Lecture 4: Learning classes of quantum states

Srinivasan Arunachalam (IBM Quantum)
Learning interesting quantum states

Fundamental *drawback* of almost all results so far in the last lecture

- For tomography on $n$ qubits, the sample complexity is $O(2^{2n})$
- For shadow tomography, PAC learning the sample complexity is poly$(n)$ but the time complexity is large

Is it possible to *time or sample-efficiently learn interesting states*?

*In this lecture.*

- Learning Gibbs states of local Hamiltonians
- Learning stabilizer states
- Statistical learning
Learning Hamiltonians. Given Gibbs states of Hamiltonians, learn the Hamiltonian?

Problem definition. Let $H$ be a $\kappa$-local Hamiltonian acting on $n$ qubits written as $H = \sum_{i=1}^{m} \mu_i E_i$ for an orthonormal $k$-local basis $\{E_i\}$. Given $T$ copies of a Gibbs state

$$\rho = \frac{e^{-\beta H}}{\text{Tr}(e^{-\beta H})},$$

output $\mu' = (\mu'_1, \ldots, \mu'_m)$ such that $\|\mu' - \mu\|_2 \leq \varepsilon$.

Motivation for this problem. Physics perspective, verification of quantum systems, Machine learning, Experimental motivation

Result [AAKS’20]: No. of copies of $\rho$ to solve HLP is $\tilde{\Theta}(\text{poly}(e^{\beta+\kappa}, 1/\beta, 1/\varepsilon, n^3))$. 
Recall: Given copies of $\rho_\mu = \frac{1}{Z_\beta} e^{-\beta H}$ where $H = \sum_i \mu_i E_i$, output approximation of $\mu$

**Sufficient statistics:** Just use shadow tomography

1. Suppose we have approximations $e'_i$ of $e_i = \text{Tr}(E_i \rho_\mu)$ for all $i \in [m]$ satisfying $|e'_i - e_i| \leq \varepsilon$, can we recover $\mu$? Using [Aar’18, HKP’20, BO’20]?

2. Classical post-processing produces $\rho' \approx \rho_\mu$, but that doesn’t even imply $\rho'$ is a Gibbs state $e^{-\beta H'}$, so approximating $\mu$ is unclear!

**Observation 1:** suppose we maximize over $\rho_\lambda = e^{-\beta H}$ where $H = \sum_i \lambda_i E_i$ s.t.

$$\text{Tr}(\rho_\lambda E_i) = \text{Tr}(\rho_\mu E_i) \quad \text{for every } i \in [m],$$

then $\rho_\lambda = \rho_\mu$ which implies $\lambda = \mu$. Isn’t this “hard”?

**Observation 2:** Maximum entropy principle $\rightarrow$ Cast as an optimization problem

$$\max_{\sigma} \quad S(\sigma)$$

$$\text{s.t.} \quad \text{Tr}[\sigma E_i] = e_i, \quad \forall i \in [m]$$

$$\sigma \succeq 0, \quad \text{Tr}[\sigma] = 1.$$  \hspace{1cm} (1)

where $S(\sigma) = -\text{Tr}[\sigma \log \sigma]$ is the quantum entropy of $\sigma$. Optimum of (1) equals $\rho_\mu$
Quantum proof: First idea (continued)

Recall: Given copies of $\rho_\mu = \frac{1}{Z_\beta} e^{-\beta H}$ where $H = \sum_i \mu_i E_i$, output approximation of $\mu$.

Maximum entropy principle: $\sigma$ with equal marginals $\{e_i\}$ & maximum entropy is $\rho_\mu$.

Given approximations $e'_i$ of $e_i = \text{Tr}(E_i \rho_\mu)$ for $i \in [m]$ satisfying $|e'_i - e_i| \leq \varepsilon$ recover $\mu$?

$$\max_{\sigma} S(\sigma) \quad \text{s.t.} \quad \text{Tr}[\sigma E_i] = e_i, \quad \forall i \in [m] \quad \sigma \succcurlyeq 0, \quad \text{Tr}[\sigma] = 1.$$  

If $\rho_\mu$ maximizes first and $\rho_\mu'$ maximizes second problem, then $\|\rho_\mu - \rho_\mu'\|_1 = O(m\varepsilon)$.

Does this suffice for our problem in approximating the $\mu$s? No.

In order to approximate $\mu$, need to bound

$$\|\log \rho_\mu - \log \rho_\mu'\|_1$$

Could be exponentially worse than $\|\rho_\mu - \rho_\mu'\|_1$.

Issue is non-Lipschitz nature of $\log(x)$ function.
Recall: Given copies of \( \rho \mu = \frac{1}{Z_\beta} e^{-\beta H} \) where \( H = \sum_i \mu_i E_i \), output approximation of \( \mu \)

**How to handle** \( \log(\rho \mu) - \log(\rho \mu') \)? Let’s take a look at the dual

\[
\max_{\sigma} \quad S(\sigma) \\
\text{s.t.} \quad \text{Tr}[\sigma E_i] = e_i, \quad \forall i \in [m] \\
\sigma \succeq 0, \quad \text{Tr}[\sigma] = 1.
\]

Dual \( \rightarrow \)

\[
\mu = \arg \min_{\lambda_1, \ldots, \lambda_m} \log Z_\beta(\lambda) + \beta \cdot \sum_i \lambda_i e_i,
\]

where \( Z_\beta(\lambda) = \text{Tr}(e^{-\beta H}) \) & \( H = \sum_i \lambda_i E_i \)

**Issue:** Don’t have \( e_i = \text{Tr}(\rho \mu E_i) \), but only \( e_i' \) satisfying \( |e_i' - e_i| \leq \varepsilon \), so we are solving

\[
\max_{\sigma} \quad S(\sigma) \\
\text{s.t.} \quad \text{Tr}[\sigma E_i] = e_i', \quad \forall i \in [m] \\
\sigma \succeq 0, \quad \text{Tr}[\sigma] = 1.
\]

Dual \( \rightarrow \)

\[
\mu' = \arg \min_{\lambda_1, \ldots, \lambda_m} \log Z_\beta(\lambda) + \beta \cdot \sum_i \lambda_i e_i'
\]

How far is \( \mu \) from \( \mu' \) given \( \varepsilon \) additive approximations of \( \{e_i\}_i \)?

**Strong convexity:** Puts a bound on how slow the function changes.

Let \( f : \mathbb{R}^m \rightarrow \mathbb{R} \). If \( \nabla^2 f \succ \alpha I \), then for every \( \nu, \nu' \in \mathbb{R}^m \)

\[
f(\nu') - f(\nu) - \nabla f(\nu)^T (\nu' - \nu) \geq \alpha \|\nu' - \nu\|_2 \quad \text{(Think of } f(\cdot) = \log Z_\beta(\cdot))
\]
Recall: Given copies of $\rho_\mu = \frac{1}{Z_\beta} e^{-\beta H}$ where $H = \sum_i \mu_i E_i$, output approximation of $\mu$

Result [AAKS’20]: No. of copies of $\rho$ to solve HLP is $\tilde{\Theta}(\text{poly}(e^{\beta+\kappa}, 1/\beta, 1/\varepsilon, n^3))$.

1. **Estimating marginals** Shadows to get $e'_i$ s.t. $|e'_i - \text{Tr}(E_i \rho_\mu)| \leq \delta$

2. **Sufficient statistics** We then solve the optimization problem

$$
\mu' = \max_{\lambda_1, \ldots, \lambda_n} \log Z_\beta(\lambda) + \beta \sum_i \lambda_i e'_i
$$

3. We show $\|\mu - \mu'\|_2 \leq \varepsilon$ by taking sufficient samples. Crucially showing log partition function is strong convex.

A few remarks:

1. Algorithm not time efficient for generic Hamiltonians
2. Except obtain measurement statistics of $\rho$, our algorithm is classical
3. Exponential in $\beta, \kappa$: Might seem bad, but cannot be generically avoided
4. [HKT’22] considered small $\beta$, the sample complexity is $(\log n)/(\beta^2 \varepsilon^2)$. 
Weyl matrices

Pauli matrices: \[ I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \]

\( n \)-qubit Pauli matrices \( \{I, X, Y, Z\}^n \) form an orthonormal basis for \( \mathbb{C}^n \).

In particular, for every \( x = (a, b) \in \mathbb{F}_2^n \), define a Weyl operator

\[ W_x = i^{a \cdot b} (X^{a_1} Z^{b_1} \otimes X^{a_2} Z^{b_2} \otimes \ldots \otimes X^{a_n} Z^{b_n}). \]

\( \{W_x\} \) are orthonormal, form a basis for quantum states, i.e., for every \( |\psi\rangle \), we have

\[ |\psi\rangle \langle \psi| = \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} \alpha_x \cdot W_x, \]

where

\[ \alpha_x = \text{Tr}(W_x |\psi\rangle \langle \psi|), \quad \frac{1}{2^n} \sum_x \alpha_x^2 = 1. \]

Below we will use \( p_{\psi}(x) = \alpha_x^2 / 2^n \), so that \( \sum_x p_{\psi}(x) = 1. \)

Note that this is similar to Fourier decomposition of a Boolean function \( f : \{0, 1\}^n \rightarrow \{0, 1\} \) as \( f(x) = \sum_S \hat{f}(S) \chi_S(x) \) where \( \sum_S \hat{f}(S)^2 = 1. \)
Recall. Pauli matrices \{\mathbb{I}, X, Y, Z\}^n form an orthonormal basis for \(\mathbb{C}^n\).
In particular, for every \(x = (a, b) \in \mathbb{F}_2^n\), define a Weyl operator
\[
W_x = i^{a \cdot b} (X^{a_1} Z^{b_1} \otimes X^{a_2} Z^{b_2} \otimes \cdots \otimes X^{a_n} Z^{b_n}).
\]
Define \(p_\psi(x) = \langle \psi | W_x | \psi \rangle^2 / 2^n\), and we have \(\sum_x p_\psi(x) = 1\).

Bell basis. Observe that \(\{ |W_b\rangle = (W_b \otimes \mathbb{I}) |\Phi^+\rangle : b \in \{0, 1\}^2 \} \) where
\( |\Phi^+\rangle = (|00\rangle + |11\rangle) / \sqrt{2} \) is an orthonormal basis for \(\mathbb{C}^2\).

Bell sampling.
- **Input:** Bell sampling takes 4 copies of an \(n\)-qubit \(|\psi\rangle\).
- **Procedure:** Using the first two copies of \(|\psi\rangle\), measure qubit \(i, n + i\) in the Bell basis to obtain \((b_i, b_i')\). Call the resulting string \(x = ((b_1, b_1'), \ldots, (b_n, b_n'))\).
  
  Using the second two copies of \(|\psi\rangle\) to obtain \(y = ((c_1, c_1'), \ldots, (c_n, c_n'))\).
- **Output:** \(x + y \in \mathbb{F}_2^2\).

**Theorem.** The output \(z \in \mathbb{F}_2^n\) above is sampled according to the distribution
\[
q_\psi(z) = \sum_{a \in \mathbb{F}_2^n} p_\psi(a) p_\psi(z + a).
\]
Bell sampling for learning stabilizer states

**Stabilizer states.** Consider a Clifford circuit $C$ (i.e., consisting of $H$, $S$, $CNOT$ gates) then output of $C |0^n\rangle$ is a stabilizer state!

Alternatively, a stabilizer state $|\psi\rangle$ is a pure states such that there is a subgroup $S \subseteq \{W_x\}$ of size $2^n$ such that $P |\psi\rangle = |\psi\rangle$ for all $P \in S$. In particular,

$$|\psi\rangle\langle\psi| = \sum_{\sigma \subseteq S} \sigma,$$

where $S \subseteq \{W_x\}_x$ has dimension $n$.

Observe

$$p_\psi(z) = 2^{-n} \cdot \langle\psi|W_z|\psi\rangle^2 = \begin{cases} 2^{-n} & z \text{ stabilizes } |\psi\rangle \\ 0 & \text{otherwise} \end{cases}$$

Hence

$$q_\psi(z) = \sum_{a \in \mathbb{F}_2^2} p_\psi(a)p_\psi(z+a) = 2^{-n} \sum_{a \in \text{Stab}(\psi)} p_\psi(z+a) = 2^{-n}[z \in \text{Stab}|\psi\rangle] = p_\psi(z).$$

Bell sampling uses 4 copies of $|\psi\rangle$ and produces a $z \sim q_\psi(z)$ (or $z \sim p_\psi(z)$). In particular, 4 copies produces a $W_z$ that stabilizes the unknown $|\psi\rangle$.

**Learn the basis.** Repeat the above for $O(n)$ times, to obtain $n$ many linearly independent $z$s that stabilize $|\psi\rangle$. Call it $\{W_1, \ldots, W_n\}$. Time complexity is $O(n^3)$
Some recent improvements

Learning beyond stabilizer states

1. Single layer of $T$ gates
   - [LC’21] considered learning states that are output of circuits that consist of Clifford + one layer of at most $O(\log n)$ many $T$ gates.
   - Write a stabilizer decomposition of $|T\rangle$ state and write the resulting state as a sum of $2^k$ many stabilizer states.
   - A technical modification of Bell sampling learns in time $\text{poly}(n, 2^k)$

2. A sequence of works [GIKL ’22,’22,’23), LOH’23, HG’23] showed how to learn Clifford + $k$ many $T$ gates in time $\text{poly}(n, 2^k)$
   - Stabilizer dimension ($\psi$) = $\dim(\{z : \langle\psi|W_z|\psi\rangle \neq 0\})$.
   - GIKL’23 showed how to learn states $|\psi\rangle$ with stabilizer dimension $\geq n - k$ in times $\text{poly}(n, 2^k)$
   - Main idea is Bell sampling
   - If $|\psi\rangle$ produced by Clifford+ $k$ many non-Clifford gates, then Stabilizer dimension($\psi$) $\geq n - k$. 
Bell sampling, viewed as taking derivatives


Every stabilizer state written as $|\psi\rangle = \sum_{x \in A} (-1)^{s \cdot x} |x\rangle$ for a subspace $A \subseteq \mathbb{F}_2^n$

A learning algorithm performs the following: (for simplicity let $A = \{0, 1\}^n, S = 0^n$)

Goal is to learn $B \in \mathbb{F}_2^{n \times n}$. Take two copies of $|\psi\rangle$

$|\psi\rangle \otimes |\psi\rangle = \sum_{x,y} (-1)^{x^t B x + y^t B y} |x, y\rangle$

$\xrightarrow{CNOT} \sum_{x,y} (-1)^{x^t B x + y^t B y} |x, x + y\rangle$

$= \sum_{x,z} (-1)^{x^t (B + B^t) z + z^t B z} |x, z\rangle = \sum_{x,z} (-1)^{x^t (B + B^t) z + z^t B z} |x, z\rangle$

Measure the second register and suppose we obtain $\tilde{z}$, resulting state is

$(-1)^{\tilde{z}^t B \tilde{z}} \left( \sum_{x} (-1)^{x^t (B + B^t) \tilde{z}} |x\rangle \right) |\tilde{z}\rangle \xrightarrow{BV} |(B + B^t) \cdot \tilde{z}\rangle |\tilde{z}\rangle$

Two copies of $|\psi\rangle$ allow to take one derivative of $x^t B x$ (in the direction of $\tilde{z}$).
Take $\tilde{O}(n)$ more copies to take $n$ derivatives and learn $B$, hence $|\psi\rangle$ completely
Learning phase states

A natural extension? Let

$$|\psi_f\rangle = \frac{1}{\sqrt{2^n}} \sum_{x \in \{0,1\}^n} (-1)^{f(x)} |x\rangle$$

where degree of \(f\) equals \(d\). How many copies of \(|\psi_f\rangle\) suffice to learn \(f\)?

Why care about phase states?

- **Pseudorandomness**: If \(f\) is a pseudorandom function, then \(|\psi_f\rangle\) is indistinguishable from a Haar random state.
- **IQP circuits**: Applying \(\{Z, CZ, CCZ\}\) to \(H^n |0^n\rangle\) produces a degree-3 phase state.
- **Quantum complexity**: Recently several results using phase states.

In [ABDY'22], we considered the learning question and showed optimal bounds.

<table>
<thead>
<tr>
<th></th>
<th>Separable</th>
<th>Entangled</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree-(d) Binary phase state</td>
<td>(\Theta(n^d))</td>
<td>(\Theta(n^{d-1}))</td>
</tr>
</tbody>
</table>
Learning phase states with entangled measurements

A first approach. Taking derivatives? let $f$ be degree-3

$$|\psi_f\rangle^\otimes2 \mapsto (s, \sum_x (-1)^f(x)+f(x+s) |x\rangle).$$

Recall that when $f$ was degree-2, then $f(x) + f(x + s)$ was degree-1, so we can learn it by applying Hadamards, but now $f(x) + f(x + s)$ is degree-2, unclear how to learn!

An entangled learning algorithm.

(1.) Pretty good measurement for the ensemble $\mathcal{E} = \{|\psi_f\rangle^\otimes k : f \in P(n,d)\}$

(2.) The failure probability of the PGM for $\mathcal{E}$ is given by

$$\frac{1}{|\mathcal{E}|} \sum_{f \neq g} \langle \psi_f | \psi_g \rangle^k = \sum_{g \in P^*(n,d)} [1 - 2 \Pr_x [g(x) \neq 0]]^k$$

$$= \sum_{g \in P^*(n,d)} [1 - 2\text{wt}(g)]^k$$

$$= \sum_{\ell=1}^{d-1} \sum_{g \in P^*(n,d)} [1 - 2|g|/2^n]^k [\text{wt}(g) \in [2^{n-\ell-1}, 2^{n-\ell}]]$$

(3.) Use weight properties of Reed-Muller codes to show above is $\leq \exp(-k + n^{d-1})$, which is $\leq 1/100$ for $k = O(n^{d-1})$

(4.) Optimal since there are $n^d$ bits of information and each $|\psi_f\rangle$ has $n$ bits
Quantum statistical query learning

**Problem.** Let $C$ be a class of functions $c : \{0, 1\}^n \rightarrow \{0, 1\}$.

In quantum learning theory, given access to $T$ copies of

$$|\psi_c\rangle = \frac{1}{\sqrt{2^n}} \sum_x |x, c(x)\rangle,$$

can we learn $c$?

**Entangled measurements:** Joint-measurement on $|\psi_c\rangle \otimes T$

**Separable measurements** The learner can only apply single-copy measurements

**QSQ model**

1. Maybe even entangled and separable measurements far from NISQ.
2. In [AGY'20], introduced the quantum statistical query model (QSQ)
3. QSQ learner can make Qstat queries: specifies $\|M\| \leq 1$ and $\tau \in [0, 1]$

$$\text{Qstat} : (M, \tau) \rightarrow \alpha_M \in \left[ \langle \psi_c | M | \psi_c \rangle - \tau, \langle \psi_c | M | \psi_c \rangle + \tau \right]$$

How many Qstat queries are necessary/sufficient to learn $c$?

4. Say someone in the “cloud” possesses $|\psi_c\rangle$ and a learner is purely classical

5. Quantum examples are useful for learning parities, juntas, DNF formulas, coupon collector: all of these algorithms need only QSQ measurements!
Let $C$ be a class of functions $c : \{0, 1\}^n \rightarrow \{0, 1\}$ and $|\psi_c\rangle = \frac{1}{\sqrt{2^n}} \sum_x |x, c(x)\rangle$.

Given access to $Qstat$ queries, can we learn $c$?

Almost all quantum learning algorithms can be converted into the QSQ model.

What separates QSQ from entangled/separable measurements?

Classically.

1. Uniform PAC learning i.e., given just uniform $(x, c(x))$
2. Classical SQ queries, i.e., $M = \text{diag}(\{\phi(x)\}_x)$ for some arbitrary $\phi : \{0, 1\}^{n+1} \rightarrow [0, 1]$

What separates classical SQ and classical PAC? Parities!

Quantumly One can show that degree-2 functions separates QSQ and QPAC!
Let $C$ be a class of functions $c : \{0, 1\}^n \rightarrow \{0, 1\}$ and $|\psi_c \rangle = \frac{1}{\sqrt{2^n}} \sum_x |x, c(x)\rangle$.

**Given access to Qstat queries, can we learn $c$?**

Almost all quantum learning algorithms can be converted into the QSQ model.

In [AHS’23], for the concept class $C = \{c(x) = x^t A x : A \in \mathbb{F}_2^{n \times n}\}$, then

- Entangled complexity: $\Theta(n)$
- Separable complexity: $\Theta(n^2)$
- QSQ complexity: $\Theta(2^n)$

**Learning with noise:** given copies of $\sum_x |x\rangle \sqrt{1 - \eta} |c(x)\rangle + \sqrt{\eta} |c(x)\rangle$, learnable in poly($n, 1/(1 - 2\eta)$) time.

**Some consequences:**

- Separates QSQ from Quantum learning with classification noise (a natural classical analogue is an open question)
- Exponential separation between Weak and strong error mitigation
Let $C$ be a class of quantum states (no longer Boolean functions).

QSQ algorithm of learning $C$ makes Qstat queries: for an unknown $\rho \in C$

$$\text{Qstat} : (M, \tau) \rightarrow \alpha_M \in [\text{Tr}(M\rho) - \tau, \text{Tr}(M\rho) + \tau]$$

[CCHL’21] showed that for few computational tasks (such as shadow tomography): $O(n)$ entangled measurements suffice but $2^n$ separable measurements are necessary.

- Introduce a statistical dimension which gives lower bounds for QSQ
- Show that the CCHL states require $\geq 4^n$ copies to learn
- Show that Abelian coset states requires $\geq 2^n$ copies to learn
- For several algorithms like learning Gibbs state, trivial states, the algorithm can be implemented in QSQ
Directions and outlook

Through these lectures.

1. Considered learning **Boolean functions** using quantum examples. Sometimes useful sometimes not

2. Considered learning **quantum states** exactly, approximately and their properties

3. Considered **interesting classes** of states and saw efficient algorithms

Open questions.

1. Learning more interesting classes of states
2. More “realistic” learning theory models motivated by near-term
3. Connections between learning and other topics
4. Several surveys on this topic, containing many interesting open questions

THANK YOU