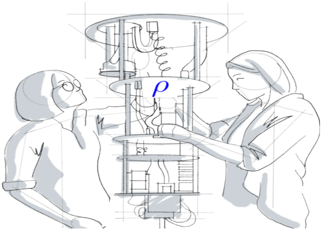
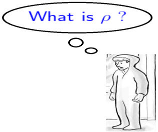
Overview of results for learning quantum states

Srinivasan Arunachalam (IBM Quantum)





Learning quantum states: overview

So far. We looked at learning Boolean functions $c : \{0,1\}^n \to \{0,1\}$ encoded as a quantum example state

$$\sum_x \sqrt{D(x)}\,|x, c(x)\rangle$$

and looked strengths and weakness of these examples for learning c.

What if we are given ρ, an unknown quantum state?

1. Copies of ρ, learn ρ

   Tomography

2. Measurement statistics of ρ, learn ρ "approximately well"

   PAC learning and shadow tomography

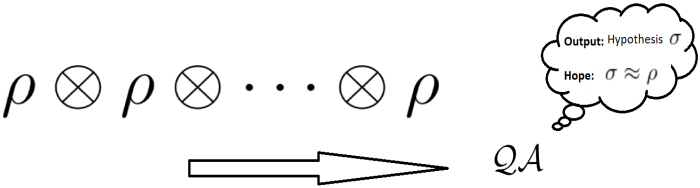
3. When ρ is an interesting class of states

   Gibbs states of local Hamiltonians, stabilizer states and their variants

Learning quantum states: Tomography

Let ρ be an n-qubit quantum state $\rho \in \mathbb{C}^{D\times D}$ with $D = 2^n$.

Tomography. How many copies of ρ are necessary and sufficient to produce classical description of a state σ that approximates ρ well enough?



Motivation. Fundamental question, experimental verification, understanding noise.

Output requirement. σ satisfies $\|\sigma - \rho\|_{tr} \leq \varepsilon$.

A trivial algorithm. Estimate each entry well enough requires $O(D^6)$ copies

Subsequent works. Using compressive sensing $O(D^4)$ copies and matrix recovery $O(D^3)$ copies

A breakthrough in 2015. Showed how to do tomography in complexity $O(D^2/\varepsilon^2)$. [OW'15, HHJWY'15]. Known to be optimal.

Tomography protocols

Two protocols for tomography using $O(d^2/\varepsilon^2)$ copies.

1. [OW'15]: Spectrum estimation using Schur sampling and the classical RSK algorithm for state reconstruction
2. [HHJWY'15]: Used the PGM, analyzed via Schur-Weyl duality

Remark: [OW'15, HHJWY'15] showed that $O(dr/\varepsilon^2)$ copies suffice when $\mathrm{rank}(\rho) = r$.

Pure state tomography. Given $|\psi\rangle^{\otimes T}$, output ϕ such that $\langle\psi|\phi\rangle \geq 1 - \varepsilon$

Protocol. Apply a natural measurement!

1. Ensemble: Uniform over $|\psi\rangle^{\otimes T}$
2. POVM: Apply the POVM $\{E_{|\phi\rangle}^{\otimes T} : |\phi\rangle\}$, where

$$E_{|\phi\rangle} = \binom{T+d-1}{d-1} |\phi\rangle^{\otimes T} \langle\phi|^{\otimes T} \, d\phi$$

   (although continuous, can implemented by taking an appropriate discretization)
3. Output: Output the resulting $|\phi\rangle$.
4. Sample complexity: If $T = O(d/\varepsilon^2)$, then $\| \, |\phi\rangle - |\psi\rangle \, \|_{tr} \leq \varepsilon$.

Remains to see. Why $\{E_\psi^{\otimes T} : \psi\}$ is a POVM and the sample complexity bound!

Tomography protocols

Why a valid POVM? One needs to show that $\int_{|\phi\rangle} E_{|\phi\rangle} = \mathbb{I}$. To this end, observe that

$$\int_{|\phi\rangle} \binom{T+d-1}{d-1} |\phi\rangle^{\otimes T} \langle\phi|^{\otimes T} \, d\phi = \binom{T+d-1}{d-1} \int_{|\phi\rangle} |\phi\rangle^{\otimes T} \langle\phi|^{\otimes T} \, d\phi$$

$$= \binom{T+d-1}{d-1} \frac{\Pi_{sym}^{T,d}}{\mathrm{Tr}(\Pi_{sym}^{T,d})} = \Pi_{sym}^{T,d},$$

since our POVM acts *only* on the symmetric subspace, this equals $\mathbb{I}$

Output of the algorithm. On input $|\psi\rangle^{\otimes T}$, the expected output $|\phi\rangle$ satisfies

$$\begin{aligned}
\mathbb{E}_{\phi\sim\mathrm{POVM}}[\langle\psi|\phi\rangle^2] &= \int_\phi \langle\psi|\phi\rangle^2 \cdot \Pr[\text{POVM outputs } |\phi\rangle] d\phi \\
&= \int_\phi \langle\psi|\phi\rangle^2 \cdot \langle\psi|\phi\rangle^{2T} \cdot \binom{d+T-1}{d-1} d\phi \\
&= \binom{d+T-1}{d-1} \cdot \int_\phi \langle\psi|\phi\rangle^{2T+2} d\phi = \binom{d+T-1}{d-1} \cdot \frac{1}{\binom{d+T}{d}} \sim 1 - d/T.
\end{aligned}$$

Hence the expected distance between $|\phi\rangle$ and $|\psi\rangle$ is $\sqrt{d/T}$, which is ε for $T = d/\varepsilon^2$.

Alternative learning models?

- Given $D = 2^n$, complexity is large for $n = 10$ (best known experiment)
- Learning ρ entirely is an overkill, maybe want to learn only certain aspects?

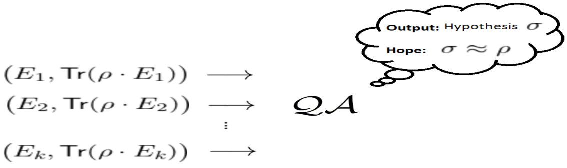
PAC learning quantum states

So far. Tomography learned the entire quantum state ρ. Producing a σ s.t. $\|\sigma - \rho\|_{\mathrm{Tr}}$ is small, means we'v learned $\mathrm{Tr}(E \cdot \rho)$ "approximately" for every E!

Relax this goal? Learn ρ approximately well for "most" 2-outcome measurements?

PAC learning quantum states. Motivation is classical PAC learning.

Valiant gave a complexity-theoretic definition of what it means to learn: introduced the Probably Approximately Correct model



1. let $D : \mathcal{E} \rightarrow [0,1]$ be a distribution over all possible two-outcome measurements
2. Given $E_1, \ldots, E_k$ sampled from D along with $\mathrm{Tr}(\rho \cdot E_1), \ldots, \mathrm{Tr}(\rho \cdot E_k)$
3. Goal is to produce σ that satisfies

$$\Pr\Big[|\mathrm{Tr}(\rho E) - \mathrm{Tr}(\sigma E)| \leq \varepsilon\Big] \geq 1 - \delta,$$

i.e., probably (with prob. $\geq 1 - \delta$) over $E \sim D$, can approximately learn $\mathrm{Tr}(\rho E)$.

PAC learning protocol

Recall. ρ is unknown. $D : \mathcal{E} \to [0,1]$ distribution over measurements. Given Given $\{(E_i, \mathrm{Tr}(\rho E_i))\}_{i=1}^k$ where $E_i \sim D$, produce σ s.t. $\Pr\Big[|\mathrm{Tr}(\rho E) - \mathrm{Tr}(\sigma E)| \le \varepsilon\Big] \ge 1 - \delta$.

Is PAC learning sample complexity smaller than $O(D^2)$ tomography complexity?

Yes! Aaronson'03 showed that $O(\log D)$ many examples suffice to produce a σ!

Proof sketch.

- Take $O(\log D)$ many examples, just find a σ that is consistent with these trace measurement outcomes!
- Why does this work? VC-theory for real-valued functions!
- Consider the function $f_\rho : \mathcal{E} \to [0,1]$ defined as $f_\rho(E) = \mathrm{Tr}(\rho \cdot E)$. Let $\mathcal{C} = \{f_\rho : \mathcal{E} \to [0,1]\}_\rho$ be the concept class of interest
- Well known that learning $\mathcal{C}$ can be done using fat-shattering dimension $\mathrm{fat}(\mathcal{C})$-many samples of the form $(E, f_\rho(E))$ where $E \sim D$
- Using random-access codes one can show $\mathrm{fat}(\mathcal{C}) = O(\log D)$

PAC learning protocol

Recall. ρ is unknown. $D : \mathcal{E} \to [0,1]$ distribution over measurements. Given Given $\{(E_i, \text{Tr}(\rho E_i))\}_{i=1}^k$ where $E_i \sim D$, produce σ s.t. $\Pr\left[|\text{Tr}(\rho E) - \text{Tr}(\sigma E)| \leq \varepsilon\right] \geq 1 - \delta$.

Is PAC learning sample complexity smaller than $O(D^2)$ tomography complexity?

Yes! Aaronson'03 showed that $\text{sfat}(\mathcal{C}) = O(\log D)$ examples suffice to produce σ!

Remarks.

- Sample complexity of PAC learning quantum states is exponentially better than tomography!
- Time complexity is still large!
- Morally,

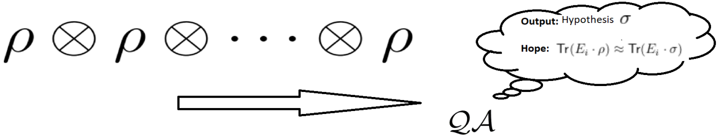
  VC dimension characterizes learning Boolean functions.

  Fat-shattering dimension characterizes learning quantum states

Shadow tomography

So far. Estimated ρ on a distribution of measurements. But if we fix a set of measurements, can we learn faster?

$$\mathcal{E} = \{E_1, \ldots, E_k\}$$



Problem. Given $\{E_1, \ldots, E_k\}$, how many copies of ρ suffice to estimate $\mathrm{Tr}(\rho E_1), \ldots, \mathrm{Tr}(\rho E_k)$?

Trivial algorithm.

Tomography: Uses D^2 copies of ρ.

Empirical: Use $1/\varepsilon^2$ copies of ρ and estimate $\mathrm{Tr}(\rho E_i)$, totally k/ε^2 copies of ρ.

Better algorithm. Using ideas from online learning and PAC learning, Aaronson'17 proposed an algorithm that can estimate $\mathrm{Tr}(\rho E_1), \ldots, \mathrm{Tr}(\rho E_k)$ using $O(\log k, \log D)$ copies of ρ (exponentially better than trivial)

Shadow tomography protocol

Problem. Given $\{E_1, \ldots, E_k\}$, how many copies of ρ suffice to estimate $\mathrm{Tr}(\rho E_1), \ldots, \mathrm{Tr}(\rho E_k)$ up to error ε?

Protocol idea

Part 1: Communication complexity

Suppose Alice has ρ, Bob has $\{E_1, \ldots, E_k\}$. Bob needs to output $\{\mathrm{Tr}(\rho \cdot E_i)\}_i$. Only Alice can communicate to Bob. Trivial protocol cost is $O(D^2)$

1. Bob guesses Alice's state sequentially $\sigma_0 = \mathbb{I}/D, \ldots, \sigma_T$ such that eventually $\mathrm{Tr}(\rho \cdot E_i) \approx \mathrm{Tr}(\sigma_T \cdot E_i)$
2. Alice sends bits in order to improve Bobs guess in each iteration :
If $|\mathrm{Tr}(\rho \cdot E_i) - \mathrm{Tr}(\sigma_i \cdot E_i)| > \varepsilon$ Alice sends $(i, \mathrm{Tr}(E_i\rho))$.
3. Bob updates his guess $\sigma_i \to \sigma_{i+1}$ as follows: consider the 2-outcome observable F that applies $\{E_i, \mathbb{I} - E_i\}$ acting on $(\log n)$ copies of σ_i and "accepts" if at least a constant fraction accepted, if so, trace out the last $(\log n) - 1$ copies and the resulting state is σ_{i+1}.
4. Clearly $|\mathrm{Tr}(\rho \cdot E_i) - \mathrm{Tr}(\sigma_T \cdot E_i)| \leq \varepsilon$ for all i
5. Main observation in [Aar'03] was it suffices to send $\mathrm{poly}(\log D, \log k)$ many bits in the communication protocol

Shadow tomography protocol: II

Problem. Given $\{E_1, \ldots, E_k\}$, no. of copies of ρ to ε-estimate $\mathrm{Tr}(\rho E_1), \ldots, \mathrm{Tr}(\rho E_k)$?

Protocol idea. 1. Communication complexity

1. Suppose Alice has ρ, Bob has $\{E_1, \ldots, E_k\}$. Approximate $\mathrm{Tr}(\rho \cdot E_i)$
2. Alice sends bits in order to improve Bobs guess in each iteration: If $|\mathrm{Tr}(\rho \cdot E_i) - \mathrm{Tr}(\sigma_i \cdot E_i)| > \delta$ Alice sends $(i, \mathrm{Tr}(E_i \rho))$.
3. Main observation in [Aar'03] was it suffices to send $\mathrm{poly}(\log D, \log k)$ many bits in the communication protocol

Protocol idea 2. Simulating this CC protocol for learning

1. In shadow tomography, there is no Alice, but just $\rho^{\otimes T}$
2. Quantum OR lemma: Given $O(\log k)$ copies of ρ decides if there

   exists $j \in [k]$ s.t. $\mathrm{Tr}(E_j \rho) \geq \Omega(1)$, or for all $j \in [k]$, $\mathrm{Tr}(E_j \rho) \ll 1/k$
3. What is j? Aar'18 used a binary-search approach to find this j
4. [Aar'18] The overall sample complexity is $O(\log^4 k \cdot \log D \cdot \varepsilon^{-5})$
5. Few works improving the dependence on these parameters.
   State of the art: $O(\log^2 k \cdot \log D \cdot \varepsilon^{-4})$ [BO'20]

Quantum hypothesis selection

Badescu & O'Donnell'20 gave a shadow tomography protocol using sample complexity using $T = O(\log^2 k \cdot \log D \cdot \varepsilon^{-4})$ copies of ρ.

Interesting corollary. Let $\mathcal{C} = \{\rho_1, \ldots, \rho_k\}$ and σ be an unknown state.

Given copies of σ, find the nearest $\rho_i \in \mathcal{C}$, i.e., find an $\ell \in [k]$ such that

$$\|\sigma - \rho_\ell\|_{tr} \leq \mathrm{OPT} + \varepsilon,$$

where $\mathrm{OPT} = \min_{i \in [k]} \|\rho_i - \sigma\|$.

Remark: If one could improve the $\log^2 k \to \log k$ in the complexity above, one can show tomography can be done using $\tilde{O}(d^2)$ copies!

1. For every $i \neq j$, by Holevo-Helstrom there exists A_{ij} such that
$$\|\rho_i - \rho_j\|_{tr} = \mathrm{Tr}(A_{ij}(\rho_i - \rho_j))$$
2. Now perform shadow tomography on $\sigma^{\otimes T}$ using the operators $\{E_{ij}\}_{i,j}$ to obtain $|\alpha_{ij} - \mathrm{Tr}(A_{ij}\sigma)| \leq \varepsilon/2$
3. Go over all $\rho \in \mathcal{C}$ to find ℓ that minimizes $\max_{ij} \mathrm{Tr}(\rho_\ell A_{ij} - \alpha_{ij})$
4. One can show that $\|\rho_i - \sigma\|_{tr} \leq 3\mathrm{OPT} + \varepsilon$.

Classical shadows

Subsequent work of [HKP'20] introduced classical shadows that

(i) given copies of ρ, creates a *classical* shadow of ρ efficiently

(ii) classical shadows used to compute expectation values of arbitrary observables

Procedure to obtain shadows.

1. Given ρ, apply a random U_i on ρ and measure to obtain $b^i \in \{0,1\}^n$
2. Classical shadows are $\{|s_1\rangle, \ldots, |s_T\rangle\}$ where $|s_i\rangle = U_i^* \left|b^i\right\rangle$
3. View the process of "average mapping" $\rho \to U|b^i\rangle\langle b^i|U^*$ as a channel $\mathbb{E}[|s_i\rangle\langle s_i|] = \mathcal{M}(\rho)$

Intuitively, one should now view $\mathbb{E}[\mathcal{M}^{-1}|s_i\rangle\langle s_i|] = \rho$, or $\mathcal{M}^{-1}|s_i\rangle\langle s_i| \approx \rho$.

Predicting expectation values. For observables E, compute

$$\mathbb{E}_i[\mathrm{Tr}(E\mathcal{M}^{-1}|s_i\rangle\langle s_i|)] := \alpha_E \approx \mathrm{Tr}(E\rho).$$

Using median-of-means estimator to output $\alpha_E \in \mathbb{R}$

Correctness. [HKP'20] showed that if $T = O(\|E\|_{shadow}/\varepsilon^2)$, then $|\alpha_E - \mathrm{Tr}(E\rho)| \leq \varepsilon$.

This bound is known to be tight

Also, given $\{E_1, \ldots, E_k\}$, the same classical shadows can be used to estimate $|\alpha_i - \mathrm{Tr}(E_i\rho)| \leq \varepsilon$ using $O((\log k) \cdot \|E\|_{shadow}/\varepsilon^2)$ copies of ρ.

Norm. We have $\|E\|_{shadow} \leq \sqrt{\mathrm{Tr}(E^2)}$. So $\|E\|_{shadow} \leq 1$ for rank-1 observables!

Some further results

So far. We saw tomography, PAC learning shadow tomography and classical shadows.

Results we didn't cover

1. Extending shadow tomography to k-outcome observables
2. Lower bounds on shadow tomography and standard tomography if allowed only separable measurements
3. Online learning quantum states
4. Learning arbitrary quantum channels or unitary channels
5. Learning matrix product states, quantum states produced by low-depth circuits
6. Learning time-dependent states
7. $\vdots$

Hamiltonian Learning Problem

Learning Hamiltonians. Given Gibbs states of Hamiltonians, learn the Hamiltonian?

Problem definition. Let H be a κ-local Hamiltonian acting on n qudits 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

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

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 ρ to solve HLP is $\widetilde{\Theta}\big(\,\mathrm{poly}(e^{\beta+\kappa}, 1/\beta, 1/\varepsilon, n^3\big)$.

Quantum proof: First idea

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

Sufficient statistics:

1. Suppose we have approximations e_i' of

$$e_i = \mathrm{Tr}(E_i \rho_\mu) \quad \text{for all } i \in [m]$$

satisfying $|e_i' - e_i| \leq \varepsilon$, can we recover μ? Using [Aar'18, HKP'20, CW'20]?

2. These works produce $\rho' \approx \rho_\mu$, but that doesn't even imply ρ' is a Gibbs state $e^{-\beta H'}$, so approximating μ is unclear!

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

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

then $\rho_\lambda = \rho_\mu$ which implies $\lambda = \mu$. Isn't this "hard"?

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

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

where $S(\sigma) = -\mathrm{Tr}[\sigma \log \sigma]$ is the *quantum entropy* of σ. Optimum of (1) equals ρ_μ

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 μ

Maximum entropy principle: σ with equal marginals $\{e_i\}$ & maximum entropy is ρ_μ

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

$$\begin{aligned} \max_\sigma \quad & S(\sigma) \\ \text{s.t.} \quad & \mathrm{Tr}[\sigma E_i] = e_i, \quad \forall i \in [m] \\ & \sigma \succcurlyeq 0, \quad \mathrm{Tr}[\sigma] = 1. \end{aligned} \quad \xrightarrow{\textit{Approximations}} \quad \begin{aligned} \max_\sigma \quad & S(\sigma) \\ \text{s.t.} \quad & \mathrm{Tr}[\sigma E_i] = e_i', \quad \forall i \in [m] \\ & \sigma \succcurlyeq 0, \quad \mathrm{Tr}[\sigma] = 1. \end{aligned}$$

If ρ_μ maximizes first and $\rho_{\mu'}$ maximizes second problem, then by Pinsker's inequality

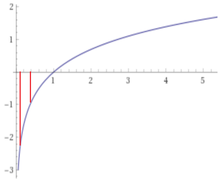
$$\|\rho_\mu - \rho_{\mu'}\|_1 \leq O(m\varepsilon)$$

Does this suffice for our problem in approximating the μs? **No**

In order to approximate μ, 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



Hamiltonian Learning algorithm

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

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

Algorithm.

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

$$\mu = \max_{\lambda_1,\ldots,\lambda_m} \log Z_\beta(\lambda) + \beta \sum_i \lambda_i e_i$$

   with errors

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

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

Hamiltonian Learning algorithm

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

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

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 ρ, our algorithm is classical
3. Exponential in β, κ: Might seem bad, but cannot be generically avoided
4. [HKT'22] considered small β, the sample complexity is $(\log n)/(\beta^2\varepsilon^2)$.