Quantum Predictive Learning and Communication Complexity with Single Input

Dmitry Gavinsky
NEC Laboratories America, Inc.
4 Independence Way, Suite 200
Princeton, NJ 08540, U.S.A.

Abstract

We define a new model of quantum learning that we call *Predictive Quantum (PQ)*. This is a quantum analogue of PAC, where during the testing phase the student is only required to answer a polynomial number of testing queries.

We demonstrate a relational concept class that is *efficiently learnable* in PQ, while in *any* "reasonable" classical model exponential amount of training data would be required. This is the first unconditional separation between quantum and classical learning.

We show that our separation is the best possible in several ways; in particular, there is no analogous result for a functional class, as well as for several weaker versions of quantum learning.

In order to demonstrate tightness of our separation we consider a special case of one-way communication that we call *single-input mode*, where Bob receives no input. Somewhat surprisingly, this setting becomes nontrivial when relational communication tasks are considered. In particular, any problem with two-sided input can be transformed into a single-input relational problem of equal *classical* one-way cost. We show that the situation is different in the *quantum* case, where the same transformation can make the communication complexity exponentially larger. This happens if and only if the original problem has exponential gap between quantum and classical one-way communication costs. We believe that these auxiliary results might be of independent interest.

1 Introduction

In this paper we compare quantum and classical modes of computational learning and give the first unconditional exponential separation between the two.

Let X be a (finite) domain and Y be a set of possible labels. Let $\mathcal{C}$ be a *concept class* consisting of functions $\ell : X \to Y$, each $\ell \in \mathcal{C}$ can be viewed as assignment of a label to every $x \in X$. The knowledge of X, Y and $\mathcal{C}$ is shared between a *teacher* and a *learner*; the teacher also knows some *target concept* $\ell_0 \in \mathcal{C}$, unknown to the learner. The learning process consists of two stages: the *learning phase*, followed by the *testing phase*. In the learning phase, the teacher and the learner communicate in order to let the latter learn ℓ_0. In the testing phase, the learner has to demonstrate that he has successfully learned ℓ_0: for example, an arbitrary $x \in X$ may be given to him, and he would have to respond with $\ell_0(x)$.

A *learning model* specifies the set of rules governing the learning and the testing phases. The teacher is, in general, viewed as an adversary that obeys the model's restrictions.

One of the most natural and widely used learning models is that of *Probably Approximately Correct (PAC)*, defined by Valiant [V84]. In the learning phase of PAC a sequence of training examples

$$(x_1, \ell_0(x_1)), \ldots, (x_k, \ell_0(x_k))$$

is sent by the teacher to the learner. The examples are independently chosen according to some distribution D over the domain X.[1] In the testing phase the learner is given a random $x \sim D$ and has to respond with $\ell_0(x)$.

[1] Several variations of PAC are studied in the literature, in particular there is a definition that allows "distribution-specific" learning algorithms. In this paper we will always fix D to be the uniform distribution over X, as that is sufficient for our purposes and simplifies the notation at the same time.

Two error parameters are present in the definition of PAC: *accuracy* $1-\varepsilon$ and *confidence* $1-\delta$. We say that learning was successful if in the testing phase the learner correctly labels a randomly chosen $x \sim D$ with probability at least $1-\varepsilon$. A learning algorithm must be successful with probability at least $1-\delta$, taken over both algorithm's randomness and the set of examples received during the learning phase.

We say that a concept class $\mathcal{C}$ is *efficiently learnable* in PAC if there exists an algorithm that runs in time at most polylogarithmic in the domain size and polynomial in $1/\varepsilon$ and $1/\delta$, and learns any $\ell \in \mathcal{C}$. Note that the running time of an algorithm is, trivially, an upper bound on the number of training examples that it uses during the learning phase.

1.1 Previous work

In [BJ95] Bshouty and Jackson introduced a natural quantum analogue of PAC, which we denote here by QAC. They gave an efficient algorithm that learns DNF formulas w.r.t. the uniform distribution from *quantum* examples – this is currently not known to be possible from classical examples (even with a quantum learning algorithm).

The question of whether quantum learning models are more efficient than the classical ones has been considered by Servedio and Gortler [SG04], who showed that the models PAC and QAC are equivalent from the information-theoretic point of view. On the other hand, they showed that quantum models are computationally more powerful than their classical analogues if certain cryptographic assumptions hold.

1.2 Our results

In the definition of a new learning model PQ (*Predictive Quantum*) we will generalize QAC in several ways.

First, we allow *relational* concept classes. Namely, the elements ℓ of $\mathcal{C}$ can be arbitrary subsets of $X \times Y$, thus allowing multiple correct labellings for every $x \in X$. During the learning phase the learner receives pairs (x_i, y_i), such that $x_i \sim D$ and y_i is a uniformly random element of $\{y | (x_i, y) \in \ell_0\}$. At the testing phase any y satisfying $(x, y) \in \ell_0$ is accepted as a correct answer to the query x.

Second, we classify all learning models as follows:

- We call *standard* a learning model where in the testing phase the learner outputs a *final hypothesis*, viewed as a function $h : X \to Y$. In the testing phase it is checked whether $h(x)$ agrees well with the target concept. The final hypothesis should be efficiently evaluatable (under the same notion of efficiency that applies to the learning algorithms in the model).
- We say that a model is *quasi-predictive* if the learner has to answer queries in the testing phase. The number of testing queries that will be asked is unknown during the learning phase.
- We call a model *predictive* if the learner should answer a single query in the testing phase.[2]

For example, the PAC model, as defined above, is predictive. If we would allow an arbitrary number of testing queries, that would make it quasi-predictive. If we require that in the end of the learning phase the learner produces a hypothesis $h : X \to Y$, such that $\mathbf{Pr}_{x \sim D}[h(x) = \ell(x)] \geq 1-\varepsilon$, that turns the model into standard.

As long as the learning phase remains unchanged, standard learnability of a concept implies its quasi-predictive learnability, which, in turn, implies predictive learnability. On the other hand, it is well known that in any "reasonable" *classical* learning model, a predictive learning algorithm can be turned into a standard one (this can be achieved by producing a final hypothesis consisting of a description of the answering subroutine, all the data available after the learning phase, and a random string, if randomness is used by the answering subroutine). Therefore, in the classical case the standard, the quasi-predictive, and the predictive modes of learning are essentially equivalent; in particular, the above three definitions of PAC give rise to the same family of efficiently learnable concept classes. We will see that the situation is different with quantum learning.

For the rest of the paper let $n \stackrel{\text{def}}{=} \lceil \log |X| \rceil$. Consider the following definition.

Definition 1 *Let D be a distribution over X. We say that a hypothesis $h : X \to Y$ approximates a concept $\ell \in \mathcal{C}$ w.r.t. D if*

[2] Note that a concept class that is efficiently learnable by our definition of predictive learning is also efficiently learnable in a version where *polynomial number* of testing queries are made. For notational convenience we will use the single-query definition of predictive learnability.

- $\mathbf{Pr}_{x\sim D}[h(x) = \ell(x)] \geq 2/3$, *when* $\ell : X \to Y$ *is a function;*
- $\mathbf{Pr}_{x\sim D}[(x, h(x)) \in \ell] \geq 2/3$, *when* $\ell \subseteq X \times Y$ *is a relation.*

A hypotheses class $\mathcal{H}$ is said to approximate $\mathcal{C}$ if for every $\ell \in \mathcal{C}$, $\mathcal{H}$ contains some h that approximates ℓ.

Any standard algorithm that learns $\mathcal{C}$ with $\varepsilon \leq 1/3$ must use a class of final hypotheses that approximates $\mathcal{C}$. An efficient algorithm can use a class of final hypotheses of size at most exponential in $\mathrm{poly}(n)$. As outlined above, efficient learnability in any classical model implies efficient learnability in the corresponding standard model, and therefore *$\mathcal{C}$ is efficiently learnable in some classical model only if there exists $\mathcal{H}$ of size at most $2^{\mathrm{poly}(n)}$ that approximates $\mathcal{C}$.*

We call a concept class $\mathcal{C}$ *unspeakable* if any class $\mathcal{H}$ that approximates it should be of size at least $2^{2^{\Omega(n)}}$. In particular, *neither a classical algorithm nor a standard quantum algorithm can efficiently learn an unspeakable concept class*.

In this paper we demonstrate an *efficient quantum predictive algorithm that learns an unspeakable relational concept class*. Therefore, *quantum predictive learnability does not imply quantum standard learnability*. On the other hand, we will show that *no quasi-predictive quantum algorithm can efficiently learn an unspeakable concept class*. We also show that *efficient quantum learning of a functional unspeakable concept class is impossible*, and therefore the combination of relational concepts and quantum predictive mode of learning is essential for learning an unspeakable class.

Following is a summary of our main results (cf. Theorem 7, Lemma 11, and Lemma 12).

Theorem 2 *There exists a relational concept class that is unspeakable but can be efficiently learned in the model of predictive quantum PAC.*

A concept class $\mathcal{C}$ that witnesses the above theorem is given in Definition 6. Its construction has been inspired by a communication problem due to Bar-Yossef, Jayram and Kerenidis [BJK04].

Theorem 3 *Classical learning of an unspeakable concept class is not possible from less than exponential amount of information from the teacher, even by a computationally unlimited learner.*

Both standard and quasi-predictive learning of an unspeakable concept class is not possible from less than exponential amount of quantum (w.l.g.) information from the teacher, even by a computationally unlimited learner.

Predictive learning of an unspeakable functional concept class is not possible from less than exponential amount of quantum (w.l.g.) information from the teacher, even by a computationally unlimited learner.

Two parts of Theorem 3 are proved by making connection to two "impossibility of separation" results in communication complexity. One of them is due to Aaronson [A04], and the other is new and might be of independent interest.

We will consider a special case of one-way communication, which will we call *single-input mode*, where Bob receives no input. We show that, somewhat surprisingly, for any single-input communication task the quantum and the classical one-way costs are asymptotically the same (the statement is trivial for functional tasks, but the relational case is more involved). More details can be found in Section 4.2.

2 Definitions and more

For $a \in \mathbb{N}$ we denote $[a] \stackrel{\text{def}}{=} \{1, \ldots, a\}$. We view the elements of $\mathbb{Z}_a$ as integers $\{0, 1, , \ldots, a-1\}$, and accordingly we define their ordering $0 < 1 < \cdots < a-1$. For any $i \in \mathbb{N}$ and $b \in \mathbb{Z}_a$, let $i \cdot b = ib$ be the i'th power of b w.r.t. the group operation $+$.

We use subscripts to address individual bits of binary strings: for $x \in \{0,1\}^n$ and $i \in [n]$, x_i stands for the i'th bit of x.

Let D be the uniform distribution over X, recall Definition 1.

Definition 4 *Let $\mathcal{C}$ be a concept class. We say that $\mathcal{C}$ is unspeakable if $|\mathcal{C}'| \in 2^{2^{\Omega(n)}}$ holds for any $\mathcal{C}'$ that approximates $\mathcal{C}$ w.r.t. D.*

2.1 Quantum learning

In [SG04] the authors provide an excellent survey of quantum vs. classical learning. Below we sketch one possible intuition behind the concepts considered in this work.[3]

Starting from PAC, how can we make it quantum? First, we can give the student ability to run any computation that a quantum computer can perform efficiently (e.g., to decide membership in any language from the complexity class BQP). Second, we can let the training examples be quantum, i.e., the student receives from the teacher *quantum bits (qubits)*. In this paper we consider the situation when *both the student and the examples are quantum*.

Information-theoretic consequences of "quantumness" stem from the facts that, on the one hand, quantum states require *exponential* (in the number of qubits) amount of classical bits for their full description, while on the other hand, the uncertainty principle dictates that given a quantum state only a (tiny) fraction of that classical data can be accessed by an observer.[4]

Note also that computational impact of a student being quantum is not necessarily captured by the power of BQP: As training examples are quantum, the student can apply quantum algorithms to quantum input, while BQP only deals with situations when quantum algorithms are fed with classical input.

What can be viewed as a reasonable model of quantum training examples? Let the target concept be ℓ_0. First, assume that $\ell_0 : X \to Y$ is a Boolean function, then a quantum example shall look like

$$\frac{1}{\sqrt{|X|}} \sum_{x \in X} |x, \ell_0(x)\rangle \,,$$

where $|\cdot\rangle$ denotes the corresponding basis state[5] of the quantum register over $n+1$ qubits. Note that the above form of training examples corresponds to the uniform distribution of $x \in X$, since measuring the first n qubits in the computational basis can return each possible $x_0 \in X$ with the same probability of $1/|X|$.

Now, let $\ell_0 \subseteq X \times Y$ be a relation. We need a quantum superposition over all possible pairs $(x, y) \in \ell_0$. Naturally, we want to choose the amplitudes such that every x_0 still shows up with probability $1/|X|$, and at the same time, conditional on obtaining x'_0, every element of $\{y | (x'_0, y) \in \ell_0\}$ appears with equal probability. It can be seen that the following quantum superposition satisfies the requirements:

$$\sum_{(x,y) \in \ell_0} \frac{1}{\sqrt{|X| \cdot |\{y' | (x, y') \in \ell_0\}|}} |x, y\rangle \,.$$

This quantum state will be used in Definition 5 below to describe the training examples that our student will receive from the teacher.

2.2 The model of predictive quantum learning

We will usually ignore normalization factors and global phases of quantum states.[6] We define a predictive quantum version of PAC, as follows.

Definition 5 *In the PQ (Predictive Quantum) learning model, a learning algorithm can ask for arbitrarily many copies of the state*

$$\sum_{(x,y) \in \ell_0} \frac{1}{\sqrt{|\{y' | (x, y') \in \ell_0\}|}} |x, y\rangle \,,$$

[3] This part is mostly meant to assist a reader whose familiarity with quantum computing is limited; analyzing the philosophical foundations of quantum mechanics is beyond our scope.

[4] To visualize the uncertainty principle, consider a classical observer who wants to measure a quantum particle moving with velocity $v(t)$ and taking position $x(t)$, as functions of time t. At the moment t_0 it is possible to measure $v(t_0)$ with high accuracy, but then $x(t_0)$ can only be determined very roughly; alternatively, it is possible to measure $x(t_0)$ with high accuracy, but that leaves $v(t_0)$ with large uncertainty. The "quantum catch" here is that the uncertainty does not result from any kind of "imperfection" in the measurement devices being used, but rather constitutes one of several fundamental principles of quantum mechanics.

[5] This is Dirac's "bra-ket" notation: *Ket*s ($|\cdot\rangle$) denote unit vectors corresponding to pure quantum states, and *bra*s ($\langle\cdot|$) stand for the complex conjugates of *ket*s. Naturally, $\langle\cdot|\cdot\rangle$ and $|\cdot\rangle\langle\cdot|$ are, respectively, the inner and the outer products of two vector operands.

[6] That is, we allow arbitrary non-zero complex vectors to represent quantum states.

where $\ell_0 \subseteq X \times Y$ is a relational target concept. In the end of the learning process the algorithm receives an element $x \in X$ and should, with probability at least $5/6$, output any y satisfying $(x, y) \in \ell_0$.

A learning algorithm is efficient if its running time is at most polynomial in $n \stackrel{def}{=} \lceil \log |X| \rceil$. A concept class $\mathcal{C}$ is efficiently learnable in PQ if there exists an efficient algorithm that PQ-learns every $\ell \in \mathcal{C}$.

In the above definition the relative amplitudes of the pairs $|x, y\rangle$ in a training example are chosen such that a projective measurement in the computational basis would result in a uniformly chosen x, and given x, all elements of $\{y' | (x, y') \in \ell_0\}$ are equally likely to come with it. Therefore, the model can be viewed as a natural quantum generalization of the relational version of PAC, as discussed in the Introduction.

The fact that all quantum training examples are the same lets us get rid of the confidence parameter (δ) in the definition of PQ (there is no such thing as "unlucky" sample of training examples). For simplicity, we choose the required accuracy (ε) to always be $5/6$. Note also that in the testing phase we want the learning algorithm to give a correct answer to *any* $x \in X$ with good probability (instead of just being able to cope with a randomly chosen x). This further simplifies the definition and also makes our result stronger (as we construct a PQ-algorithm, and do not state any lower bound against this model).

3 Concept class $\mathcal{C}$

We define a concept class $\mathcal{C}$ that will be shown to be both unspeakable and efficiently PQ-learnable. Our definition has been inspired by a communication problem considered in [BJK04].

Definition 6 *Let N be prime. Every concept in the class $\mathcal{C}$ is represented by $C \in \{0,1\}^N$. The set of queries is $[N-1]$, represented by binary strings of length $n = \lceil \log N \rceil$. A pair $(x, b) \in \mathbb{Z}_N \times \{0, 1\}$ is a valid answer to query j w.r.t. $C \in \mathcal{C}$ if $C_x \oplus C_{x+j} = b$.*

We slightly abuse the notation by viewing each $C \in \mathcal{C}$ either as a binary string of length N or as a set $\{(j, x, b) | (x, b)$ is a valid answer to j w.r.t. $C\}$.

Theorem 7 *The concept class $\mathcal{C}$ is unspeakable. On the other hand, $\mathcal{C}$ is efficiently learnable in PQ.*

The two parts of the theorem will be proved in Sections 3.1 and 3.2, respectively. The key observation that we use to efficiently learn $\mathcal{C}$ is the following (originating from [KW04]). Let a binary string $x \in \{0,1\}^n$ be represented as a quantum state $|\alpha(x)\rangle = \sum(-1)^{x_i} |i\rangle$, where i ranges in $[n]$. Even though it is impossible to recover individual bits of x by measuring $|\alpha(x)\rangle$, there is something nontrivial about x that can be learned from $|\alpha(x)\rangle$. Namely, given any perfect matching M over $[n]$, it is possible to measure $|\alpha(x)\rangle$ in such a way that for some $(i, j) \in M$ the value of $x_i \oplus x_j$ would become known after the measurement. The quantum state $|\alpha(x)\rangle$ fits in $\lceil \log n \rceil$ qubits; on the other hand, it can be shown that the amount of classical information needed to allow similar type of access to x is $n^{\Omega(1)}$, and this is used to show that $\mathcal{C}$ is unspeakable.

3.1 Efficient PQ-learning of $\mathcal{C}$

Our learner will need k PQ-examples in order to answer to the testing query with probability $1 - 1/2^k$, and whenever an answer is given it is correct.[7] Fix $C \in \mathcal{C}$, then the training examples are of the form

$$\left|\alpha^C\right\rangle \stackrel{\text{def}}{=} \sum_{(j,x,i) \in C} |j, x, i\rangle .$$

The learner measures the last register of each of the k instances of $\left|\alpha^C\right\rangle$ in the basis $\{|0\rangle + |1\rangle, |0\rangle - |1\rangle\}$. With probability $1 - 1/2^k$ at least one measurement results in $|0\rangle - |1\rangle$, then the learner keeps that copy and abandons the rest (otherwise he gives up). Next, the learner measures the second register in the computational basis, thus obtaining in the first two registers

$$\sum_{(j,x_0,i) \in C} (-1)^i |j, x_0\rangle = \sum_{j \in [N-1]} (-1)^{C_{x_0} \oplus C_{x_0+j}} |j, x_0\rangle = \sum_{j \in [N-1]} (-1)^{C_{x_0+j}} |j, x_0\rangle$$

[7] If we allow a slightly modified form of training examples, where i is represented through the amplitude as $\sum_{(j,x,i) \in C} (-1)^i |j, x\rangle$, then it is possible to PQ-learn $\mathcal{C}$ *exactly* from one such example.

for some $x_0 \in \mathbb{Z}_N$. Then he performs the transformation $|j, x_0\rangle \to |j + x_0, x_0\rangle$, and the state of the first register becomes

$$\left|\alpha_{x_0}^C\right\rangle \stackrel{\text{def}}{=} \sum_{j\in[N-1]} (-1)^{C_{x_0+j}} |x_0 + j\rangle = \sum_{k\in\mathbb{Z}_N\setminus\{x_0\}} (-1)^{C_k} |k\rangle .$$

At this point the learner is ready for the testing phase. Assume that a question $q \in [N-1]$ has been asked. Define the following perfect matching over $\mathbb{Z}_N \setminus \{x_0\}$:

$$m_q \stackrel{\text{def}}{=} \left\{ \big(x_0 + (2i+1)q, x_0 + (2i+2)q\big) \middle| 0 \le i \le \frac{N-3}{2} \right\}.$$

Pairwise disjointness of the edges and the fact that x_0 is isolated follow from primality of N. The learner performs projective measurement of $\left|\alpha_{x_0}^C\right\rangle$ onto $(N-1)/2$ subspaces, each spanned by a pair of vectors $|a\rangle$ and $|b\rangle$ where a and b are connected in m_q (to make the measurement complete we add $|x_0\rangle\langle x_0|$ to it, but this outcome never occurs).

Assume that the outcome of the last measurement corresponds to the edge $(a, a+q) \in m_q$. Then the state of the register that contained $\left|\alpha_{x_0}^C\right\rangle$ becomes either $|a\rangle + |a+q\rangle$ or $|a\rangle - |a+q\rangle$, the former corresponding to $C_a \oplus C_{a+q} = 0$ and the latter to $C_a \oplus C_{a+q} = 1$. As the two states are orthogonal, the learner is able to distinguish and, respectively, answer $(a, 0)$ in the first case and $(a, 1)$ in the second, and that is a correct answer.

All quantum operations involved in the algorithm can be performed efficiently.

3.2 $\mathcal{C}$ is unspeakable

Let us see that the concept class $\mathcal{C}$ is unspeakable. The following proof uses some ideas from [BJK04] and [GKRW06].

Assume that $\mathcal{C}$ is approximated by a class $\mathcal{D}$. Then there exists some $h_0 \in \mathcal{D}$ that simultaneously approximates at least $2^N / |\mathcal{D}|$ elements of $\mathcal{C}$, denote the set of those elements by $\mathcal{C}_0$.

Consider the answers that h_0 gives to all possible queries $q \in [N-1]$. Denote $(x_q, i_q) \stackrel{\text{def}}{=} h_0(q)$ and let

$$Q_0 \stackrel{\text{def}}{=} \big\{ q \big| (x_q, i_q) \text{ is a good answer to } q \text{ w.r.t. at least } 3/5\text{'th of } \mathcal{C}_0\text{'s elements} \big\} .$$

Counting reveals that $|Q_0| \ge \frac{N-1}{6}$.

Let $e_q \stackrel{\text{def}}{=} (x_q, x_q + q)$ and $E_0 \stackrel{\text{def}}{=} \big\{ e_q \big| q \in Q_0 \big\}$. Every edge e_q corresponds to at most 2 different values of $q \in [N-1]$, therefore $|E_0| \ge \frac{N-1}{12}$. Consider a graph G_0 over N nodes, whose edges are the elements of E_0. Observe that G_0 contains at least $\sqrt{2\,|E_0|} \ge \sqrt{\frac{N-1}{6}}$ non-isolated vertices.

Let $F_0 \subseteq G_0$ be a forest consisting of a spanning tree for each connected component of G_0. Then F_0 contains at least $\sqrt{\frac{N-1}{24}}$ edges, denote them by E_0'. Let $Q_0' \subseteq Q_0$ be a subset of size $|E_0'|$, such that

$$E_0' = \big\{ e_q \big| q \in Q_0' \big\} .$$

View the elements of $\mathcal{C}$ as binary strings of length N. Let us consider two probability distributions, one corresponding to uniformly choosing $C \in \mathcal{C}$ and the other corresponding to uniformly choosing $C \in \mathcal{C}_0$ – denote them by D^C and D_0^C, respectively. Then

$$\log\left(\frac{|\mathcal{C}|}{|\mathcal{C}_0|}\right) = \mathbf{H}\left[D^C\right] - \mathbf{H}\left[D_0^C\right] ,$$

where $\mathbf{H}[\cdot]$ denotes the binary entropy.

For every $e_q = (a, b)$ put $I_q \stackrel{\text{def}}{=} C_a \oplus C_b$, and let $J \stackrel{\text{def}}{=} (I_q)_{q\in Q_0'}$. It is straightforward from the construction of Q_0' that if $C \sim D^C$ then the collection $\big\{ I_q \big| q \in Q_0' \big\}$ consists of mutually independent unbiased Boolean random variables, and therefore $\mathbf{H}_{D^C}[J] = |Q_0'|$.

As $\mathbf{H}[C] = \mathbf{H}[J] + \mathbf{H}\left[C\middle|J\right]$ holds w.r.t. any distribution of C,

$$\begin{aligned}\log\left(\frac{|\mathcal{C}|}{|\mathcal{C}_0|}\right) &= \underset{D^C}{\mathbf{H}}[C] - \underset{D_0^C}{\mathbf{H}}[C] = \underset{D^C}{\mathbf{H}}[J] - \underset{D_0^C}{\mathbf{H}}[J] + \underset{D^C}{\mathbf{H}}\left[C\middle|J\right] - \underset{D_0^C}{\mathbf{H}}\left[C\middle|J\right] \\ &\geq \underset{D^C}{\mathbf{H}}[J] - \underset{D_0^C}{\mathbf{H}}[J] = |Q_0'| - \underset{D_0^C}{\mathbf{H}}[J] \\ &\geq |Q_0'| - \sum_{q\in Q_0'} \underset{D_0^C}{\mathbf{H}}[I_q] = \sum_{q\in Q_0'}\left(1 - \underset{D_0^C}{\mathbf{H}}[I_q]\right),\end{aligned} \tag{1}$$

where the first inequality follows from the fact that $\mathbf{H}_{D^C}\left[C\middle|J\right] = N - |Q_0'|$, which is the maximum of $\mathbf{H}\left[C\middle|J\right]$ under any distribution of C.

From the definition of Q_0 (and the fact that $Q_0' \subseteq Q_0$), we know that each of $\left\{I_q\middle|q \in Q_0'\right\}$ is at least $3/5$-biased, therefore $\mathbf{H}_{D_0^C}[I_q] \leq \frac{49}{50}$, and (1) leads to

$$\log\left(\frac{|\mathcal{C}|}{|\mathcal{C}_0|}\right) \geq \frac{|Q_0'|}{50} = \frac{|E_0'|}{50} > \frac{\sqrt{N}}{250},$$

for sufficiently large N. According to our choice of h_0,

$$|\mathcal{D}| \geq \frac{|\mathcal{C}|}{|\mathcal{C}_0|} \in 2^{N^{\Omega(1)}} \subseteq 2^{2^{\Omega(n)}},$$

which means that the class $\mathcal{C}$ is unspeakable.

4 Optimality of our separation

The model of PQ where we demonstrated learnability of $\mathcal{C}$ is computationally feasible. But in the definition of PQ we have modified what is probably the most usual learning setting in several ways: Besides being quantum, our algorithm is *predictive*; moreover, the concept class that we learn is a *relational* one. In this section we will see that all these "enhancements" are essential in order to be able to learn an unspeakable class efficiently.

We already know that classical learning of an unspeakable class cannot be efficient. We will show that exponential amount of training data is required in order to learn a functional unspeakable concept (Lemma 11), as well as to learn any unspeakable concept in quasi-predictive setting (Lemma 12). The both results are established through making a connection to one-way communication complexity: Our proof of Lemma 11 is based on Aaronson's [A04], and in order to prove Lemma 12 we establish a new fact about one-way communication complexity that might be of independent interest (Theorem 9, Corollary 10).

4.1 Quantum and classical one-way communication complexity

The one-way model of communication complexity is defined as follows. Let $P \subseteq X \times Y \times Z$ be a (relational) two-party communication problem. Input to P has the form $(x, y) \in X \times Y$, in the beginning it is split between two players: Alice receives x and Bob receives y. The goal is for Bob to produce $z \in Z$, such that $(x, y, z) \in P$. The players cooperate to achieve it, namely Alice sends a message m to Bob, and he outputs $z \in Z$ based on the message m and his portion of input y.

Assume for convenience that both the length of y and the length of m are functions of the lengths of x, and denote the latter by $n = \lceil \log |X| \rceil$. Both Alice and Bob are all-powerful computationally, and their goal is to solve the problem using as short m as possible. There are two versions of this model that we are interested in, namely *quantum* and *classical*. In the former the action of the players should obey the laws of quantum mechanics, in particular the message m is quantum and its "length" is measured in qubits; in the latter the message is classical and consists of bits. We let our protocols employ mixed strategies, i.e., shared randomness is allowed.

For any ε we say that *a protocol $\mathcal{T}$ solves P with error ε* if Alice and Bob, who behave according to $\mathcal{T}$, produce a correct answer to every input $(x, y) \in X \times Y$ with probability at least $1 - \varepsilon$. For a distribution μ over $X \times Y$ we say that *$\mathcal{T}$ solves P with error ε w.r.t. μ* if a correct answer is produced with probability at least $1 - \varepsilon$ when $(x, y) \sim \mu$. The *ε-error communication cost* of P is the smallest possible message length of a protocol that solves P with error ε, and *ε-error communication cost w.r.t. μ* is defined similarly. We say that the *bounded-error cost* of P is at most k if for any $\varepsilon \in \Omega(1)$ its ε-error cost is at most k.

Denote by $\mathcal{R}^{1}_{\varepsilon}(P)$ ($\mathcal{R}^{1}_{\mu,\varepsilon}(P)$) the classical one-way ε-error communication cost of P (w.r.t. μ), and by $\mathcal{R}^{1}(P)$ its bounded-error classical cost. Denote by $\mathcal{Q}^{1}_{\varepsilon}(P)$, $\mathcal{Q}^{1}_{\mu,\varepsilon}(P)$ and $\mathcal{Q}^{1}(P)$ the corresponding quantum analogs.

An important special case of relational communication problems are *functional* problems (partial or total). The following theorem follows readily from Theorem 6 of [A04]:

Theorem 8 [A04] *For any functional two-party communication problem $F : X \times Y \to Z$, it holds that $\mathcal{R}^{1}(F) \in O\left(\log(|Y|)\mathcal{Q}^{1}_{\varepsilon}(F)\log\mathcal{Q}^{1}_{\varepsilon}(F)\right)$ for any $\varepsilon < 1/2 - \Omega(1)$.*

4.2 One-way communication when Bob receives no input

In this section we present a new result in communication complexity, it will be used later to prove Lemma 12.

Consider a special case of one-way communication that we call *single-input mode*, where Bob receives no input. Denote $\mathbf{0} \stackrel{\text{def}}{=} \{0\}$, and let $P \subseteq X \times \mathbf{0} \times Z$ be a communication task where Alice receives x and sends a single message m to Bob, who has to output $z \in Z$ based on the message m alone.

This setting is not as trivial as it may appear at first glance.[8] For instance, any communication problem with two-sided input $P \subseteq X \times Y \times Z$ has a single-input analogue $P' \subseteq X \times \mathbf{0} \times Z^{Y}$, where Bob has to produce a list of answers to the original P w.r.t. all $y \in Y$. Namely, let

$$P'_{\mu,\varepsilon} \stackrel{\text{def}}{=} \left\{ (x, 0, (z_y)_{y\in Y}) \middle| \underset{y\sim\mu_x}{\mathbf{Pr}}\left[(x,y,z_y) \in P\right] \geq \varepsilon \right\},$$

where μ is a distribution on $X \times Y$ and μ_x is the marginal distribution of B when $(A,B) \sim \mu$ and $A = x$. Note that for any μ and $\varepsilon \in \Omega(1)$, $\mathcal{R}^{1}(P'_{\mu,\varepsilon}) \leq \mathcal{R}^{1}(P)$, and on the other hand, by the Minimax theorem $\mathcal{R}^{1}(P) = \sup\left\{\mathcal{R}^{1}(P'_{\mu,\varepsilon})\right\}$, where the supremum is taken w.r.t. all possible μ and $\varepsilon \in \Omega(1)$.

In other words, $P'_{\mu,\varepsilon}$ is essentially as difficult to solve in the model of one-way *classical* communication as P is. Somewhat surprisingly, the same is not true in the case of quantum communication. More generally, below we show that for any single-input communication task the quantum and the classical one-way costs are asymptotically the same. In particular, this means that $\mathcal{Q}^{1}(P)$ can be exponentially smaller than $\mathcal{Q}^{1}(P'_{\mu,\varepsilon})$ for some $\varepsilon \in \Omega(1)$ – this happens if and only if the gap between $\mathcal{Q}^{1}(P)$ and $\mathcal{R}^{1}(P)$ is exponential (examples of such P were given in [BJK04], [GKKRW07]).

Theorem 9 *For any relational two-party communication problem $P \subseteq X \times \mathbf{0} \times Z$, any distribution μ over $x \in X$ and any $\Omega(1) < \varepsilon < 1 - \Omega(1)$, it holds that $\mathcal{R}^{1}_{\mu,\varepsilon}(P) \in O\left(\mathcal{Q}^{1}_{\mu,\varepsilon}(P)\right)$.*

Corollary 10 *For any $P \subseteq X \times \mathbf{0} \times Z$, it holds that $\mathcal{R}^{1}(P) \in O\left(\mathcal{Q}^{1}(P)\right)$.*

Proof: By the Minimax theorem, for every ε there exists μ such that $\mathcal{R}^{1}_{\varepsilon}(P) = \mathcal{R}^{1}_{\mu,\varepsilon}(P)$. ■

If P is a function then Corollary 10 is a very trivial special case of Theorem 8. On the other hand, Corollary 10 applies to the much more general case of relational problems, where a statement analogous to Theorem 8 provably does not hold.

Proof:(Theorem 9) Let $\mathcal{W}$ be a valid $\mathcal{Q}^{1}_{\mu,\varepsilon}$-protocol of cost m for P, i.e., $\mathcal{W}$ guarantees error at most ε w.r.t. $x \sim \mu$. We want to build an $\mathcal{R}^{1}_{\mu,\varepsilon}$-protocol of cost $O(m)$.

Let A and B be random variables taking the value of Alice's input $x \in X$ and Bob's answer $z \in Z$, respectively. Assume $A \sim \mu$ and let μ^{B} be the corresponding distribution of B. Conditional upon $A = x$ let $B \sim \mu^{B}_{x}$. Define a random variable B' as a "refined version" of B, namely: if $A = x$ then the conditional distribution of B' is

$$\mu^{B'}_{x}(z) \stackrel{\text{def}}{=} \begin{cases} \mu^{B}_{x}(z)/(1-\varepsilon_x) & \text{if } (x,0,z) \in P \\ 0 & \text{otherwise} \end{cases},$$

where $1 - \varepsilon_x$ is the probability that $\mathcal{W}$ returns a correct answer on input x.

By the Holevo bound and the information processing principle,

$$m \geq \mathbf{I}[A:B] = \underset{A=x}{\mathbf{E}}\left[d_{KL}\left(\mu^{B}_{x}\middle\|\mu^{B}\right)\right].$$

[8] It is important that we consider relational problems, for functions the single-input mode is indeed uninteresting.

For every $x \in X$,

$$\begin{aligned} d_{KL}\left(\mu_x^{B'}\middle\|\mu^B\right) &= \sum_z \mu_x^{B'}(z)\log\frac{\mu_x^{B'}(z)}{\mu^B(z)} \\ &\leq \frac{1}{1-\varepsilon_x}\sum_z \mu_x^B(z)\log\left(\frac{\mu_x^B(z)}{\mu^B(z)}\cdot\frac{1}{1-\varepsilon_x}\right) \\ &\leq \frac{1}{1-\varepsilon}\cdot d_{KL}\left(\mu_x^B\middle\|\mu^B\right) + \frac{1}{1-\varepsilon}\log\frac{1}{1-\varepsilon}. \end{aligned}$$

By linearity of expectation,

$$\mathop{\mathbf{E}}_{A=x}\left[d_{KL}\left(\mu_x^{B'}\middle\|\mu^B\right)\right] \leq \frac{m}{1-\varepsilon} + \frac{1}{1-\varepsilon}\log\frac{1}{1-\varepsilon} < \frac{2m}{1-\varepsilon}, \tag{2}$$

for sufficiently large m.

We claim that there exists an $\mathcal{R}^1_{\mu,\varepsilon}$-protocol for P of cost $\left\lceil \frac{11m}{\varepsilon(1-\varepsilon)} \right\rceil$. By the definition of B', any z in the support of $\mu_x^{B'}$ is a correct answer to $x \in X$. The key observation is that $\mu_x^{B'}$ is not too far from μ^B, by (2). Therefore, if Alice and Bob sample sufficiently many elements from μ^B, with high probability at least one of them would belong to the support of $\mu_x^{B'}$. Such sampling can be performed by the players locally, using shared randomness. Then Alice can send a pointer to an element which is a good answer w.r.t. her x.

Let us estimate the probability that a randomly chosen $z \sim \mu^B$ satisfies $\mu_x^{B'}(z) > 0$. Let

$$X' \stackrel{\text{def}}{=} \left\{ x \in X \middle| d_{KL}\left(\mu_x^{B'}\middle\|\mu^B\right) < \frac{5m}{\varepsilon(1-\varepsilon)} \right\},$$

then it follows from (2) that $\mu(X') > 1 - \varepsilon/2$. Fix any $x_0 \in X'$ and let $Z' \stackrel{\text{def}}{=} \left\{ z \middle| \mu_{x_0}^{B'}(z) > 0 \right\}$. From

$$\sum_{z\in Z'} \mu_{x_0}^{B'}(z)\log\frac{\mu_{x_0}^{B'}(z)}{\mu^B(z)} = d_{KL}\left(\mu_{x_0}^{B'}\middle\|\mu^B\right) < \frac{5m}{\varepsilon(1-\varepsilon)}$$

it follows that

$$\mathop{\mathbf{Pr}}_{z\sim\mu_{x_0}^{B'}}\left[\frac{\mu_{x_0}^{B'}(z)}{\mu^B(z)} < 2^{\frac{10m}{\varepsilon(1-\varepsilon)}}\right] \geq \frac{1}{2}.$$

Let $Z'' \stackrel{\text{def}}{=} \left\{ z \in Z' \middle| \mu^B(z) > \mu_{x_0}^{B'}(z)\cdot 2^{-\frac{10m}{\varepsilon(1-\varepsilon)}} \right\}$, then $\mu^B(Z'') > 2^{-\frac{10m}{\varepsilon(1-\varepsilon)}-1}$.

We have that for any $x_0 \in X'$,

$$\mathop{\mathbf{Pr}}_{z\sim\mu^B}\left[(x_0,0,z)\in P\right] = \mu^B(Z') \geq \mu^B(Z'') > 2^{-\frac{10m}{\varepsilon(1-\varepsilon)}-1}.$$

If we sample $M \stackrel{\text{def}}{=} \left\lceil 2^{\frac{11m}{\varepsilon(1-\varepsilon)}} \right\rceil$ elements from μ^B then with probability greater than $1-\varepsilon/3$ at least one of them is a correct answer w.r.t. to the given x_0, whenever $x_0 \in X'$. As the latter happens with probability at least $1-\varepsilon/2$, the unconditional probability that one of the M elements is a correct answer is greater than $1-\varepsilon$. A pointer to one of M elements requires $\left\lceil \frac{11m}{\varepsilon(1-\varepsilon)} \right\rceil$ bits, and that is the cost of our $\mathcal{R}^1_{\mu,\varepsilon}$-protocol for P, as required. ■

4.3 Connection to learnability of unspeakable concepts

Let us see how Theorem 8 and Corollary 10 imply that our construction in Theorem 2 is tight. First, let us see that no unspeakable *functional* concept class can be efficiently learned even in a quantum predictive learning model.

Lemma 11 *Predictive learning of an unspeakable functional concept class is not possible from less than exponential amount of quantum (w.l.g.) information from the teacher, even by a computationally unlimited learner.*

Proof: Assume that for some functional concept class $\mathcal{F}$ that is unspeakable, the following holds. A teacher $\mathcal{T}$ knows some $f_0 \in \mathcal{F}$, hidden from a learner $\mathcal{S}$. Then $\mathcal{T}$ exchanges at most k_q qubits with $\mathcal{S}$. Finally, $\mathcal{S}$ is given some x_0 from the domain X of the functions in $\mathcal{F}$, and is able to compute $f_0(x_0)$ with confidence at least $5/6$.

Consider the following two-party communication task $\mathcal{G}$. Alice receives $f_0 \in \mathcal{F}$, Bob receives $x_0 \in X$ and they have to output $f_0(x_0)$. Clearly, $\mathcal{Q}^1_{5/6}(\mathcal{G}) \leq k_q$.

Let $k_c = \mathcal{R}^1(\mathcal{G})$. As $\mathcal{F}$ is unspeakable, $k_c \in 2^{\Omega(n)}$. By Theorem 8, $k_c \in O\left(n \cdot k_q \log(k_q)\right)$, and so $k_q \in 2^{\Omega(n)}$, as required. ■

Now we show that unspeakable concepts cannot be efficiently learned in the *quasi-predictive* (or standard) setting:

Lemma 12 *Both standard and quasi-predictive learning of an unspeakable concept class is not possible from less than exponential amount of quantum (w.l.g.) information from the teacher, even by a computationally unlimited learner.*

Proof: It is enough to prove the statement only for quasi-predictive learning, and the standard model can be viewed as a special case.

Let $\mathcal{C}$ be an unspeakable concept class consisting of relations over $X \times Y$, assume that it is learnable in the quasi-predictive model by a protocol of cost k_q. Then there exists a protocol, according to which a teacher $\mathcal{T}$ who knows some $\ell_0 \in \mathcal{C}$ exchanges at most k_q qubits with a learner $\mathcal{S}$ who doesn't know ℓ_0. Nevertheless, afterward $\mathcal{S}$ is able to answer with sufficient confidence any number of testing questions regarding ℓ_0.

For us it is enough to consider the testing phase where all possible $x \in X$ are asked (say, in the lexicographic order) and the learner responds with $(y_x)_{x \in X}$, such that

$$\forall (\ell_0, x) \in \mathcal{C} \times X : \mathbf{Pr}\left[(x, y_x) \in \ell_0\right] \geq 5/6,$$

where the probability is taken w.r.t. possible runs of the learning protocol for the given $\ell_0 \in \mathcal{C}$.

Define a relational single-input communication problem $P_{\mathcal{C}} \subseteq \mathcal{C} \times \mathbf{0} \times Y^X$ as

$$P_{\mathcal{C}} \stackrel{\text{def}}{=} \left\{ (\ell_0, 0, (y_x)_{x \in X}) \middle| \left|\left\{x \middle| (x, y_x) \in \ell_0\right\}\right| \geq \frac{4}{5}\left|X\right| \right\}.$$

The learning protocol for $\mathcal{C}$ that we considered above can be turned into a $\mathcal{Q}^1$-protocol of cost k_q for $P_{\mathcal{C}}$ that is correct with probability $1 - o(1)$ w.r.t. every $\ell_0 \in \mathcal{C}$, in particular $\mathcal{Q}^1(P_{\mathcal{C}}) \leq k_q$. By Corollary 10, $\mathcal{R}^1(P_{\mathcal{C}}) \in O(k_q)$.

Any $\mathcal{R}^1$-protocol of cost k_c for $P_{\mathcal{C}}$ readily leads to an approximating class for $\mathcal{C}$ of size 2^{k_c}. As $\mathcal{C}$ is unspeakable, $k_c \in 2^{\Omega(n)}$, where $n = \log|X|$. Therefore, $k_q \in 2^{\Omega(n)}$, as required. ■

For simplicity, in the two proofs above we assumed distribution-free mode of learning, where the learner in the testing phase had to give correct answer to any $x \in X$ with high probability. Distributional versions of Lemmas 11 and 12 can be proved similarly.

5 Open problems

We demonstrated that efficient quantum predictive learning of an unspeakable relational concept class is possible. The following questions seem interesting.

When we considered the limitations of quantum quasi-predictive learning (in the proof of Lemma 12), we argued that certain "quasi-hypothesis" of polynomial length can be extracted from an efficient quantum quasi-predictive learning algorithm. But our construction does not rely upon the efficiency of the learning algorithm, and on the other hand, the quasi-hypothesis we construct cannot, in general, be efficiently evaluated. It would be interesting to come up with a stronger argument that would "preserve efficiency"; or otherwise, to give an example of an interesting quantum quasi-predictive learning algorithm. Similar observations can be made w.r.t. our proof of Lemma 11. The transformation in [A04] is, in general, not efficient. Are there interesting quantum predictive (or even quasi-predictive) learning algorithms for functional concepts?

In the above questions by "interesting" we meant quantum algorithms for learning a concept class that *admits* concise hypotheses, but only those that cannot be efficiently evaluated. Observe that such "quasi-unspeakable" concept classes cannot be learned efficiently in any reasonable classical model (in the classical setting the equivalence between standard and predictive learning is efficiency-preserving).

Note that a trivial positive answer to these questions would follow, e.g., from an assumption that $BQP \not\subseteq P/poly$. Therefore the goal should be to weaken the assumptions.

More generally, give new examples of efficient quantum (quasi-)predictive learning of concept classes that are not efficiently learnable classically. Such examples might be interesting even for models stronger than PQ (e.g., one may allow the learner to make *membership queries*).

Acknowledgments

I thank Rahul Jain for helpful discussions.

References

[A04] S. Aaronson. Limitations of Quantum Advice and One-Way Communication. *Proceedings of the 19th IEEE Conference on Computational Complexity, pages 320-332*, 2004.

[BJ95] N. Bshouty and J. Jackson. Learning DNF over the Uniform Distribution using a Quantum Example Oracle. *Proceedings of the 8th Annual Conference on Computational Learning Theory, pages 118-127*, 1995.

[BJK04] Z. Bar-Yossef, T. S. Jayram and I. Kerenidis. Exponential Separation of Quantum and Classical One-Way Communication Complexity. *Proceedings of 36th Symposium on Theory of Computing, pages 128-137*, 2004.

[GKKRW07] D. Gavinsky, J. Kempe, I. Kerenidis, R. Raz and R. de Wolf. Exponential Separations for One-Way Quantum Communication Complexity, with Applications to Cryptography. *Proceedings of the 39th Symposium on Theory of Computing, pages 516-525*, 2007.

[GKRW06] D. Gavinsky, J. Kempe, O. Regev and R. de Wolf. Bounded-error Quantum State Identification and Exponential Separations in Communication Complexity. *Proceedings of the 38th Symposium on Theory of Computing, pages 594-603*, 2006.

[KW04] I. Kerenidis and R. de Wolf. Exponential Lower Bound for 2-Query Locally Decodable Codes via a Quantum Argument. *Journal of Computer and System Sciences, 69(3), pages 395-420*, 2004.

[SG04] R. Servedio and S. Gortler. Equivalences and Separations Between Quantum and Classical Learnability. *SIAM Journal on Computing 33(5), pages 1067-1092*, 2004.

[V84] L. Valiant. A Theory of Learnable. *Communications of the ACM 27(11), pages 1134-1142*, 1984.