QUANTUM SAMPLE COMPLEXITY LECTURE NOTES

FARIS SBAHI

We will mainly be talking about the works of Srinivasan Arunachalam Sourav Chakraborty Troy Lee Ronald de Wolf

1. Definitions

Let us first explain the setting of distribution-dependent learning from examples.

Definition 1.1. *(Concept class) Let $\mathcal{C}$ be a class of functions. For concreteness assume they are ± 1-valued functions on a domain of size $N = 2^n$.*

Example 1.2. *A halfspace is specified by a vector $w \in \mathbb{R}^p$ and decision rule*

$$f(x) = sgn(w \cdot x + b)$$

Definition 1.3. *(Shattering) Let $S = \{s_1, \cdots, s_d\} \subseteq \{0,1\}^n$ be data points on our domain. Over this set, we can consider all of its possible ± 1 labelings $\{a_1, a_2, \cdots, a_{2^d}\} \cong \{0,1\}^d$. If for each a_i there exists a $c_j \in \mathcal{C}$ s.t.*

$$c_j(x_k) = (a_i)_k, \forall k$$

then S is shattered by $\mathcal{C}$

Definition 1.4. *(VC Dimension) The size of the largest subset S of $\{0,1\}^n$ which is shattered by $\mathcal{C}$ is the VC dimension of $\mathcal{C}$*

Example 1.5. *The VC dimension of halfspaces in $\mathbb{R}^2$ is 3.*

The VC dimension of halfspaces in $\mathbb{R}^p$ is $p+1$. To see this, note that the definition of VC dimension asks us to find any dataset in our domain with size $p+1$. In which case, we can choose a dataset with points $\{e_i\}_{1 \leq i \leq d}$ where e_i is the standard basis vector with a 1 in the ith position and 0 elsewhere and $e_0 := 0$.

A halfspace is specified by a vector $w \in \mathbb{R}^p$ and decision rule

$$f(x) = sgn(w \cdot x + b)$$

Because f is a boolean function, we can decompose it in the Fourier basis and simply give positive Fourier coefficient to each e_i that S_i specifies has label $+1$ and negative coefficient otherwise. The bias b is determined by the label of $e_0 = 0$.

A boolean function is a function

$$f : \{0,1\}^n \to \{0,1\}$$

or by relabeling

$$f : \{-1,+1\}^n \to \{-1,+1\}$$

The domain of a boolean function is the "Hamming cube", $\mathbb{B}^n$.

Definition 1.6. *Let $S \subseteq [n]$ with $n \in \mathbb{Z}$. Then,*

(with $x^{\emptyset} = 1$ by convention)
$$\chi_S(x) := \prod_{i \in S} x_i$$

Hence, it is clear that we can view $\chi_S(x) : \{-1,+1\}^n \to \{-1,+1\}$ as computing the parity of x over the set S.

Theorem 1.7. *Every function $f : \{-1,+1\} \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial,*

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S(x)$$

We denote this expansion as the Fourier expansion of f, and term the real coefficent $\hat{f}(S)$ the Fourier coefficient of f on S.

Definition 1.8. *Define an inner product $\langle \cdot, \cdot \rangle$ on $f, g : \{-1,+1\} \to \mathbb{R}$ by*

$$\begin{aligned}\langle f, g \rangle &= 2^{-n} \sum_{x \in \{-1,+1\}^n} f(x)g(x) \\ &= E_{x \sim \{-1,+1\}^n}[f(x)g(x)]\end{aligned}$$

where $x \sim \{-1,+1\}^n$ denotes that x is chosen uniformly at random from $\{-1,+1\}^n$.

Theorem 1.9. *The set of all functions $f : \{-1,+1\}^n \to \mathbb{R}$ forms a vector space V such that $\dim V = 2^n$. The $\{\chi_S\}, S \subseteq [n]$ form an orthonormal basis for V.*

Definition 1.10. *(Bshouty and Jackson 1999) A quantum example oracle $QPEX(c, D)$ acts on $|0\rangle^{\otimes n} |0\rangle$ and produces a quantum example*

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

A quantum learner is given access to some copies of the state generated by $QPEX(c, D)$ and performs a POVM where each outcome is associated with a hypothesis.

Example 1.11.

$$\frac{1}{2^{n/2}} \sum_{x \in \{0,1\}^n} |x, a \cdot x\rangle$$

a small modification of the Bernstein-Vazirani algorithm [BV97] can recover a (and hence c) with probability 1/2

Lemma 1.12. *Let $f : \{0,1\}^n \to \{-1,1\}$. There exists a procedure that uses one uniform quantum example and satisfies the following: with probability $1/2$ it outputs an S drawn from the distribution*

$$\{\hat{f}(S)^2\}_{S \in \{0,1\}^n}$$

otherwise it rejects.

Proof. Using a uniform quantum example $\frac{1}{\sqrt{2^n}} \sum_x |x, f(x)\rangle$, one can obtain $\frac{1}{\sqrt{2^n}} \sum_x f(x) |x\rangle$ with probability $1/2$: unitarily replace $f(x)$ by $(1f(x))/2$, apply the Hadamard transform to the last qubit and measure it. With probability $1/2$ we obtain the outcome 0, in which case our procedure rejects. Otherwise the remaining state is $\frac{1}{\sqrt{2^n}} \sum_x f(x) |x\rangle$. Apply Hadamard transforms to all n qubits to obtain $\sum_S \hat{f}(S) |S\rangle$. Measuring this quantum state gives S with probability $\hat{f}(S)^2$ □

Definition 1.13. *(PAC model) A learning algorithm $\mathcal{A}$ is an (ϵ, δ)-PAC quantum learner for $\mathcal{C}$ if for every $c \in \mathcal{C}$ and distribution D, given access to the $QPEX(c, D)$ oracle, $\mathcal{A}$ outputs an h such that*

$$err_D(h, c) \leq \epsilon \tag{2}$$

with probability at least $1 - \delta$.

Definition 1.14. *We term $(\epsilon = 0, \delta)$-PAC learning as exact learning.*

Definition 1.15. *The sample complexity of $\mathcal{A}$ is the maximum number invocations of the $QPEX(c, D)$ oracle, maximized over all $c \in \mathcal{C}$, distributions D, and the learners internal randomness. The (ϵ, δ)-PAC quantum sample complexity of a concept class $\mathcal{C}$ is the minimum sample complexity over all (ϵ, δ)-PAC quantum learners for $\mathcal{C}$.*

2. Overview

In the classical case, the sample complexity of concept class $\mathcal{C}$ with VC dimension d in the PAC setting is

$$\Theta(\frac{d}{\epsilon} + \frac{\log(1/\delta)}{\epsilon})$$

($\epsilon \mapsto \epsilon^2$ in agnostic case)

where ϵ is the approximation coefficient and δ is the probability of success, as usual. We can show the same bound in the quantum PAC case.

Can also show similar bound using VC dimension comparing training error and test error (Vapnik 2000)

Notice that we are looking for a bound on an arbitrary distribution, with the oracle as specified, and notion of complexity given by the VC dimension. These are the three essential parameters one varies to look at sample complexity bounds.

Later: exact learning of k-sparse functions with uniform samples (we will define but if familiar k is the number of nonzero Fourier coefficients)

Classically $O(nk \log k)$ and $\Omega(nk)$

Quantum $O(k^{1.5}(\log k)^2)$ and (key) $\Omega(k \log k)$. Great for $k \ll n^2$.

Lemma 2.1. *Let $\mathcal{C}$ be a non-trivial concept class. For every $\delta \in (0, 1/2), \epsilon \in (0, 1/4)$, a (ϵ, δ)-PAC quantum learner for $\mathcal{C}$ has sample complexity $\Omega(\frac{1}{\epsilon} \log \frac{1}{\delta})$*

Proof. Since $\mathcal{C}$ is non-trivial, we may assume there are two concepts $c_1, c_2 \in C$ defined on two inputs $\{x_1, x_2\}$ as follows:

$$c_1(x_1) = c_2(x_1) = 0$$
$$c_1(x_2) = 0, c_2(x_2) = 1$$

Consider the distribution D such that

$$D(x_1) = 1 - \epsilon$$
$$D(x_2) = \epsilon$$

For $i \in \{1, 2\}$, the state of the algorithm after T queries to $QPEX(c_i, D)$ is

$$|\psi_i\rangle = (\sqrt{1-\epsilon}\,|x_1, 0\rangle + \sqrt{\epsilon}\,|x_2, c_i(x_2)\rangle)^{\otimes T}$$

Therefore, $\langle\psi_1|\psi_2\rangle = (1-\epsilon)^T$. Since the success probability of an (ϵ, δ)-PAC quantum learner is $\geq 1 - \delta$, Corollary **??** implies $\langle\psi_1|\psi_2\rangle \leq 2\sqrt{\delta(1-\delta)}$.

$\therefore T = \Omega(\frac{1}{\epsilon} \log \frac{1}{\delta})$ □

2.1. VC-dependent lower bounds.

Theorem 2.2. *Let C be a concept class with $dim_{VC}(\mathcal{C}) = d+1$. Then for every $\delta \in (0, 1/2)$ and $\epsilon \in (0, 1/4)$, every (ϵ, δ)-PAC learner for $\mathcal{C}$ has sample complexity $\Omega(\frac{d}{\epsilon} + \frac{\log(1/\delta)}{\epsilon})$.*

Consider an (ϵ, δ)-PAC learner for $\mathcal{C}$ that uses T examples. The d-independent part of the lower bound, $T = \Omega(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$, was proven in Lemma 2.1. Hence it remains to prove $T = \Omega(d/\epsilon)$.

It suffices to show this for a specific distribution D, defined as follows. Let $S = \{s_0, s_1, \cdots, s_d\} \subseteq \{0, 1\}^n$ be some $(d+1)$-element set shattered by $\mathcal{C}$. Define

$$D(s_0) = 1 - 4\epsilon \tag{3}$$

$$D(s_i) = 4\epsilon/d \tag{4}$$

for all $i \in [d]$.

Because S is shattered by $\mathcal{C}$, for each string $a \in \{0,1\}^d$, there exists a concept $c_a \in \mathcal{C}$ such that $c_a(s_0) = 0$ and $c_a(s_i) = a_i$ for all $i \in [d]$.

We essentially have two distributions of interest: the distribution which describes sampling the $\{s_i\}$ and the distribution which describes the uniform samples a. Hence, define A to be a random variable uniformly distributed over $\{0,1\}^d$. For fixed a, define $B = B_1 \otimes \cdots \otimes B_T$ as T i.i.d. quantum examples from $QPEX(c_a, D)$.

Hence, each B_i is the quantum sample

$$|\psi_a\rangle = \sum_{i \in \{0,1\}^d} \sqrt{D(s_i)}\, |i, c_a(s_i)\rangle$$

and so the AB bipartite system can be written as

$$\frac{1}{2^d} \sum_{a \in \{0,1\}^d} |a\rangle\langle a| \otimes |\psi_a\rangle\langle\psi_a|^{\otimes T}$$

To prove this theorem, we use the following lemmas.

Lemma 2.3.

$$I(A:B) \geq (1-\delta)(1 - H(1/4))d - H(\delta) = \Omega(d)$$

Proof. The learner will output a hypothesis $h(B) \in \{0,1\}^d$ based upon the observed samples B. Allow us to restrict our view to error on $s_1, \cdots, s_d$. In this case, the error of the hypothesis weighted by the distribution is simply the hamming distance $d_H(h(B), A)$ multiplied by $4\epsilon/d$:

$$err_D(h(B), A) = d_H(h(B), A)$$

where we are comparing with A because $c_a(A) = A$ on our restriction.

So, define indicator variable Z as indicating whether our restricted error is $\leq \epsilon$. Then, this is equivalent to requiring that

$$d_H(h(B), A) \leq d/4 \tag{5}$$

Since we are in search of a PAC learner, we require that $P(Z) \geq 1 - \delta$ which implies that Z has binary entropy

$$H_2(Z) \leq H_2(\delta)$$

If $h(B)$ satisfies (5), then A ranges over only $\sum_{i=0}^{d/4} \binom{d}{i} \leq 2^{H_2(1/4)d}$ (**??**) bit-strings given $h(B)$. Therefore,

$$H_2(A \mid h(B), Z = 1) \leq H_2(1/4)d$$

Furthermore, because $h(B)$ is assumed close to A we also have $H_2(A \mid B, Z = 1) \leq H_2(A \mid h(B), Z = 1)$. Now,

$$
\begin{aligned}
I(A:B) &= H(A) - H(A \mid B) \\
&\geq H(A) - H(A \mid B, Z) - H(Z) \\
&= H(A) - Pr(Z=1)H(A \mid B, Z=1) - Pr(Z=0)H(A \mid B, Z=0) - H(Z) \\
&\geq d - d(1-\delta)H(1/4) - d\delta - H(\delta) \\
&= \Omega(d)
\end{aligned}
$$

□

Lemma 2.4.

$$I(A:B) \leq T \cdot I(A:B_1)$$

Proof.

$$
\begin{aligned}
I(A:B) &= H(B) - H(B \mid A) \\
\text{(independence)} \quad &= H(B) - \sum_{i=1}^{T} H(B_i \mid A) \\
\text{(??)} \quad &\leq \sum_{i=1}^{T} H(B_i) - \sum_{i=1}^{T} H(B_i \mid A) \\
&= \sum_{i=1}^{T} I(A:B_i) \\
(I(A:B_i) = I(A:B_1), \forall i) \quad &= \sum_{i=1}^{T} I(A:B_1)
\end{aligned}
$$

□

Lemma 2.5.

$$I(A:B_1) = 4\epsilon \log(2d)$$

Proof. Since AB is a classical-quantum state, we have $I(A:B_1) = H(A) + H(B_1)H(AB_1) = H(B_1)$, where the first equality follows from definition and the second equality uses $S(A) = d$ since A is uniformly distributed in $\{0,1\}^d$, and $S(AB_1) = d$ since the matrix $\sigma = \frac{1}{2^d}\sum_{a\in\{0,1\}^d} |a\rangle\langle a| \otimes |\psi_a\rangle\langle\psi_a|$ is block diagonal with 2^d rank-1 blocks on the diagonal. It thus suffices to bound the entropy of the singular values of the reduced state of B_1, which is

$$\rho = \frac{1}{2^d} \sum_{a\in\{0,1\}^d} |\psi_a\rangle\langle\psi_a|$$

Let $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{2d} \geq 0$ be its singular values. Since ρ is a density matrix, these form a probability distribution. Note that the upper-left entry of the matrix $|\psi_a\rangle\langle\psi_a|$ is $D(s_0) = 1 - 4\epsilon$, hence so is the upper-left entry of ρ. This implies $\sigma_0 \geq 14\epsilon$.

Consider sampling a number $N \in \{0, 1, ..., 2d\}$ according to the σ-distribution. Let Z be the indicator random variable for the event $N \neq 0$, which has probability $1 - \sigma_0 \leq 4\epsilon$. Note that $H(N \mid Z = 0) = 0$, because $Z = 0$ implies $N = 0$. Also, $H(N \mid Z = 1) \leq \log(2d)$, because if $Z = 1$ then N ranges over $2d$ elements.

We now have

$$\begin{aligned} H(\rho) = H(N) = H(N,Z) &= H(Z) + H(N|Z) \\ &= H(Z) + Pr[Z=0]\cdot H(N|Z=0) + Pr[Z=1]\cdot H(N|Z=1) \\ &\leq H(4\epsilon) + 4\epsilon\log(2d) \\ \text{(Using the Taylor series of the logarithm)} \quad &= O(\epsilon\log(d/\epsilon)) \end{aligned}$$

□

Now, we are prepared to prove the theorem.

Proof. (Theorem 2.2) Combining these three lemmas, we have that

$$\begin{aligned} I(A:B) &\leq T\cdot I(A:B_1) \\ \Omega(\epsilon\log(d/\epsilon)) &= T\cdot 4\epsilon \\ T &= \Omega\Big(\frac{d}{\epsilon\log(d/\epsilon)}\Big) \end{aligned}$$

which comes close to optimal bound $\Omega(d/\epsilon)$ □

2.2. **State Identification for Optimal Bound.** Given a density matrix ensemble $\mathcal{E} = \{p_i, \sigma_i\}$ and a quantum state ρ we are promised that ρ is in state σ_i with probability p_i. In the general case we have $i \in [m]$ and of course $\sum_{i=1}^{m} p_i = 1$. Our goal is then to successfully identify which of the σ_i that our state ρ is actually in. This is known as Quantum Hypothesis Testing. Define $S = \sum_i \sigma_i$

$$E_i = S^{-1/2} p_i \sigma_i S^{-1/2}$$

for our original problem. Positive semidefiniteness is clear, so it remains to show that we have completeness

$$\begin{aligned} \sum_i E_i &= \sum_i S^{-1/2} p_i \sigma_i S^{-1/2} \\ &= S^{-1/2}\sum_i p_i\sigma_i S^{-1/2} \\ &= S^{-1/2} S S^{-1/2} = I \end{aligned}$$

Theorem 2.6. *Let* $\Pr_{opt}(\mathcal{E})$ *by the optimal success probability for our quantum hypothesis testing problem. Define* $\Pr_{PGM}(\mathcal{E})$ *to be the average success probability using the PGM POVM. Then,*

$$Pr_{opt}(\mathcal{E})^2 \leq Pr_{PGM}(\mathcal{E}) \leq Pr_{opt}(\mathcal{E})$$

Definition 2.7. *(Trace Distance)*

The trace distance $T(\cdot,\cdot)$ *is a metric on the space of density operators and gives a measure of distinguishability between states. In particular, let* ρ, σ *be density operators,*

$$\begin{aligned} T(\rho,\sigma) &= \frac{1}{2}\operatorname{Tr}\Big[\sqrt{(\rho-\sigma)^2}\Big] \\ &= \frac{1}{2}\sum_i |\lambda_i| \end{aligned}$$

where λ_i *are the eigenvalues of Hermitian* $\rho - \sigma$.

Hence, it is simply the trace norm of the positivization of the difference of matrices.

Proposition 2.8. *The maximum probability of distinguishing between two states with an optimal measurement is given by*

$$1/2[1 + T(\rho_1, \rho_2)]$$

In order to get rid of the logarithmic factor we then try another proof approach, which views learning from quantum examples as a quantum state identification problem: we are given T copies of the quantum example for some concept c and need to ϵ-approximate c from this. In order to render ϵ-approximation of c equivalent to exact identification of c, we use good linear error-correcting codes, restricting to concepts whose d-bit labeling of the elements of the shattered set $s_1, \cdots, s_d$ corresponds to a codeword. We then have $2^{\Omega(d)}$ possible concepts, one for each codeword, and need to identify the target concept from a quantum state that is the tensor product of T identical quantum examples.

We can use Pretty Good Measurement (PGM, also known as square root measurement) introduced by Hausladen and Wootters. The PGM is a specific measurement that one can always use for state identification, and whose success probability is no more than quadratically worse than that of the very best measurement. Authors use Fourier analysis to give an exact analysis of the average success probability of the PGM on the state-identification problems that come from both the PAC. This analysis could be useful in other settings as well. Here it implies that the number of quantum examples, T, is lower bounded by equations shown previously.

3. Exact learning lower bound

Assume for simplicity that k is a power of 2, so $\log k$ is an integer. We prove the lower bound for the following concept class, which was also used for the classical lower bound of Haviv and Regev [HR16]: let $\mathcal{V}$ be the set of distinct subspaces in $0,1^n$ with dimension $n \log k$ and

$$\mathcal{C} = \{c_V : \{0,1\}^n\{1,1\} : cV(x) = -1 \text{ iff } x \in V, \text{ where } V \in \mathcal{V}\}$$

Note that $|\mathcal{C}| = |\mathcal{V}|$, and each $c_V \in \mathcal{C}$ evaluates to 1 on a $(1 - 1/k)$-fraction of its domain.

We use almost the same approach as before... Let A be a random variable that is uniformly distributed over $\mathcal{C}$. Suppose $A = c_V$, then let $B = B_1 \cdots B_T$ be T copies of the quantum example

$$|\psi_V\rangle = \frac{1}{2^{n/2}} \sum_{x \in \{0,1\}^n} |x, c_V(x)\rangle$$

for c_V . The random variable B is a function of the random variable A. The following upper and lower bounds on $I(A : B)$ are similar to previous proof and we omit the details of the first two steps here.

(1) $I(A : B) \geq \Omega(\log |\mathcal{V}|)$ because B allows one to recover A with high probability.
(2) $I(A : B) \leq T \cdot I(A : B_1)$ using a chain rule for mutual information.

Lemma 3.1. $I(A : B_1) \leq O(n/k)$

Proof. Since AB is a classical-quantum state, we have

$$I(A : B1) = S(A) + S(B1)S(AB1) = S(B1),$$

where the first equality is by definition and the second equality uses $S(A) = log|V|$ since A is uniformly distributed over $\mathcal{C}$, and $S(AB_1) = \log |V|$ since the matrix

$$\sigma = \frac{1}{\mathcal{V}} \sum_{V \in \mathcal{V}} |V\rangle \langle V| \otimes |\psi_V\rangle \langle \psi_V|$$

is block-diagonal with $|V|$ rank-1 blocks on the diagonal. It thus suffices to bound the entropy of the (vector of singular values of) the reduced state of B_1, which is

$$\rho = \frac{1}{\mathcal{V}} |\psi_V\rangle \langle \psi_V|$$

Let $\sigma_0 \geq \sigma_1 \geq \cdot \geq \sigma_{2^{n+1}-1} \geq 0$ be the singular values of ρ. Since ρ is density matrix, these form a probability distribution. Now observe that $\sigma_0 \geq 1 - 1/k$ since the inner product between 1 n+1 2n x0,1n —x,1 and every —V is1$-1/k$. Let $N \in 0, 1, ..., 21$ be a random variable with probabilities $\sigma_0, \sigma_1, \cdots, \sigma_{2^{n+1}-1}$, and Z an indicator for the event $N \neq 0$ (note that $Z = 0$ with probability $\sigma_0 \geq 1 - 1/k$). By a similar argument as in before, we have

$$S(\rho) = H(N) = H(N, Z) = H(Z) + H(N \mid Z)$$
$$= H(\sigma_0) + \sigma_0 \cdot H(N \mid Z = 0) + (1 - \sigma_0) \cdot H(N \mid Z = 1) \leq H(1/k) + \frac{n+1}{k} \leq O(\frac{n + \log k}{k})$$

using $H(N \mid Z = 0) = 0$ in the first inequality, $H(\alpha) \leq O(\alpha \log(1/\alpha))$ in the second.
Combining these three steps implies $T = \Omega(k(\log |\mathcal{V}|)/n)$. It remains to lower bound $\mathcal{V}$ (not hard).
$\Omega(k \log k)$

4. Further developments

Quantum membership queries

$$O_c : |x, b\rangle \mapsto |x, b \cdot c(x)\rangle$$

for x of choice!
Noisy conditions
Other notions of function complexity
Conjecture by authors: bounds can be made tight