^d$

These $d$ column indices are shattered by $C$

Fundamental theorem of PAC learning
VC dimension characterizes PAC sample complexity

**VC dimension of $\mathcal{C}$**

$M$ is the $|\mathcal{C}| \times 2^n$ Boolean matrix whose $c$-th row is the truth table of $c$

VC-dim($\mathcal{C}$): largest $d$ s.t. the $|\mathcal{C}| \times d$ rectangle in $M$ contains $\{0, 1\}^d$

These $d$ column indices are shattered by $\mathcal{C}$

**Fundamental theorem of PAC learning**

Suppose VC-dim($\mathcal{C}$) = $d$
VC dimension characterizes PAC sample complexity

**VC dimension of \( C \)**

\( M \) is the \(|C| \times 2^n\) Boolean matrix whose \( c \)-th row is the truth table of \( c \)

\( \text{VC-dim}(C) \): largest \( d \) s.t. the \(|C| \times d\) rectangle in \( M \) contains \( \{0, 1\}^d \)

These \( d \) column indices are shattered by \( C \)

**Fundamental theorem of PAC learning**

Suppose \( \text{VC-dim}(C) = d \)

- Blumer-Ehrenfeucht-Haussler-Warmuth’86:
  every \((\varepsilon, \delta)\)-PAC learner for \( C \) needs \( \Omega \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right) \) examples
VC dimension characterizes PAC sample complexity

<table>
<thead>
<tr>
<th>VC dimension of $\mathcal{C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$ is the $</td>
</tr>
<tr>
<td>VC-dim($\mathcal{C}$): largest $d$ s.t. the $</td>
</tr>
<tr>
<td>These $d$ column indices are shattered by $\mathcal{C}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fundamental theorem of PAC learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suppose VC-dim($\mathcal{C}$) = $d$</td>
</tr>
<tr>
<td>- Blumer-Ehrenfeucht-Haussler-Warmuth’86: every $(\varepsilon, \delta)$-PAC learner for $\mathcal{C}$ needs $\Omega \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ examples</td>
</tr>
<tr>
<td>- Hanneke’16: there exists an $(\varepsilon, \delta)$-PAC learner for $\mathcal{C}$ using $O \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ examples</td>
</tr>
</tbody>
</table>
Quantum PAC learning

(Bshouty-Jackson’95): Quantum generalization of classical PAC
Quantum PAC learning

- (Bshouty-Jackson’95): **Quantum generalization** of classical PAC
- **Learner is quantum**:

\[
\sum_{x \in \{0, 1\}^n} \sqrt{D(x)} |x, c(x)\rangle
\]

Measuring this state gives \((x, c(x))\) with probability \(D(x)\), so quantum examples are at least as powerful as classical
Quantum PAC learning

- (Bshouty-Jackson’95): Quantum generalization of classical PAC
- Learner is quantum:
  \[
  \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle
  \]
  Measuring this state gives \((x, c(x))\) with probability \(D(x)\), so quantum examples are at least as powerful as classical.
Quantum PAC learning

- (Bshouty-Jackson’95): Quantum generalization of classical PAC
- Learner is quantum:

\[
\sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle
\]

Data is quantum: Quantum example is a superposition

Measuring this state gives \((x, c(x))\) with probability \(D(x)\),
Quantum PAC learning

- (Bshouty-Jackson’95): Quantum generalization of classical PAC
- Learner is quantum:

\[ \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle \]

Data is quantum: Quantum example is a superposition

Measuring this state gives \((x, c(x))\) with probability \(D(x)\), so quantum examples are at least as powerful as classical
Classical vs. Quantum PAC learning algorithm!

Question

Can quantum sample complexity be significantly smaller than classical?
Quantum PAC learning

Quantum Data

- Quantum example: $|E_{c,D}\rangle = \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle$
- Quantum examples are at least as powerful as classical examples

Quantum is indeed more powerful for learning! (for uniform distribution)
Quantum Data

- Quantum example: $|E_{c,D}\rangle = \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle$
- Quantum examples are at least as powerful as classical examples

Quantum is indeed more powerful for learning! (for uniform distribution)

- Sample complexity: Learning class of linear functions
Quantum Data

- Quantum example: \( |E_{c,D}⟩ = \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)⟩ \)
- Quantum examples are at least as powerful as classical examples

Quantum is indeed more powerful for learning! (for uniform distribution)

- Sample complexity: Learning class of linear functions
  - Classical: \( \Omega(n) \) classical examples needed
  - Quantum: \( O(1) \) quantum examples suffice (Bernstein-Vazirani’93)
Quantum PAC learning

Quantum Data

- **Quantum example:** \( |E_{c,D} \rangle = \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x) \rangle \)
- Quantum examples are at least as powerful as classical examples

Quantum is indeed more powerful for learning! (for *uniform distribution*)

- **Sample complexity:** Learning class of linear functions
  - Classical: \( \Omega(n) \) classical examples needed
  - Quantum: \( O(1) \) quantum examples suffice (Bernstein-Vazirani’93)
- **Time complexity:** Learning DNFs
Quantum PAC learning

Quantum Data

- Quantum example: \(|E_{c,D}\rangle = \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)\rangle\)
- Quantum examples are at least as powerful as classical examples

Quantum is indeed more powerful for learning! (for uniform distribution)

- Sample complexity: Learning class of linear functions
  Classical: \(\Omega(n)\) classical examples needed
  Quantum: \(O(1)\) quantum examples suffice (Bernstein-Vazirani’93)
- Time complexity: Learning DNFs
  Classical: Best known upper bound is quasi-poly. time (Verbeugt’90)
  Quantum: Polynomial-time (Bshouty-Jackson’95)
Quantum PAC learning

Quantum Data

- Quantum example: $|E_{c,D}⟩ = \sum_{x \in \{0,1\}^n} \sqrt{D(x)} |x, c(x)⟩$
- Quantum examples are at least as powerful as classical examples

Quantum is indeed more powerful for learning! (for a fixed distribution)

- Learning class of linear functions under uniform $D$:
  - Classical: $\Omega(n)$ classical examples needed
  - Quantum: $O(1)$ quantum examples suffice (Bernstein-Vazirani'93)
- Learning DNF under uniform $D$:
  - Classical: Best known upper bound is quasi-poly. time (Verbeugt’90)
  - Quantum: Polynomial-time (Bshouty-Jackson’95)

But in the PAC model, learner has to succeed for all $D$!
Quantum sample complexity

Quantum upper bound

Classical upper bound $O \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ carries over to quantum
Quantum sample complexity

Quantum upper bound

Classical upper bound $O\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ carries over to quantum

Best known quantum lower bounds

Atici & Servedio’04: lower bound $\Omega\left(\frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right)$
Quantum sample complexity

Quantum upper bound

Classical upper bound \( O \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right) \) carries over to quantum

Best known quantum lower bounds

Atici & Servedio’04: lower bound \( \Omega \left( \frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon} \right) \)

Zhang’10: improved first term to \( \frac{d^{1-\eta}}{\varepsilon} \) for all \( \eta > 0 \)
Quantum sample complexity = Classical sample complexity

Quantum upper bound

Classical upper bound $O \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ carries over to quantum

Best known quantum lower bounds

Atici & Servedio’04: lower bound $\Omega \left( \frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon} \right)$

Zhang’10 improved first term to $\frac{d^{1-\eta}}{\varepsilon}$ for all $\eta > 0$

Our result: Tight lower bound

We show: $\Omega \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ quantum examples are necessary
Quantum sample complexity = Classical sample complexity

Quantum upper bound

Classical upper bound $O\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ carries over to quantum

Best known quantum lower bounds

Atici & Servedio'04: lower bound $\Omega\left(\frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right)$

Zhang'10 improved first term to $\frac{d^{1-\eta}}{\varepsilon}$ for all $\eta > 0$

Our result: Tight lower bound

We show: $\Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ quantum examples are necessary

Two proof approaches

- Information theory: conceptually simple, nearly-tight bounds
### Quantum upper bound

Classical upper bound $O \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ carries over to quantum.

### Best known quantum lower bounds

Atici & Servedio'04: lower bound $\Omega \left( \frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon} \right)$

Zhang’10 improved first term to $\frac{d^{1-\eta}}{\varepsilon}$ for all $\eta > 0$

### Our result: Tight lower bound

We show: $\Omega \left( \frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \right)$ quantum examples are necessary

Two proof approaches

- Information theory: conceptually simple, nearly-tight bounds
- Optimal measurement: tight bounds, some messy calculations
First, we consider the problem of probably **exactly** learning: quantum learner should *identify* the concept.
First, we consider the problem of probably **exactly** learning: quantum learner should *identify* the concept.

Here, quantum learner is given one out of $|\mathcal{C}|$ quantum states. Identify the target concept using copies of the quantum state.
First, we consider the problem of probably exactly learning: quantun learner should *identify* the concept.

Here, quantum learner is given one out of $|C|$ quantum states. **Identify the target concept** using copies of the quantum state.

Quantum state identification has been well-studied.
First, we consider the problem of probably exactly learning: quantum learner should identify the concept.

Here, quantum learner is given one out of $|C|$ quantum states. Identify the target concept using copies of the quantum state.

Quantum state identification has been well-studied.

We’ll get to probably approximately learning soon!
State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$
State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal:** identify $z$
Proof sketch: Quantum sample complexity \( T \geq \frac{\text{VC-dim}(C)}{\varepsilon} \)

State identification: Ensemble \( \mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]} \)

- Given state \( |\psi_z\rangle \in \mathcal{E} \) with prob \( p_z \). **Goal:** identify \( z \)
- Optimal measurement could be quite complicated,
State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. Goal: identify $z$
- Optimal measurement could be quite complicated, but we can always use the Pretty Good Measurement
- Crucial property: if $P_{opt}$ is the optimal success probability, then

Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$
State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm}$
State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill’02)

Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\epsilon$
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

**State identification: Ensemble** $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill’02)

**How does learning relate to identification?**
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill’02)

How does learning relate to identification?

- Quantum PAC: **Given** $|\psi_c\rangle = |E_c,D\rangle^{\otimes T}$, learn $c$ *approximately*
State identification: Ensemble $E = \{(pz, |\psi_z\rangle)\}_{z \in [m]}

- Given state $|\psi_z\rangle \in E$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill’02)

How does learning relate to identification?

- Quantum PAC: **Given** $|\psi_c\rangle = |E_{c,D}\rangle^\otimes T$, learn $c$ **approximately**
- Let VC-dim$(C) = d$. Suppose $\{s_0, \ldots, s_d\}$ is shattered by $C$. 

Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

**State identification**: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then
  $$P_{opt} \geq P_{pgm} \geq P^2_{opt} \quad \text{(Barnum-Knill’02)}$$

**How does learning relate to identification?**

- **Quantum PAC**: Given $|\psi_c\rangle = |E_c,D\rangle \otimes^T$, learn $c$ **approximately**
- Let $\text{VC-dim}(C) = d$. Suppose $\{s_0, \ldots, s_d\}$ is shattered by $C$. Fix $D(s_0) = 1 - \varepsilon$, $D(s_i) = \varepsilon/d$ on $\{s_1, \ldots, s_d\}$
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then
  \[ P_{opt} \geq P_{pgm} \geq P_{opt}^2 \] (Barnum-Knill’02)

How does learning relate to identification?

- **Quantum PAC**: Given $|\psi_c\rangle = |E_{c,D}\rangle^\otimes T$, learn $c$ **approximately**
- Let $\text{VC-dim}(C) = d$. Suppose $\{s_0, \ldots, s_d\}$ is shattered by $C$. Fix
  \[ D(s_0) = 1 - \varepsilon, \quad D(s_i) = \varepsilon/d \] on $\{s_1, \ldots, s_d\}$
- Let $k = \Omega(d)$ and $E : \{0,1\}^k \to \{0,1\}^d$ be an error-correcting code
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. Goal: identify $z$
- Optimal measurement could be quite complicated, but we can always use the Pretty Good Measurement
- Crucial property: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill’02)

How does learning relate to identification?

- Quantum PAC: Given $|\psi_c\rangle = |E_{c,D}\rangle^\otimes T$, learn $c$ approximately
- Let VC-dim$(C) = d$. Suppose $\{s_0, \ldots, s_d\}$ is shattered by $C$. Fix $D(s_0) = 1 - \varepsilon$, $D(s_i) = \varepsilon/d$ on $\{s_1, \ldots, s_d\}$
- Let $k = \Omega(d)$ and $E : \{0,1\}^k \rightarrow \{0,1\}^d$ be an error-correcting code
- Pick $2^k$ codeword concepts $\{c_z\}_{z \in \{0,1\}^k} \subseteq C$:
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal**: identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- **Crucial property**: if $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill’02)

How does learning relate to identification?

- Quantum PAC: Given $|\psi_c\rangle = |E_{c,D}\rangle^\otimes T$, learn $c$ approximately
- Let VC-dim($C$) = $d$. Suppose $\{s_0, \ldots, s_d\}$ is shattered by $C$. Fix $D(s_0) = 1 - \varepsilon$, $D(s_i) = \varepsilon/d$ on $\{s_1, \ldots, s_d\}$
- Let $k = \Omega(d)$ and $E : \{0, 1\}^k \rightarrow \{0, 1\}^d$ be an error-correcting code
- Pick $2^k$ codeword concepts $\{c_z\}_{z \in \{0,1\}^k} \subseteq C$:
  
  $c_z(s_0) = 0$, $c_z(s_i) = E(z)_i \quad \forall \ i \in b$
Pick concepts \( \{ c_z \} \subseteq C \): \( c_z(s_0) = 0 \), \( c_z(s_i) = E(z)_i \ \forall \ i \)

Suppose \( VC(C) = d + 1 \) and \( \{ s_0, \ldots, s_d \} \) is shattered by \( C \), i.e.,
\[
|C| \times (d + 1) \text{ rectangle of } \{ s_0, \ldots, s_d \} \text{ contains } \{0, 1\}^{d+1}
\]
Pick concepts \( \{ c_z \} \subseteq C: c_z(s_0) = 0, c_z(s_i) = E(z)_i \ \forall \ i \)

Suppose \( VC(C) = d + 1 \) and \( \{ s_0, \ldots, s_d \} \) is shattered by \( C \), i.e., 
\(|C| \times (d + 1)\) rectangle of \( \{ s_0, \ldots, s_d \} \) contains \( \{0, 1\}^{d+1} \)

<table>
<thead>
<tr>
<th>Concepts ( c \in C )</th>
<th>Truth table</th>
</tr>
</thead>
</table>
| \( c_1 \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |
| \( c_2 \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |
| \( c_3 \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |
| \( \vdots \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |
| \( c_2^{d+1} \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |
| \( c_2^{d+1} \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |
| \( \vdots \) | \( \begin{array}{cccc}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\end{array} \) |

\( c(s_0) = 0 \)
Pick concepts \( \{ c_z \} \subseteq C: \ c_z(s_0) = 0, \ c_z(s_i) = E(z)_i \ \forall \ i \)

Suppose \( VC(C) = d + 1 \) and \( \{ s_0, \ldots, s_d \} \) is shattered by \( C \), i.e.,
\( |C| \times (d + 1) \) rectangle of \( \{ s_0, \ldots, s_d \} \) contains \( \{0, 1\}^{d+1} \)

<table>
<thead>
<tr>
<th>Concepts ( c \in C )</th>
<th>Truth table</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_1 )</td>
<td>( s_0 \ s_1 \ \ldots \ s_{d-1} \ s_d \ \ldots \ \ldots )</td>
</tr>
<tr>
<td>( c_2 )</td>
<td>( 0 \ 0 \ \ldots \ 1 \ 0 \ \ldots \ \ldots )</td>
</tr>
<tr>
<td>( c_3 )</td>
<td>( 0 \ 0 \ \ldots \ 1 \ 1 \ \ldots \ \ldots )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots \ \vdots \ \ldots \ \vdots \ \ldots \ \ldots )</td>
</tr>
<tr>
<td>( c_{2^d} )</td>
<td>( 0 \ 1 \ \ldots \ 1 \ 1 \ \ldots \ \ldots )</td>
</tr>
</tbody>
</table>

Among \( \{c_1, \ldots, c_{2^d}\} \), pick \( 2^k \) concepts that correspond to codewords of \( E: \{0, 1\}^k \rightarrow \{0, 1\}^d \) on \( \{s_1, \ldots, s_d\} \)
Proof sketch: Quantum sample complexity $T \geq \text{VC-dim}(C)/\varepsilon$

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob $p_z$. **Goal:** identify $z$
- Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**
- If $P_{opt}$ is the optimal success probability, then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$

How does learning relate to identification?

- Quantum PAC: **Given** $|\psi_c\rangle = |E_{c,D}\rangle^\otimes T$, learn $c$ **approximately**
- Let $\text{VC-dim}(C) = d$. Suppose $\{s_0, \ldots, s_d\}$ is shattered by $C$. Fix $D : D(s_0) = 1 - \varepsilon$, $D(s_i) = \varepsilon/d$ on $\{s_1, \ldots, s_d\}$
- Let $k = \Omega(d)$ and $E : \{0,1\}^k \rightarrow \{0,1\}^d$ be an error-correcting code
- Pick $2^k$ concepts $\{c_z\}_{z \in \{0,1\}^k} \subseteq C$: $c_z(s_0) = 0$, $c_z(s_i) = E(z)_i \ \forall \ i$
Proof sketch: Quantum sample complexity \( T \geq VC\text{-}dim(C)/\varepsilon \)

**State identification: Ensemble** \( \mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]} \)

- Given state \( |\psi_z\rangle \in \mathcal{E} \) with prob \( p_z \). **Goal:** identify \( z \)
- Optimal measurement could be quite complicated, but we can always use the Pretty Good Measurement
- If \( P_{opt} \) is the optimal success probability, then \( P_{opt} \geq P_{pgm} \geq P_{opt}^2 \)

**How does learning relate to identification?**

- **Quantum PAC:** *Given* \( |\psi_c\rangle = |E_{c,D}\rangle^T \), learn \( c \) approximately
- Let \( VC\text{-}dim(C) = d \). Suppose \( \{s_0, \ldots, s_d\} \) is shattered by \( C \). Fix \( D : D(s_0) = 1 - \varepsilon, D(s_i) = \varepsilon/d \) on \( \{s_1, \ldots, s_d\} \)
- Let \( k = \Omega(d) \) and \( E : \{0, 1\}^k \rightarrow \{0, 1\}^d \) be an error-correcting code
- Pick \( 2^k \) concepts \( \{c_z\}_{z \in \{0,1\}^k} \subseteq C : c_z(s_0) = 0, \ c_z(s_i) = E(z)_i \ \forall \ i \)
- **Learning** \( c_z \) approximately (wrt \( D \)) is equivalent to **identifying** \( z \)!
Recap

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
Recap

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{cz}\rangle = |E_{c_z,D}\rangle^\otimes T$ w.p. $\geq 1 - \delta$
Sample complexity lower bound via PGM

Recap

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{cz}\rangle = |E_{cz},D\rangle^\otimes T$ w.p. $\geq 1 - \delta$
- **Goal**: Show $T \geq d/\varepsilon$
Recap

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}⟩ = |E_{c_z,D}⟩^\otimes T$ w.p. $\geq 1 - \delta$
- Goal: Show $T \geq d/\varepsilon$

Analysis of PGM

- For the ensemble $\{ |\psi_{c_z}⟩ : z \in \{0, 1\}^k \}$ with uniform probabilities $p_z = 1/2^k$, we have
  \[ P_{\text{pgm}} \geq P_{\text{opt}} \geq (1 - \delta)^2 \]
  Recall $k = \Omega(d)$ because we used a good ECC
  \[ P_{\text{pgm}} \leq \cdots \leq \exp\left( T^2 \frac{\varepsilon^2}{d} + \sqrt{Td} \varepsilon - d - T \varepsilon \right) \]
  This implies $T = \Omega(d/\varepsilon)$
Recap

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$.
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}\rangle = |E_{c_z,D}\rangle^\otimes T$ w.p. $\geq 1 - \delta$.
- **Goal**: Show $T \geq d/\varepsilon$

Analysis of PGM

- For the ensemble $\{ |\psi_{c_z}\rangle : z \in \{0, 1\}^k \}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{\text{pgm}}$. 

Sample complexity lower bound via PGM
Sample complexity lower bound via PGM

Recap

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}\rangle = |E_{c_z,D}\rangle^\otimes T$ w.p. $\geq 1 - \delta$
- Goal: Show $T \geq d/\varepsilon$

Analysis of PGM

- For the ensemble $\{|\psi_{c_z}\rangle : z \in \{0, 1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{pgm} \geq P_{opt}^2 \geq (1 - \delta)^2$
**Recap**

- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}\rangle = |E_{c_z,D}\rangle \otimes T$ w.p. $\geq 1 - \delta$
- **Goal**: Show $T \geq d/\varepsilon$

**Analysis of PGM**

- For the ensemble $\{|\psi_{c_z}\rangle : z \in \{0, 1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{pgm} \geq P_{opt}^2 \geq (1 - \delta)^2$
- Recall $k = \Omega(d)$ because we used a good ECC
Recap

- Learning \( c_z \) approximately (wrt \( D \)) is equivalent to identifying \( z \)!
- If sample complexity is \( T \), then there is a good learner that identifies \( z \) from \( |\psi_{c_z}\rangle = |E_{c_z,D}\rangle^\otimes T \) w.p. \( \geq 1 - \delta \)
- Goal: Show \( T \geq d/\varepsilon \)

Analysis of PGM

- For the ensemble \( \{|\psi_{c_z}\rangle : z \in \{0, 1\}^k \} \) with uniform probabilities \( p_z = 1/2^k \), we have \( P_{\text{pgm}} \geq P_{\text{opt}}^2 \geq (1 - \delta)^2 \)
- Recall \( k = \Omega(d) \) because we used a good ECC
- \( P_{\text{pgm}} \leq \)
Recap

- Learning \( c_z \) approximately (wrt \( D \)) is equivalent to identifying \( z \)!
- If sample complexity is \( T \), then there is a good learner that identifies \( z \) from \( |\psi_{c_z}\rangle = |E_{c_z,D}\rangle^\otimes T \) w.p. \( \geq 1 - \delta \)
- **Goal:** Show \( T \geq d/\varepsilon \)

Analysis of PGM

- For the ensemble \( \{|\psi_{c_z}\rangle : z \in \{0, 1\}^k \} \) with uniform probabilities \( p_z = 1/2^k \), we have \( P_{pgm} \geq P^2_{opt} \geq (1 - \delta)^2 \)
- Recall \( k = \Omega(d) \) because we used a good ECC
- \( P_{pgm} \leq \cdots \) 4-page calculation \( \cdots \leq \exp( T^2 \varepsilon^2 / d + \sqrt{Td\varepsilon} - d - T \varepsilon ) \)
Sample complexity lower bound via PGM

Recap
- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}\rangle = |E_{c_z,D}\rangle \otimes^T$ w.p. $\geq 1 - \delta$
- **Goal**: Show $T \geq d/\epsilon$

Analysis of PGM
- For the ensemble $\{|\psi_{c_z}\rangle : z \in \{0, 1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{pgm} \geq P_{opt}^2 \geq (1 - \delta)^2$
- Recall $k = \Omega(d)$ because we used a good ECC
- $P_{pgm} \leq \cdots$ 4-page calculation $\cdots \leq \exp( T^2\epsilon^2/d + \sqrt{Td\epsilon} - d - T \epsilon)$
- This implies $T = \Omega(d/\epsilon)$
Sample complexity lower bound via PGM

Recap
- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}⟩ = |E_{c_z,D}\rangle^\otimes T$ with probability $\geq 1 - \delta$

Analysis of PGM
- For the ensemble $\{|\psi_{c_z}\rangle : z \in \{0, 1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{pgm} \geq P_{opt}^2 \geq (1 - \delta)^2$
- $P_{pgm} \leq \cdots$ 4-page calculation $\cdots \leq \exp(\frac{T^2\varepsilon^2}{d} + \sqrt{Td\varepsilon} - d - T\varepsilon)$
- This implies $T = \Omega(\frac{d}{\varepsilon})$

Quantum PAC learning
Sample complexity lower bound via PGM

### Recap
- Learning \( c_z \) approximately (wrt \( D \)) is equivalent to identifying \( z \)!
- If sample complexity is \( T \), then there is a good learner that identifies \( z \) from \( |\psi_{c_z}\rangle = |E_{c_z}, D\rangle^\otimes T \) with probability \( \geq 1 - \delta \)

### Analysis of PGM
- For the ensemble \( \{|\psi_{c_z}\rangle : z \in \{0, 1\}^k \} \) with uniform probabilities \( p_z = 1/2^k \), we have \( P_{pgm} \geq P_{opt}^2 \geq (1 - \delta)^2 \)
- \( P_{pgm} \leq \cdots \) 4-page calculation \( \cdots \leq \exp(T^2 \epsilon^2/d + \sqrt{Td\epsilon - d - T\epsilon}) \)
- This implies \( T = \Omega(d/\epsilon) \)

---

Quantum PAC learning → Hard distribution → Codeword concepts → Error-correcting codes → State identification
Sample complexity lower bound via PGM

Recap
- Learning $c_z$ approximately (wrt $D$) is equivalent to identifying $z$!
- If sample complexity is $T$, then there is a good learner that identifies $z$ from $|\psi_{c_z}\rangle = |E_{c_z,D}\rangle \otimes^T$ with probability $\geq 1 - \delta$

Analysis of PGM
- For the ensemble $\{\psi_{c_z} : z \in \{0, 1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{pgm} \geq P_{opt}^2 \geq (1 - \delta)^2$
- $P_{pgm} \leq \cdots$ 4-page calculation $\cdots \leq \exp(T^2\epsilon^2/d + \sqrt{Td\epsilon} - d - T\epsilon)$
- This implies $T = \Omega(d/\epsilon)$
Conclusion and future work

Further results

Classical PAC
Sample complexity

Quantum PAC
Sample complexity

Future work

Quantum machine learning is still young!

Theoretically, one could consider more optimistic PAC-like models where learner need not succeed for all $c \in C$ and all $D$.
Conclusion and future work

Further results

- **Agnostic learning**: No quantum bounds known before (unlike PAC model).
Conclusion and future work

Further results

- **Agnostic learning**: No quantum bounds known before (unlike PAC model). Showed quantum examples do not reduce sample complexity.
**Conclusion and future work**

**Classical PAC**
Sample complexity


**Quantum PAC**
Sample complexity

**Further results**

- **Agnostic learning**: No quantum bounds known before (unlike PAC model). Showed quantum examples do not reduce sample complexity.

- Also studied the model with **random classification noise** and show that quantum examples are no better than classical examples.
Conclusion and future work

**Further results**

- **Agnostic learning**: No quantum bounds known before (unlike PAC model). Showed quantum examples do not reduce sample complexity.
- Also studied the model with random classification noise and show that quantum examples are no better than classical examples.

**Future work**

- Quantum machine learning is still young! Theoretically, one could consider more optimistic PAC-like models where learner need not succeed for every $c \in C$ and $D$. 
- Efficient quantum PAC learnability of $AC^0$ under uniform $D$?
Further results

- **Agnostic learning**: No quantum bounds known before (unlike PAC model). Showed quantum examples do not reduce sample complexity.
- Also studied the model with random classification noise and show that quantum examples are no better than classical examples.

Future work

- Quantum machine learning is still young!
Conclusion and future work

Further results

- **Agnostic learning**: No quantum bounds known before (unlike PAC model). Showed quantum examples do not reduce sample complexity.
- Also studied the model with random classification noise and show that quantum examples are no better than classical examples.

Future work

- Quantum machine learning is still young!
- Theoretically, one could consider more optimistic PAC-like models where learner need not succeed $\forall c \in C$ and $\forall D$.
**Conclusion and future work**

**Further results**
- **Agnostic learning:** No quantum bounds known before (unlike PAC model). Showed quantum examples do not reduce sample complexity.
- Also studied the model with random classification noise and showed that quantum examples are no better than classical examples.

**Future work**
- Quantum machine learning is still young!
- Theoretically, one could consider more optimistic PAC-like models where learner need not succeed $\forall c \in C$ and $\forall D$.
- **Efficient** quantum PAC learnability of $AC^0$ under uniform $D$?
Suppose \( \{s_0, \ldots, s_d\} \) is shattered by \( C \). By definition:
\[
\forall a \in \{0, 1\}^d \ \exists c \in C \text{ s.t. } c(s_0) = 0, \text{ and } c(s_i) = a_i \ \forall \ i \in [d]
\]

Fix a nasty distribution \( D \):
\[
D(s_0) = 1 - 4\varepsilon, \ D(s_i) = 4\varepsilon/d \text{ on } \{s_1, \ldots, s_d\}.
\]

Good learner produces hypothesis \( h \) s.t.
\[
h(s_i) = c(s_i) = a_i \text{ for } \geq \frac{3}{4} \text{ of } i
\]

Think of \( c \) as uniform \( d \)-bit string \( A \), approximated by \( h \in \{0, 1\}^d \) that depends on examples \( B = (B_1, \ldots, B_T) \)

\[
\begin{align*}
1 \quad & I(A : B) \geq I(A : h(B)) \geq \Omega(d) & \text{[because } h \approx A]\n2 \quad & I(A : B) \leq \sum_{i=1}^T I(A : B_i) = T \cdot I(A : B_1) & \text{[subadditivity]}\n3 \quad & I(A : B_1) \leq 4\varepsilon & \text{[because prob of useful example is } 4\varepsilon]\n\end{align*}
\]

This implies \( \Omega(d) \leq I(A : B) \leq 4T\varepsilon \), hence \( T = \Omega(\frac{d}{\varepsilon}) \)

For analyzing quantum examples, only step 3 changes:
\[
I(A : B_1) \leq O(\varepsilon \log(d/\varepsilon)) \Rightarrow T = \Omega(\frac{d}{\varepsilon} \frac{1}{\log(d/\varepsilon)})
\]
Suppose we’re given state $|\psi_i\rangle$ with prob $p_i$, $i = 1, \ldots, m$. Goal: learn $i$.

Optimal measurement could be quite complicated, but we can always use the **Pretty Good Measurement**. This has POVM operators

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

where $\rho = \sum_i p_i |\psi_i\rangle \langle \psi_i|$

Success probability of PGM: $P_{PGM} = \sum_i p_i \text{Tr}(M_i |\psi_i\rangle \langle \psi_i|)$

Crucial property (BK’02): if $P_{OPT}$ is the success probability of the optimal POVM, then $P_{OPT} \geq P_{PGM} \geq P_{OPT}^2$

Let $G$ be the $m \times m$ Gram matrix of the vectors $\sqrt{p_i} |\psi_i\rangle$, then $P_{PGM} = \sum_i \sqrt{G}(i, i)^2$
For the ensemble $\{|\psi_{cz}\rangle : z \in \{0, 1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{PGM} \geq (1 - \delta)^2$.

Let $G$ be the $2^k \times 2^k$ Gram matrix of the vectors $\sqrt{p_z} |\psi_{cz}\rangle$, then $P_{PGM} = \sum_z \sqrt{G}(z, z)^2$.

$G_{xy} = g(x \oplus y)$. Can diagonalize $G$ using Hadamard transform, and its eigenvalues will be $2^k \hat{g}(s)$. This gives $\sqrt{G}$.

$\sum_z \sqrt{G}(z, z)^2 \leq \cdots$ 4-page calculation $\cdots \leq \exp(T^2 \varepsilon^2 / d + \sqrt{Td} \varepsilon - d - T \varepsilon)$

This implies $T = \Omega(d/\varepsilon)$.