Learning from Quantum Data: Strengths and Weaknesses

Ronald de Wolf

joint with Srinivasan Arunachalam and others

CWI QuSoft Universiteit van Amsterdam

Quantum machine learning

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

	Classical learner	*Quantum learner*
Classical data	Classical ML	?
Quantum data	?	This talk

- We will look at the strengths and weaknesses of quantum learning from quantum examples (mostly supervised learning)

Supervised learning

- Concept: some function $f : \{0,1\}^n \to \{0,1\}$.
  Think of $x \in \{0,1\}^n$ as an object described by n "features", and concept f as describing a set of related objects
- **Goal**: learn f from a small number of examples: $(x, f(x))$

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

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

Making this precise: Valiant's "theory of the learnable"

- Concept class $\mathcal{C}$: set of concepts (small circuits, DNFs,...)
- Example for an unknown target concept $f \in \mathcal{C}$:
  $(x, f(x))$, where $x \sim$ unknown distribution D on $\{0,1\}^n$
- Goal: using some i.i.d. examples, learner for $\mathcal{C}$ should output hypothesis h that is probably approximately correct (PAC).

  h is a function of examples and of learner's randomness.

  Error of h w.r.t. target f: $\text{err}_D(f,h) = \Pr_{x\sim D}[f(x) \neq h(x)]$
- An algorithm (ε,δ)-PAC-learns $\mathcal{C}$ if:

$$\forall D \ \ \forall f \in \mathcal{C}: \quad \Pr[\quad \underbrace{\text{err}_D(f,h) \leq \varepsilon}_{h \text{ is approximately correct}} \quad] \geq 1-\delta$$

- A good learner has small time & sample complexity

Quantum data

- Much interesting quantum ML assumes classical data can be turned into quantum superposition.
  But in general this is expensive

- Let's try to circumvent the problem of putting classical data in superposition, by assuming we start from quantum data

- Bshouty-Jackson'95: suppose example is a superposition

$$\sum_{x\in\{0,1\}^n} \sqrt{D(x)}|x, f(x)\rangle$$

  Measuring this $(n+1)$-qubit state gives a classical example, so quantum examples are at least as powerful as classical

- Next slides: some cases where quantum examples are more powerful than classical for a fixed distribution D

Uniform quantum examples help some learning problems

- Quantum example under uniform D:

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

- Key subroutine: Fourier sampling (Bernstein-Vazirani'92): assume range of f is $\{\pm 1\}$. Can convert quantum example to

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

Hadamard transform turns this into $\displaystyle\sum_{s\in\{0,1\}^n}\widehat{f}(s)|s\rangle$,

$\widehat{f}(s)=\frac{1}{2^n}\sum_x f(x)(-1)^{s\cdot x}$ are the Fourier coefficients of f

- This allows us to sample s from distribution $\widehat{f}(s)^2$

Using Fourier sampling for learning

- If f is linear mod 2 $(f(x) = s \cdot x$ for one $s)$, then the Fourier distribution $\widehat{f}(s)^2$ is peaked at s.
  We can learn f from one quantum example!
- Bshouty-Jackson'95: learn Disjunctive Normal Form (DNF) formulas in poly-time: Fourier sampling + classical "boosting"
  Best known classical learner takes time $n^{O(\log n)}$
- Next slides: two new examples
  - Learning Fourier-sparse functions
  - Improving coupon collector

Learning Fourier-sparse Boolean functions

- $f : \{0,1\}^n \to \{\pm 1\}$ is k-Fourier-sparse if it has $\leq k$ non-zero Fourier coefficients
- Haviv-Regev'15:
  we can exactly learn such a function from $O(nk \log k)$ uniform samples $(x, f(x))$, and $\Omega(nk)$ samples are necessary
- Uniform quantum examples should be able to improve this.
  In particular, $k = 1$ is the special case of learning linear functions, where 1 quantum example suffices
- Next slide: learning f using $\widetilde{O}(k^{1.5})$ uniform quantum examples (Arunachalam-Chakraborty-Lee-dW'18)

Learning Fourier-sparse f from quantum examples

- Fourier span of f: $V = \text{span}\{s : \widehat{f}(s) \neq 0\}$.
  Let $r = \dim(V)$. Sanyal'15: $r = O(\sqrt{k}\log k)$
- Our learner:
  1. Fourier sample $O(rk)$ times. W.h.p.: span of the results $= V$.
     Now we can transform f by an $\mathbb{F}_2$-linear map M to a function $f_M : \{0,1\}^r \to \{\pm 1\}$
  2. Now use Haviv-Regev to learn f_M using $O(rk\log k)$ classical uniform examples (M converts examples between f and f_M).
     Transform f_M back to get f.

  Hence $\widetilde{O}(k^{1.5})$ quantum examples suffice for learning f exactly
- Lower bound: $\Omega(k\log k)$ quantum examples needed

Quantum superposition helps the coupon collector

- Coupon collector: sample uniformly from N elements. How many samples before you've seen each element at least once?

  Simple analysis:

  $\Pr[\text{see a new element} \mid \text{have already seen } i \text{ elements}] = \dfrac{N-i}{N}$

  $$\mathbb{E}[\#\text{samples}] = \sum_{i=0}^{N-1} \mathbb{E}[\#\text{samples to see } (i+1)\text{st element}]$$
  $$= \sum_{i=0}^{N-1} \frac{N}{N-i} = N \sum_{k=1}^{N} \frac{1}{k} \sim N \ln N$$

- Variation: sample uniformly from $[N] \backslash \{i\}$.

  How many samples before you know i? Still $\sim N \ln N$

- Suppose given superpositions instead of random samples.

  How many such quantum examples to learn i? $O(N)$ suffice!

Proof: use Pretty Good Measurement

- Define $|\psi_i\rangle = \left(\frac{1}{\sqrt{N-1}} \sum_{j \in [N] \setminus \{i\}} |j\rangle\right)^{\otimes T}$.
  Goal: do state identification on ensemble $\{|\psi_i\rangle, 1/N\}$
- Pretty good measurement has success probability at least square of the best-possible measurement (Barnum-Knill'02)
- Let $G_{i,j} = \frac{1}{N}\langle\psi_i|\psi_j\rangle$ be normalized Gram matrix of N states.
  Average success probability of PGM is $P_{PGM} = \sum_i (\sqrt{G}_{ii})^2$
  $\sqrt{G}$ is easy to compute here, can show $P_{PGM} \approx 1 - e^{-T/N}$.
  Setting $T = 2N$ gives $P_{PGM} \geq 2/3$
- Arunachalam-Childs-Kothari-dW'18: working on efficient implementation + tight analysis for all k, N

Ideally, we want our learner to work for all distributions D

- Remember Valiant's model:
  an algorithm (ε, δ)-PAC-learns concept class $\mathcal{C}$ if

$$\forall D \ \ \forall f \in \mathcal{C} : \quad \Pr[\quad \underbrace{\mathrm{err}_D(f,h) \leq \varepsilon}_{h \text{ is approximately correct}} \quad] \geq 1 - \delta$$

- We've seen examples where quantum examples help
  for a specific fixed D
- But in the PAC model, the learner has to succeed for all D
- Do quantum examples help also in this
  distribution-independent setting?

VC-dimension determines classical sample complexity

- Cornerstone of classical sample complexity: VC-dimension

  Set $S = \{s_1, \ldots, s_d\} \subseteq \{0,1\}^n$ is shattered by $\mathcal{C}$ if for all $a \in \{0,1\}^d$, there is $c \in \mathcal{C}$ s.t. $\forall i \in [d] : c(s_i) = a_i$

  VC-dim($\mathcal{C}$) $= \max\{d : \exists S$ of size d shattered by $\mathcal{C}\}$

- Equivalently, let M be the $|\mathcal{C}| \times 2^n$ matrix whose c-row is the truth-table of c. Then M contains complete $2^d \times d$ rectangle

- Blumer-Ehrenfeucht-Haussler-Warmuth'86: every (ε, δ)-PAC-learner for $\mathcal{C}$ needs $\Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ examples

- Hanneke'16: for every concept class $\mathcal{C}$, there exists an (ε, δ)-PAC-learner using $O\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ examples

Quantum sample complexity

Could quantum sample complexity be significantly smaller than classical sample complexity in the PAC model?

- Classical sample complexity is $\Theta\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$
- Classical upper bound carries over to quantum examples
- Atici & Servedio'04: lower bound $\Omega\left(\frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right)$
- Arunachalam & dW'17: tight bounds $\Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ quantum examples are necessary to learn $\mathcal{C}$

Hence in distribution-independent learning:
quantum examples are not significantly better than classical examples

Sketch of lower bound on quantum sample complexity

- Let $S = \{s_0, s_1, \ldots, s_d\}$ be shattered by $\mathcal{C}$.
  Define distribution D with $1 - 8\varepsilon$ probability on s_0,
  and $8\varepsilon/d$ probability on each of $\{s_1, \ldots, s_d\}$.
- ε-error learner takes T quantum examples and produces hypothesis h that agrees with $c(s_i)$ for $\geq \frac{7}{8}$ of $i \in \{1, \ldots, d\}$.
  This is an approximate state identification problem
- Take a good error-correcting code $E : \{0,1\}^k \to \{0,1\}^d$, with $k = d/4$, distance between any two codewords $> d/4$:
  approximating codeword $E(z) \Leftrightarrow$ exactly identifying $E(z)$
- We now have an exact state identification problem with 2^k possible states. Quantum learner cannot be much better than the Pretty Good Measurement, and we can analyze precisely how well PGM can do as a function of T.

  High success probability $\Rightarrow T \geq \Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$

Similar results for agnostic learning

- Agnostic learning: unknown distribution D generates examples (x, ℓ). We want to learn to predict bit ℓ from x. This allows to model situations where we only have "noisy" examples for target concept (maybe no fixed target exists)
- Best concept from $\mathcal{C}$ has error OPT $= \min_{c \in \mathcal{C}} \Pr_{(x,\ell)\sim D}[c(x) \neq \ell]$
- Goal of the learner: output $h \in \mathcal{C}$ with error $\leq \text{OPT} + \varepsilon$
- Classical sample complexity: $T = \Theta\left(\frac{d}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$

  NB: generalization error $\varepsilon = O(1/\sqrt{T})$, not $O(1/T)$ as in PAC
- Again, we show the quantum sample complexity is the same, by analyzing PGM to get optimal quantum bound

Summary & Outlook

- Quantum machine learning combines two great fields
- With classical data, you can get quadratic speed-ups for some ML problems, exponential speed-up under strong assumptions
  Biggest issue: how to put big classical data in superposition
- This talk: assume we start from data in superposition
- Positive result: for fixed distributions (e.g., uniform) quantum examples can be very helpful: learning linear functions, DNF, k-sparse functions, coupon collector
- Negative result: for distribution-independent learning (PAC and agnostic), quantum does not reduce sample complexity