Optimal Quantum Sample Complexity of Learning Algorithms

Srinivasan Arunachalam* Ronald de Wolf†

Abstract

In learning theory, the *VC dimension* of a concept class $\mathscr{C}$ is the most common way to measure its "richness." A fundamental result says that the number of examples needed to learn an unknown target concept $c \in \mathscr{C}$ under an unknown distribution D, is tightly determined by the VC dimension d of the concept class $\mathscr{C}$. Specifically, in the PAC model

$$\Theta\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$$

examples are necessary and sufficient for a learner to output, with probability $1-\delta$, a hypothesis h that is ε-close to the target concept c (measured under D). In the related *agnostic* model, where the samples need not come from a $c \in \mathscr{C}$, we know that

$$\Theta\left(\frac{d}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$$

examples are necessary and sufficient to output an hypothesis $h \in \mathscr{C}$ whose error is at most ε worse than the error of the best concept in $\mathscr{C}$.

Here we analyze *quantum* sample complexity, where each example is a coherent quantum state. This model was introduced by Bshouty and Jackson [BJ99], who showed that quantum examples are more powerful than classical examples in some fixed-distribution settings. However, Atıcı and Servedio [AS05], improved by Zhang [Zha10], showed that in the PAC setting (where the learner has to succeed for every distribution), quantum examples cannot be much more powerful: the required number of quantum examples is

$$\Omega\left(\frac{d^{1-\eta}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right) \text{ for arbitrarily small constant } \eta > 0.$$

Our main result is that quantum and classical sample complexity are in fact equal up to constant factors in both the PAC and agnostic models. We give two proof approaches. The first is a fairly simple information-theoretic argument that yields the above two classical bounds and yields the same bounds for quantum sample complexity up to a $\log(d/\varepsilon)$ factor. We then give a second approach that avoids the log-factor loss, based on analyzing the behavior of the "Pretty Good Measurement" on the quantum state identification problems that correspond to learning. This shows classical and quantum sample complexity are equal up to constant factors for every concept class $\mathscr{C}$.

*QuSoft, CWI, Amsterdam, the Netherlands. Supported by ERC Consolidator Grant QPROGRESS.

†QuSoft, CWI and University of Amsterdam, the Netherlands. Partially supported by ERC Consolidator Grant QPROGRESS.

1 Introduction

1.1 Sample complexity and VC dimension

Machine learning is one of the most successful parts of AI, with impressive practical applications in areas ranging from image processing, speech recognition, to even beating Go champions. Its theoretical aspects have been deeply studied, revealing beautiful structure and mathematical characterizations of when (efficient) learning is or is not possible in various settings.

1.1.1 The PAC setting

Leslie Valiant's Probably Approximately Correct (PAC) model [Val84] gives a precise complexity-theoretic definition of what it means for a concept class to be (efficiently) learnable. For simplicity we will (without loss of generality) focus on concepts that are Boolean functions, $c : \{0,1\}^n \to \{0,1\}$. Equivalently, a concept c is a subset of $\{0,1\}^n$, namely $\{x : c(x) = 1\}$. Let $\mathscr{C} \subseteq \{f : \{0,1\}^n \to \{0,1\}\}$ be a concept class. This could for example be the class of functions computed by disjunctive normal form (DNF) formulas of a certain size, or Boolean circuits or decision trees of a certain depth.

The goal of a learning algorithm (the learner) is to probably approximate some unknown *target concept* $c \in \mathscr{C}$ from random *labeled examples*. Each labeled example is of the form $(x, c(x))$ where x is distributed according to some unknown distribution D over $\{0,1\}^n$. After processing a number of such examples (hopefully not too many), the learner outputs some *hypothesis* h. We say that h is *ε-approximately correct* (w.r.t. the target concept c) if its error probability under D is at most ε: $\Pr_{x\sim D}[h(x) \neq c(x)] \leq \varepsilon$. Note that the learning phase and the evaluation phase (i.e., whether a hypothesis is approximately correct) are according to the same distribution D—as if the learner is taught and then tested by the same teacher. An (ε, δ)-learner for the concept class $\mathscr{C}$ is one whose hypothesis is probably approximately correct:

> For all target concepts $c \in \mathscr{C}$ and distributions D:
> Pr[the learner's output h is ε-approximately correct] $\geq 1 - \delta$,

where the probability is over the sequence of examples and the learner's internal randomness. Note that we leave the learner the freedom to output an h which is not in $\mathscr{C}$. If always $h \in \mathscr{C}$, then the learner is called a *proper* PAC-learner.

Of course, we want the learner to be as efficient as possible. Its *sample complexity* is the worst-case number of examples it uses, and its *time complexity* is the worst-case running time of the learner. In this paper we focus on sample complexity. This allows us to ignore technical issues of how the runtime of an algorithm is measured, and in what form the hypothesis h is given as output by the learner.

The sample complexity of a concept class $\mathscr{C}$ is the sample complexity of the most efficient learner for $\mathscr{C}$. It is a function of ε, δ, and of course of $\mathscr{C}$ itself. One of the most fundamental results in learning theory is that the sample complexity of $\mathscr{C}$ is tightly determined by a combinatorial parameter called the *VC dimension* of $\mathscr{C}$, due to and named after Vapnik and Chervonenkis [VC71]. The VC dimension of $\mathscr{C}$ is the size of the biggest $\mathcal{S} \subseteq \{0,1\}^n$ that can be labeled in all $2^{|\mathcal{S}|}$ possible ways by concepts from $\mathscr{C}$: for each sequence of $|\mathcal{S}|$ binary labels for the elements of $\mathcal{S}$, there is a $c \in \mathscr{C}$ that has that labeling (such an $\mathcal{S}$ is said to be *shattered* by $\mathscr{C}$). Knowing this VC dimension (and ε, δ) already tells us the sample complexity of $\mathscr{C}$ up to constant factors. Blumer et al. [BEHW89] proved that the sample complexity of $\mathscr{C}$ is lower bounded by $\Omega(d/\varepsilon + \log(1/\delta)/\varepsilon)$, and they proved an upper bound that was worse by a $\log(1/\varepsilon)$-factor. In very recent work, Han-

neke [Han16] (improving on Simon [Sim15]) got rid of this $\log(1/\varepsilon)$-factor for PAC learning,[1] showing that the lower bound of Blumer et al. is in fact optimal: the sample complexity of $\mathscr{C}$ in the PAC setting is

$$\Theta\left(\frac{d}{\varepsilon}+\frac{\log(1/\delta)}{\varepsilon}\right). \tag{1}$$

1.1.2 The agnostic setting

The PAC model assumes that the labeled examples are generated according to a target concept $c \in \mathscr{C}$. However, in many learning situations that is not a realistic assumption, for example when the examples are noisy in some way or when we have no reason to believe there is an underlying target concept at all. The *agnostic* model of learning, introduced by Haussler [Hau92] and Kearns et al. [KSS94], takes this into account. Here, the examples are generated according to a distribution D on $\{0,1\}^{n+1}$. The error of a specific concept $c : \{0,1\}^n \to \{0,1\}$ is defined to be $\mathrm{err}_D(c) = \Pr_{(x,b)\sim D}[c(x) \neq b]$. When we are restricted to hypotheses in $\mathscr{C}$, we would like to find the hypothesis that minimizes $\mathrm{err}_D(c)$ over all $c \in \mathscr{C}$. However, it may require very many examples to do that exactly. In the spirit of the PAC model, the goal of the learner is now to output an $h \in \mathscr{C}$ whose error is at most an additive ε worse than that of the best (= lowest-error) concepts in $\mathscr{C}$.

Like in the PAC model, the optimal sample complexity of such agnostic learners is tightly determined by the VC dimension of $\mathscr{C}$: it is

$$\Theta\left(\frac{d}{\varepsilon^2}+\frac{\log(1/\delta)}{\varepsilon^2}\right), \tag{2}$$

where the lower bound was proven by Vapnik and Chervonenkis [VC74] (see also Simon [Sim96]), and the upper bound was proven by Talagrand [Tal94]. Shalev-Shwartz and Ben-David [SB14, Section 6.4] call Eq. (1) and Eq. (2) the "Fundamental Theorem of PAC learning."

1.2 Our results

In this paper we are interested in *quantum* sample complexity. Here a *quantum example* for some concept $c : \{0,1\}^n \to \{0,1\}$, according to some distribution D, corresponds to an $(n+1)$-qubit state

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

In other words, instead of a random labeled example, an example is now given by a coherent quantum superposition where the square-roots of the probabilities become the amplitudes.[2] This model was introduced by Bshouty and Jackson [BJ99], who showed that DNF formulas are learnable in polynomial time from quantum examples when D is uniform. For learning DNF under the uniform distribution from *classical* examples, the best upper bound is quasipolynomial time [Ver90]. With the added power of "membership queries," where the learner can actively ask for the label of any x of his choice, DNF formulas are known to be learnable in polynomial time under uniform D [Jac97], but *without* membership queries polynomial-time learnability is a longstanding open problem (see [DS16] for a recent hardness result).

[1] Hanneke's learner is not proper, meaning that its hypothesis h is not always in $\mathscr{C}$. It is still an open question whether the $\log(1/\varepsilon)$-factor can be removed for proper PAC learning. Our lower bounds in this paper hold for all learners, quantum as well as classical, and proper as well as improper.

[2] We could allow more general quantum examples $\sum_{x\in\{0,1\}^n} \alpha_x|x, c(x)\rangle$, where we only require $|\alpha_x|^2 = D(x)$. However, that will not affect our results since our lower bounds apply to quantum examples where we know the amplitudes are square-rooted probabilities. Adding more degrees of freedom to quantum examples does not make learning easier.

How reasonable are examples that are given as a coherent superposition rather than as a random sample? They may seem unreasonable a priori because quantum superpositions seem very fragile and are easily collapsed by measurement, but if we accept the "church of the larger Hilbert space" view on quantum mechanics, where the universe just evolves unitarily without any collapses, then they may become more palatable. It is also possible that the quantum examples are generated by some coherent quantum process that acts like the teacher.

How many quantum examples are needed to learn a concept class $\mathscr{C}$ of VC dimension d? Since a learner can just measure a quantum example in order to obtain a classical example, the *upper* bounds on classical sample complexity trivially imply the same upper bounds on quantum sample complexity. But what about the lower bounds? Are there situations where quantum examples are more powerful than classical? Indeed there are. We already mentioned the results of Bshouty and Jackson [BJ99] for learning DNF under the uniform distribution without membership queries. Another good example is the learnability of the concept class of linear functions over $\mathbb{F}_2$, $\mathscr{C} = \{c(x) = a \cdot x : a \in \{0,1\}^n\}$, again under the uniform distribution D. It is easy to see that a classical learner needs about n examples to learn an unknown $c \in \mathscr{C}$ under this D. However, if we are given one quantum example

$$\sum_{x \in \{0,1\}^n} \sqrt{D(x)}|x, c(x)\rangle = \frac{1}{\sqrt{2^n}} \sum_{x \in \{0,1\}^n} |x, a \cdot x\rangle,$$

then a small modification of the Bernstein-Vazirani algorithm [BV97] can recover a (and hence c) with probability $1/2$. Hence $O(1)$ quantum examples suffice to learn c exactly, with high probability, under the uniform distribution. Atıcı and Servedio [AS09] used similar ideas to learning *k-juntas* (concepts depending on only k of their n variables) from quantum examples under the uniform distribution. However, PAC learning requires a learner to learn c under *all possible* distributions D, not just the uniform one. The success probability of the Bernstein-Vazirani algorithm deteriorates sharply when D is far from uniform, but that does not rule out the existence of other quantum learners that use $o(n)$ quantum examples and succeed for all D.

Our main result in this paper is that quantum examples are not actually more powerful than classical labeled examples in the PAC model and in the agnostic model: we prove that the lower bounds on classical sample complexity of Eq. (1) and Eq. (2) hold for quantum examples as well. Accordingly, despite several distribution-specific speedups, quantum examples do not significantly reduce sample complexity if we require our learner to work for all distributions D. This should be contrasted with the situation when considering the *time complexity* of learning. Servedio and Gortler [SG04] considered a concept class (already known in the literature [KV94a, Chapter 6]) that can be PAC-learned in polynomial time by a quantum computer, even with only classical examples, but that cannot be PAC-learned in polynomial time by a classical learner unless Blum integers can be factored in polynomial time (which is widely believed to be false).

Earlier work on quantum sample complexity had already gotten close to extending the lower bound of Eq. (1) to PAC learning from quantum examples. Atıcı and Servedio [AS05] first proved a lower bound of $\Omega(\sqrt{d}/\varepsilon + d + \log(1/\delta)/\varepsilon)$ using the so-called "hybrid method." Their proof technique was subsequently pushed further by Zhang [Zha10] to

$$\Omega\left(\frac{d^{1-\eta}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right) \text{ for arbitrarily small constant } \eta > 0. \tag{3}$$

Here we optimize these bounds, removing the η and achieving the optimal lower bound for quantum sample complexity in the PAC model (Eq. (1)).

We also show that the lower bound (Eq. (2)) for the agnostic model extends to quantum examples. As far as we know, in contrast to the PAC model, no earlier results were known for quantum

sample complexity in the agnostic model.

We have two different proof approaches, which we sketch below.

1.2.1 An information-theoretic argument

In Section 3 we give a fairly intuitive information-theoretic argument that gives optimal lower bounds for classical sample complexity, and that gives nearly-optimal lower bounds for quantum sample complexity. Let us first see how we can prove the classical PAC lower bound of Eq. (1). Suppose $S = \{s_0, s_1, \ldots, s_d\}$ is shattered by $\mathscr{C}$ (we now assume VC dimension $d+1$ for ease of notation). Then we can consider a distribution D that puts probability $1-4\varepsilon$ on s_0 and probability $4\varepsilon/d$ on each of $s_1, \ldots, s_d$.[3] For every possible labeling $(\ell_1 \ldots \ell_d) \in \{0,1\}^d$ of $s_1, \ldots, s_d$ there will be a concept $c \in \mathscr{C}$ that labels s_0 with 0, and labels s_i with ℓ_i for all $i \in \{1, \ldots, d\}$. Under D, most examples will be $(s_0, 0)$ and hence give us no information when we are learning one of those 2^d concepts. Suppose we have a learner that ε-approximates c with high probability under this D using T examples. Informally, our information-theoretic argument has the following three steps:

1. In order to ε-approximate c, the learner has to learn the c-labels of at least 3/4 of the $s_1, \ldots, s_d$ (since together these have 4ε of the D-weight, and we want an ε-approximation). As all 2^d labelings are possible, the T examples together contain $\Omega(d)$ bits of information about c.

2. T examples give at most T times as much information about c as one example.

3. One example gives only $O(\varepsilon)$ bits of information about c, because it will tell us one of the labels of $s_1, \ldots, s_d$ only with probability 4ε (and otherwise it just gives $c(s_0) = 0$).

Putting these steps together implies $T = \Omega(d/\varepsilon)$.[4] This argument for the PAC setting is similar to an algorithmic-information argument of Apolloni and Gentile [AG98] and an information-theoretic argument for variants of the PAC model with noisy examples of Gentile and Helmbold [GH01].

As far as we know, this type of reasoning has not yet been applied to the sample complexity of *agnostic* learning. To get good lower bounds there, we consider a set of distributions D_a, indexed by d-bit string a. These distributions still have the property that if a learner gets ε-close to the minimal error, then it will have to learn $\Omega(d)$ bits of information about the distribution (i.e., about a). Hence the first step of the argument remains the same. The second step of our argument also remains the same, and the third step shows an upper bound of $O(\varepsilon^2)$ on the amount of information that the learner can get from one example. This then implies $T = \Omega(d/\varepsilon^2)$. We can also reformulate this for the case where we want the *expected* additional error of the hypothesis over the best classifier in $\mathscr{C}$ to be at most ε, which is how lower bounds are often stated in learning theory. We emphasize that our information-theoretic proof is simpler than the proofs in [AB09, Aud09, SB14, KP16].

This information-theoretic approach recovers the optimal classical bounds on sample complexity, but also generalizes readily to the quantum case where the learner gets T quantum examples. To obtain lower bounds on quantum sample complexity we use the same distributions D (now corresponding to a coherent quantum state) and basically just need to re-analyze the third step of the argument. In the PAC setting we show that one quantum example gives at most $O(\varepsilon \log(d/\varepsilon))$ bits of information about c, and in the agnostic setting it gives $O(\varepsilon^2 \log(d/\varepsilon))$ bits.

[3] We remark that the distributions used here for proving lower bounds on quantum sample complexity have been used by Ehrenfeucht et al. [EHKV89] for analyzing classical PAC sample complexity.

[4] The other part of the lower bound of Eq. (1) does not depend on d and is fairly easy to prove.

This implies lower bounds on sample complexity that are only a logarithmic factor worse than the optimal classical bounds for the PAC setting (Eq. (1)) and the agnostic setting (Eq. (2)). This is not quite optimal yet, but already better than the previous best known lower bound (Eq. (3)). The logarithmic loss in step 3 is actually inherent in this information-theoretic argument: in some cases a quantum example *can* give roughly $\varepsilon \log d$ bits of information about c, for example when c comes from the concept class of linear functions.

1.2.2 A state-identification argument

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 $s_1, \ldots, s_d$ 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.

State-identification problems have been well studied, and many tools are available for analyzing them. In particular, we will use the so-called "Pretty Good Measurement" (PGM, also known as "square root measurement" [HJS+96]) introduced by Hausladen and Wootters [HW94]. 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.[5] In Section 4 we 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 and the agnostic model. This analysis could be useful in other settings as well. Here it implies that the number of quantum examples, T, is lower bounded by Eq. (1) in the PAC setting, and by Eq. (2) in the agnostic setting.

Using the Pretty Good Measurement, we are also able to prove lower bounds for PAC learning under *random classification noise*, which models the real-world situation that the learning data can have some errors. Classically in the random classification noise model (introduced by Angluin and Laird [AL88]), instead of obtaining labeled examples $(x, c(x))$ for some unknown $c \in \mathscr{C}$, the learner obtains *noisy examples* (x, b_x), where $b_x = c(x)$ with probability $1-\eta$ and $b_x = 1 - c(x)$ with probability η, for some *noise rate* $\eta \in [0, 1/2)$. Similarly, in the quantum learning model we could naturally define a *noisy quantum example* as an $(n+1)$-qubit state

$$\sum_{x \in \{0,1\}^n} \sqrt{(1-\eta)D(x)}|x, c(x)\rangle + \sqrt{\eta D(x)}|x, 1-c(x)\rangle.$$

Using the PGM, we are able to show that the quantum sample complexity of PAC learning a concept class $\mathscr{C}$ under random classification noise is:

$$\Omega\Big(\frac{d}{(1-2\eta)^2\varepsilon} + \frac{\log(1/\delta)}{(1-2\eta)^2\varepsilon}\Big). \tag{4}$$

We remark here that the best known classical sample complexity lower bound (see [Sim96]) under the random classification noise is equal to the quantum sample complexity lower bound proven in Eq. (4).

[5] Even better, in our application the PGM *is* the optimal measurement, though this is not essential for our proof.

- Passive vs active learning
  of classical states
- exact learning / oracle identification
- time complexity
- learnability of quantum states

1.3 Related work

Let us briefly discuss some related work in quantum learning theory, referring to our recent survey [AdW17] for more. In this paper we focus on *sample* complexity, which is a fundamental information-theoretic quantity. Sample complexity concerns a form of "passive" learning: the learner gets a number of examples at the start of the process, and then has to extract enough information about the target concept from these. We may also consider more active learning settings, in particular ones where the learner can make membership queries (i.e., learn the label $c(x)$ for any x of his choice). Servedio and Gortler [SG04] showed that in this setting, classical and quantum complexity are polynomially related. They also exhibit an example of a factor-n speed-up from quantum membership queries using the Bernstein-Vazirani algorithm. Jackson et al. [JTY02] showed how quantum membership queries can improve Jackson's classical algorithm for learning DNF with membership queries under the uniform distribution [Jac97].

For *quantum exact learning* (also referred to as the *oracle identification* problem in the quantum literature), Kothari [Kot14] resolved a conjecture of Hunziker et al. [HMP+10], that states that for any concept class $\mathscr{C}$, the number of quantum membership queries required to exactly identify a concept $c \in \mathscr{C}$ is $O(\frac{\log|\mathscr{C}|}{\sqrt{\hat{\gamma}^{\mathscr{C}}}})$, where $\hat{\gamma}^{\mathscr{C}}$ is a combinatorial parameter of the concept class $\mathscr{C}$ which we shall not define here (see [AS05] for a precise definition). Montanaro [Mon12] showed how low-degree polynomials over a finite field can be identified more efficiently using quantum algorithms.

In many ways the *time* complexity of learning is at least as important as the sample complexity. We already mentioned that Servedio and Gortler [SG04] exhibited a concept class based on factoring Blum integers that can be learned in quantum polynomial time but not in classical polynomial time, unless Blum integers can be factored efficiently. Under the weaker (but still widely believed) assumption that one-way functions exist, they exhibited a concept class that can be learned exactly in polynomial time using quantum membership queries, but that takes superpolynomial time to learn from classical membership queries. Gavinsky [Gav12] introduced a model of learning called "Predictive Quantum" (PQ), a variation of quantum PAC learning, and exhibited a *relational* concept class that is polynomial-time learnable in PQ, while any "reasonable" classical model requires an exponential number of classical examples to learn the concept class.

Aïmeur et al. [ABG06, ABG13] consider a number of quantum algorithms in learning contexts such as clustering via minimum spanning tree, divisive clustering, and k-medians, using variants of Grover's algorithm [Gro96] to improve the time complexity of the analogous classical algorithms. Recently, there have been some quantum machine learning algorithms based on the HHL algorithm [HHL09] for solving (in a weak sense) very well-behaved linear systems. However, these algorithms often come with some fine print that limits their applicability, and their advantage over classical is not always clear. We refer to Aaronson [Aar15] for references and caveats. There has also been some work on quantum training of neural networks [WKS14, WKS16].

In addition to learning classical objects such as Boolean functions, one may also study the learnability of *quantum* objects. In particular, Aaronson [Aar07] studied how well n-qubit quantum states can be learned from measurement results. In general, an n-qubit state ρ is specified by $\exp(n)$ many parameters, and $\exp(n)$ measurement results on equally many copies of ρ are needed to learn a good approximation of ρ (say, in trace distance). However, Aaronson showed an interesting and surprisingly efficient PAC-like result: from $O(n)$ measurement results, with measurements chosen i.i.d. according to an unknown distribution D on the set of all possible two-outcome measurements, we can learn an n-qubit quantum state $\widetilde{\rho}$ that has roughly the same expectation value as ρ for "most" possible two-outcome measurements. In the latter, "most" is again measured under D, just like in the usual PAC learning the error of the learner's hypothesis

is evaluated under the same distribution D that generated the learner's examples. Accordingly, $O(n)$ rather than $\exp(n)$ measurement results suffice to approximately learn an n-qubit state for most practical purposes.

The use of Fourier analysis in analyzing the success probability of the Pretty Good Measurement in quantum state identification appears in a number of earlier works. By considering the dihedral hidden subgroup problem (DHSP) as a state identification problem, Bacon et al. [BCD06] show that the PGM is the optimal measurement for DHSP and prove a lower bound on the sample complexity of $\Omega(\log|\mathcal{G}|)$ for a dihedral group $\mathcal{G}$ using Fourier analysis. Ambainis and Montanaro [AM14] view the "search with wildcard" problem as a state identification problem. Using ideas similar to ours, they show that the (x,y)-th entry of the Gram matrix for the ensemble depends on the Hamming distance between x and y, allowing them to use Fourier analysis to obtain an upper bound on the success probability of the state identification problem using the PGM.

1.4 Organization

In Section 2 we formally define the classical and quantum learning models and introduce the Pretty Good Measurement. In Section 3 we prove our information-theoretic lower bounds both for classical and quantum learning. In Section 4 we prove an optimal quantum lower bound for PAC and agnostic learning by viewing the learning process as a state identification problem. We conclude in Section 5 with some open questions for further work.

2 Preliminaries

2.1 Notation

Let $[n] = \{1,\ldots,n\}$. For $x,y \in \{0,1\}^d$, the bit-wise sum $x+y$ is over $\mathbb{F}_2$, the *Hamming distance* $d(x,y)$ is the number of indices on which x and y differ, $|x+y|$ is the Hamming weight of the string $x+y$ (which equals $d_H(x,y)$), and $x\cdot y = \sum_i x_i y_i$ (where the sum is over $\mathbb{F}_2$). For an n-dimensional vector space, the standard basis is denoted by $\{e_i \in \{0,1\}^n : i \in [n]\}$, where e_i is the vector with a 1 in the i-th coordinate and 0's elsewhere. We write log for logarithm to base 2, and ln for base e. We will often use the bijection between the sets $\{0,1\}^k$ and $[2^k]$ throughout this paper. Let $1_{[A]}$ be the indicator for an event A, and let $\delta_{x,y} = 1_{[x=y]}$. We denote random variables in bold, such as $\mathbf{A}$, $\mathbf{B}$.

For a Boolean function $f : \{0,1\}^m \to \{0,1\}$ and $M \in \mathbb{F}_2^{m\times k}$ we define $f\circ M : \{0,1\}^k \to \{0,1\}$ as $(f\circ M)(x) := f(Mx)$ (where the matrix-vector product is over $\mathbb{F}_2$) for all $x \in \{0,1\}^k$. For a distribution $D : \{0,1\}^n \to [0,1]$, let $\mathrm{supp}(D) = \{x \in \{0,1\}^n : D(x) \neq 0\}$. By $x \sim D$, we mean x is sampled according to the distribution D, i.e., $\Pr[\mathbf{X} = x] = D(x)$.

If M is a positive semidefinite (psd) matrix, we define $\sqrt{M}$ as the unique psd matrix that satisfies $\sqrt{M}\cdot\sqrt{M} = M$, and $\sqrt{M}(i,j)$ as the (i,j)-th entry of $\sqrt{M}$. For a matrix $A \in \mathbb{R}^{m\times n}$, we denote the singular values of A by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A) \geq 0$. The spectral norm of A is $\|A\| = \max_{x\in\mathbb{R}^n,\|x\|=1} \|Ax\| = \sigma_1$. Given a set of d-dimensional vectors $U = \{u_1,\ldots,u_n\} \in \mathbb{R}^d$, the Gram matrix V corresponding to the set U is the $n\times n$ psd matrix defined as $V(i,j) = u_i^t u_j$ for $i,j \in [n]$, where u_i^t is the row vector that is the transpose of the column vector u_i.

A technical tool used in our analysis of state identification problems is Fourier analysis on the Boolean cube. We will just introduce the basics of Fourier analysis here, referring to [O'D14] for more. Define the inner product between functions $f,g : \{0,1\}^n \to \mathbb{R}$ as

$$\langle f,g\rangle = \mathop{\mathbb{E}}_x[f(x)\cdot g(x)]$$

where the expectation is uniform over $x \in \{0,1\}^n$. For $S \subseteq [n]$ (equivalently $S \in \{0,1\}^n$), let $\chi_S(x) := (-1)^{S \cdot x}$ denote the parity of the variables (of x) indexed by the set S. It is easy to see that the set of functions $\{\chi_S\}_{S \subseteq [n]}$ forms an orthonormal basis for the space of real-valued functions over the Boolean cube. Hence every f can be decomposed as

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S)(-1)^{S \cdot x} \qquad \text{for all } x \in \{0,1\}^n,$$

where $\widehat{f}(S) = \langle f, \chi_S \rangle = \mathbb{E}_x[f(x) \cdot \chi_S(x)]$ is called a *Fourier coefficient* of f.

2.2 Learning in general

In machine learning, a concept class $\mathscr{C}$ over $\{0,1\}^n$ is a set of concepts $c : \{0,1\}^n \to \{0,1\}$. We refer to a concept class $\mathscr{C}$ as being *trivial* if either $\mathscr{C}$ contains only one concept, or $\mathscr{C}$ contains two concepts c_0, c_1 with $c_0(x) = 1 - c_1(x)$ for every $x \in \{0,1\}^n$. For $c : \{0,1\}^n \to \{0,1\}$, we will often refer to the tuple $(x, c(x)) \in \{0,1\}^{n+1}$ as a *labeled example*, where $c(x)$ is the *label* of x.

A central combinatorial concept in learning theory is the Vapnik-Chervonenkis (VC) dimension [VC71]. Fix a concept class $\mathscr{C}$ over $\{0,1\}^n$. A set $\mathcal{S} = \{s_1, \ldots, s_t\} \subseteq \{0,1\}^n$ is said to be *shattered* by a concept class $\mathscr{C}$ if $\{(c(s_1), \ldots, c(s_t)) : c \in \mathscr{C}\} = \{0,1\}^t$. In other words, for every labeling $\ell \in \{0,1\}^t$, there exists a $c \in \mathscr{C}$ such that $(c(s_1), \ldots, c(s_t)) = \ell$. The VC dimension of a concept class $\mathscr{C}$ is the size of the largest $\mathcal{S} \subseteq \{0,1\}^n$ that is shattered by $\mathscr{C}$.

2.3 Classical learning models

In this paper we will be concerned mainly with the PAC (Probably Approximately Correct) model of learning introduced by Valiant [Val84], and the agnostic model of learning introduced by Haussler [Hau92] and Kearns et al. [KSS94]. For further reading, see standard textbooks in computational learning theory such as [KV94b, AB09, SB14].

In the classical PAC model, a learner $\mathcal{A}$ is given access to a *random example oracle* $\mathrm{PEX}(c, D)$ which generates labeled examples of the form $(x, c(x))$ where x is drawn from an unknown distribution $D : \{0,1\}^n \to [0,1]$ and $c \in \mathscr{C}$ is the *target concept* that $\mathcal{A}$ is trying to learn. For a concept $c \in \mathscr{C}$ and hypothesis $h : \{0,1\}^n \to \{0,1\}$, we define the error of h compared to the target concept c, under D, as $\mathrm{err}_D(h, c) = \Pr_{x \sim D}[h(x) \neq c(x)]$. A learning algorithm $\mathcal{A}$ is an (ε, δ)-*PAC learner* for $\mathscr{C}$, if the following holds:

For every $c \in \mathscr{C}$ and distribution D, given access to the $\mathrm{PEX}(c, D)$ oracle:
$\mathcal{A}$ outputs an h such that $\mathrm{err}_D(h, c) \leq \varepsilon$ with probability at least $1 - \delta$.

The *sample complexity* of $\mathcal{A}$ is the maximum number of invocations of the $\mathrm{PEX}(c, D)$ oracle which the learner makes, over all concepts $c \in \mathscr{C}$, distributions D, and the internal randomness of the learner. The (ε, δ)-*PAC sample complexity* of a concept class $\mathscr{C}$ is the minimum sample complexity over all (ε, δ)-PAC learners for $\mathscr{C}$.

Agnostic learning is the following model: for a distribution $D : \{0,1\}^{n+1} \to [0,1]$, a learner $\mathcal{A}$ is given access to an $\mathrm{AEX}(D)$ oracle that generates examples of the form (x, b) drawn from the distribution D. We define the error of $h : \{0,1\}^n \to \{0,1\}$ under D as $\mathrm{err}_D(h) = \Pr_{(x,b) \sim D}[h(x) \neq b]$. When h is restricted to come from a concept class $\mathscr{C}$, the minimal error achievable is $\mathrm{opt}_D(\mathscr{C}) = \min_{c \in \mathscr{C}}\{\mathrm{err}_D(c)\}$. In agnostic learning, a learner $\mathcal{A}$ needs to output a hypothesis h whose error is not much bigger than $\mathrm{opt}_D(\mathscr{C})$. A learning algorithm $\mathcal{A}$ is an (ε, δ)-*agnostic learner* for $\mathscr{C}$ if:

For every distribution D on $\{0,1\}^{n+1}$, given access to the AEX(D) oracle:
$\mathcal{A}$ outputs an $h \in \mathscr{C}$ such that $\mathrm{err}_D(h) \leq \mathrm{opt}_D(\mathscr{C}) + \varepsilon$ with probability at least $1-\delta$.

Note that if there is a $c \in \mathscr{C}$ which perfectly classifies every x with label y for $(x,y) \in \mathrm{supp}(D)$, then $\mathrm{opt}_D(\mathscr{C}) = 0$ and we are in the setting of proper PAC learning. The *sample complexity* of $\mathcal{A}$ is the maximum number of invocations of the AEX(c,D) oracle which the learner makes, over all distributions D and over the learner's internal randomness. The (ε,δ)-*agnostic sample complexity* of a concept class $\mathscr{C}$ is the minimum sample complexity over all (ε,δ)-agnostic learners for $\mathscr{C}$.

2.4 Quantum information theory

Throughout this paper we will assume the reader is familiar with the following quantum terminology. An n-dimensional *pure state* is $|\psi\rangle = \sum_{i=1}^{n} \alpha_i |i\rangle$, where $|i\rangle$ is the n-dimensional unit vector that has a 1 only at position i, the α_i's are complex numbers called the *amplitudes*, and $\sum_{i\in[n]} |\alpha_i|^2 = 1$. An n-dimensional *mixed state* (or *density matrix*) $\rho = \sum_{i=1}^{n} p_i |\psi_i\rangle\langle\psi_i|$ is a mixture of pure states $|\psi_1\rangle,\ldots,|\psi_n\rangle$ prepared with probabilities $p_1,\ldots,p_n$, respectively. The eigenvalues $\lambda_1,\ldots,\lambda_n$ of ρ are non-negative reals and satisfy $\sum_{i\in[n]} \lambda_i = 1$. If ρ is pure (i.e., $\rho = |\psi\rangle\langle\psi|$ for some $|\psi\rangle$), then one of the eigenvalues is 1 and the others are 0.

To obtain classical information from ρ, one could apply a POVM (positive-operator-valued measure) to the state ρ. An m-outcome POVM is specified by a set of positive semidefinite matrices $\{M_i\}_{i\in[m]}$ with the property $\sum_i M_i = \mathrm{Id}$. When this POVM is applied to the mixed state ρ, the probability of the j-th outcome is given by $\mathrm{Tr}(M_j\rho)$.

For a probability vector $(p_1,\ldots,p_k)$ (where $p_i \geq 0$ and $\sum_{i\in[k]} p_i = 1$), the entropy function is defined as $H(p_1,\ldots,p_k) = -\sum_{i\in[k]} p_i \log p_i$. When $k=2$, with $p_1 = p$ and $p_2 = 1-p$, we denote the binary entropy function as $H(p)$. For a state ρ_{AB} on the Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$, we let ρ_A be the reduced state after taking the partial trace over $\mathcal{H}_B$. The entropy of a quantum state ρ_A is defined as $S(\mathbf{A}) = -\mathrm{Tr}(\rho_A \log \rho_A)$. The mutual information is defined as $I(\mathbf{A}:\mathbf{B}) = S(\mathbf{A}) + S(\mathbf{B}) - S(\mathbf{AB})$, and conditional entropy is defined as $S(\mathbf{A}|\mathbf{B}) = S(\mathbf{AB}) - S(\mathbf{B})$. Classical information-theoretic quantities correspond to the special case where ρ is a diagonal matrix whose diagonal corresponds to the probability distribution of the random variable. Writing ρ_A in its eigenbasis, it follows that $S(\mathbf{A}) = H(\lambda_1,\ldots,\lambda_{\dim(\rho_A)})$, where $\lambda_1,\ldots,\lambda_{\dim(\rho_A)}$ are the eigenvalues of ρ. If ρ_A is a pure state, $S(\mathbf{A}) = 0$.

2.5 Quantum learning models

The quantum PAC learning model was introduced by Bshouty and Jackson in [BJ99]. The quantum PAC model is a generalization of the classical PAC model, instead of having access to random examples $(x,c(x))$ from the PEX(c,D) oracle, the learner now has access to superpositions over all $(x,c(x))$. For an unknown distribution $D : \{0,1\}^n \to [0,1]$ and concept $c \in \mathscr{C}$, a *quantum example oracle* QPEX(c,D) acts on $|0^n,0\rangle$ and produces a *quantum example* $\sum_{x\in\{0,1\}^n} \sqrt{D(x)}|x,c(x)\rangle$ (we leave QPEX undefined on other basis states). 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. A learning algorithm $\mathcal{A}$ is an (ε,δ)-*PAC quantum learner* for $\mathscr{C}$ if:

For every $c \in \mathscr{C}$ and distribution D, given access to the QPEX(c,D) oracle:
$\mathcal{A}$ outputs an h such that $\mathrm{err}_D(h,c) \leq \varepsilon$, with probability at least $1-\delta$.

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

We define quantum agnostic learning now. For a joint distribution $D : \{0,1\}^{n+1} \to [0,1]$ over the set of examples, the learner has access to an QAEX(D) oracle which acts on $|0^n, 0\rangle$ and produces a quantum example $\sum_{(x,b)\in\{0,1\}^{n+1}} \sqrt{D(x,b)}|x,b\rangle$. A learning algorithm $\mathcal{A}$ is an (ε,δ)-*agnostic quantum learner* for $\mathscr{C}$ if:

> For every distribution D, given access to the QAEX(D) oracle:
> $\mathcal{A}$ outputs an $h \in \mathscr{C}$ such that $\mathrm{err}_D(h) \le \mathrm{opt}_D(\mathscr{C}) + \varepsilon$ with probability at least $1-\delta$.

The *sample complexity* of $\mathcal{A}$ is the maximum number invocations of the QAEX(D) oracle over all distributions D and over the learner's internal randomness. The (ε,δ)-*agnostic quantum sample complexity* of a concept class $\mathscr{C}$ is the minimum sample complexity over all (ε,δ)-agnostic quantum learners for $\mathscr{C}$.

2.6 Pretty Good Measurement

Consider an ensemble of d-dimensional states, $\mathcal{E} = \{(p_i, |\psi_i\rangle)\}_{i\in[m]}$, where $\sum_{i\in[m]} p_i = 1$. Suppose we are given an unknown state $|\psi_i\rangle$ sampled according to the probabilities and we are interested in maximizing the average probability of success to identify the state that we are given. For a POVM specified by positive semidefinite matrices $\mathcal{M} = \{M_i\}_{i\in[m]}$, the probability of obtaining outcome j equals $\langle\psi_i|M_j|\psi_i\rangle$. The average success probability is defined as

$$P_{\mathcal{M}}(\mathcal{E}) = \sum_{i=1}^{m} p_i \langle\psi_i|M_i|\psi_i\rangle.$$

Let $P^{opt}(\mathcal{E}) = \max_{\mathcal{M}} P_{\mathcal{M}}(\mathcal{E})$ denote the optimal average success probability of $\mathcal{E}$, where the maximization is over the set of valid m-outcome POVMs.

For every ensemble $\mathcal{E}$, the so-called *Pretty Good Measurement* (PGM) is a specific POVM (depending on the ensemble $\mathcal{E}$), which we shall define shortly, that does *reasonably* well against $\mathcal{E}$. Suppose $P^{PGM}(\mathcal{E})$ is defined as the average success probability of identifying the states in $\mathcal{E}$ using the PGM, then we have that

$$P^{opt}(\mathcal{E})^2 \le P^{PGM}(\mathcal{E}) \le P^{opt}(\mathcal{E}),$$

where the second inequality follows because $P^{opt}(\mathcal{E})$ is a maximization over all valid POVMs and the first inequality was shown by Barnum and Knill [BK02].

For completeness we give a simple proof of $P^{opt}(\mathcal{E})^2 \le P^{PGM}(\mathcal{E})$ below (similar to [Mon07]). Let $|\psi_i'\rangle = \sqrt{p_i}|\psi_i\rangle$, and $\mathcal{E}' = \{|\psi_i'\rangle : i \in [m]\}$ be the set of states in $\mathcal{E}$, renormalized to reflect their probabilities. Define $\rho = \sum_{i\in[m]} |\psi_i'\rangle\langle\psi_i'|$. The PGM is defined as the set of measurement operators $\{|v_i\rangle\langle v_i|\}_{i\in[m]}$ where $|v_i\rangle = \rho^{-1/2}|\psi_i'\rangle$ (the inverse square root of ρ is taken over its non-zero eigenvalues). We first verify this is a valid POVM:

$$\sum_{i=1}^{m} |v_i\rangle\langle v_i| = \rho^{-1/2}\left(\sum_{i=1}^{m} |\psi_i'\rangle\langle\psi_i'|\right)\rho^{-1/2} = \mathrm{Id}.$$

Let G be the Gram matrix for the set $\mathcal{E}'$, i.e., $G(i,j) = \langle\psi_i'|\psi_j'\rangle$ for $i,j \in [m]$. It can be verified that $\sqrt{G}(i,j) = \langle\psi_i'|\rho^{-1/2}|\psi_j'\rangle$. Hence

$$\begin{aligned} P^{PGM}(\mathcal{E}) = \sum_{i\in[m]} p_i |\langle v_i|\psi_i\rangle|^2 &= \sum_{i\in[m]} |\langle v_i|\psi_i'\rangle|^2 \\ &= \sum_{i\in[m]} \langle\psi_i'|\rho^{-1/2}|\psi_i'\rangle^2 = \sum_{i\in[m]} \sqrt{G}(i,i)^2. \end{aligned}$$

We now prove $P^{opt}(\mathcal{E})^2 \leq P^{PGM}(\mathcal{E})$. Suppose $\mathcal{M}$ is the optimal measurement. Since $\mathcal{E}$ consists of pure states, by a result of Eldar et al. [EMV03], we can assume without loss of generality that the measurement operators in $\mathcal{M}$ are rank-1, so $M_i = |\mu_i\rangle\langle\mu_i|$ for some $|\mu_i\rangle$. Note that

$$\begin{aligned} 1 = \mathrm{Tr}(\rho) &= \mathrm{Tr}\Big(\sum_{i\in[m]} |\mu_i\rangle\langle\mu_i|\rho^{1/2} \sum_{j\in[m]} |\mu_j\rangle\langle\mu_j|\rho^{1/2} \Big) \\ &= \sum_{i,j\in[m]} |\langle\mu_i|\rho^{1/2}|\mu_j\rangle|^2 \\ &\geq \sum_{i\in[m]} \langle\mu_i|\rho^{1/2}|\mu_i\rangle^2. \end{aligned} \tag{5}$$

Then, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned} P^{opt}(\mathcal{E}) = \sum_{i\in[m]} |\langle\mu_i|\psi_i'\rangle|^2 &= \sum_{i\in[m]} |\langle\mu_i|\rho^{1/4}\rho^{-1/4}|\psi_i'\rangle|^2 \\ &\leq \sum_{i\in[m]} \langle\mu_i|\rho^{1/2}|\mu_i\rangle\langle\psi_i'|\rho^{-1/2}|\psi_i'\rangle \\ &\leq \sqrt{\sum_{i\in[m]} \langle\mu_i|\rho^{1/2}|\mu_i\rangle^2} \sqrt{\sum_{i\in[m]} \langle\psi_i'|\rho^{-1/2}|\psi_i'\rangle^2} \\ &\overset{\text{Eq. (5)}}{\leq} \sqrt{\sum_{i\in[m]} \langle\psi_i'|\rho^{-1/2}|\psi_i'\rangle^2} \\ &= \sqrt{P^{PGM}(\mathcal{E})}. \end{aligned}$$

The above shows that for all ensembles $\mathcal{E}$, the PGM for that ensemble is not much worse than the optimal measurement. In some cases the PGM *is* the optimal measurement. In particular, an ensemble $\mathcal{E}$ is called *geometrically uniform* if $\mathcal{E} = \{U_i|\varphi\rangle : i \in [m]\}$ for some Abelian group of matrices $\{U_i\}_{i\in[m]}$ and state $|\varphi\rangle$. Eldar and Forney [EF01] showed $P^{opt}(\mathcal{E}) = P^{PGM}(\mathcal{E})$ for such $\mathcal{E}$.

2.7 Known results and required claims

The following theorems characterize the sample complexity of classical PAC and agnostic learning.

Theorem 1 ([BEHW89, Han16]). *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d+1$. In the PAC model, $\Theta\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ examples are necessary and sufficient for a classical (ε,δ)-PAC learner for $\mathscr{C}$.*

Theorem 2 ([VC74, Sim96, Tal94]). *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d$. In the agnostic model, $\Theta\left(\frac{d}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$ examples are necessary and sufficient for a classical (ε,δ)-agnostic learner for $\mathscr{C}$.*

We will use the following well-known theorem from the theory of error-correcting codes:

Theorem 3. *For every sufficiently large integer n, there exists an integer $k \in [n/4, n]$ and a matrix $M \in \mathbb{F}_2^{n\times k}$ of rank k, such that the associated $[n,k,d]_2$ linear code $\{Mx : x \in \{0,1\}^k\}$ has minimal distance $d \geq n/8$.*

We will need the following claims later

Claim 4. *Let $f : \{0,1\}^m \to \mathbb{R}$ and let $M \in \mathbb{F}_2^{m\times k}$. Then the Fourier coefficients of $f \circ M$ are $\widehat{f \circ M}(Q) = \sum_{S\in\{0,1\}^m : M^t S = Q} \widehat{f}(S)$ for all $Q \subseteq [k]$ (where M^t is the transpose of the matrix M).*

Proof. Writing out the Fourier coefficients of $f \circ M$

$$\begin{aligned}
\widehat{f \circ M}(Q) &= \mathop{\mathbb{E}}_{z\in\{0,1\}^k}[(f \circ M)(z)(-1)^{Q\cdot z}] \\
&= \mathop{\mathbb{E}}_{z\in\{0,1\}^k}\Big[\sum_{S\in\{0,1\}^m} \widehat{f}(S)(-1)^{S\cdot(Mz)+Q\cdot z}\Big] && \text{(Fourier expansion of } f\text{)} \\
&= \sum_{S\in\{0,1\}^m} \widehat{f}(S) \mathop{\mathbb{E}}_{z\in\{0,1\}^k}[(-1)^{(M^tS+Q)\cdot z}] && \text{(using } \langle S, Mz\rangle = \langle M^tS, z\rangle\text{)} \\
&= \sum_{S : M^tS=Q} \widehat{f}(S). && \text{(using } \mathbb{E}_{z\in\{0,1\}^k}(-1)^{(z_1+z_2)\cdot z} = \delta_{z_1,z_2}\text{)}
\end{aligned}$$

□

Claim 5. $\max\{(c/\sqrt{t})^t : t \in [1, c^2]\} = e^{c^2/(2e)}$.

Proof. The value of t at which the function $\Big(c/\sqrt{t}\Big)^t$ is the largest, is obtained by differentiating the function with respect to t,

$$\frac{d}{dt}\Big(c/\sqrt{t}\Big)^t = (c/\sqrt{t})^t\Big(\ln(c/\sqrt{t}) - 1/2\Big).$$

Equating the derivative to zero we obtain the maxima (the second derivative can be checked to be negative) at $t = c^2/e$. □

Fact 6. *For all $\varepsilon \in [0, 1/2]$ we have $H(\varepsilon) \le O(\varepsilon \log(1/\varepsilon))$, and (from the Taylor series)*

$$1 - H(1/2 + \varepsilon) \le 2\varepsilon^2/\ln 2 + O(\varepsilon^4).$$

Fact 7. *For every positive integer n, we have that $\binom{n}{k} \le 2^{nH(k/n)}$ for all $k \le n$ and $\sum_{i=0}^{m}\binom{n}{i} \le 2^{nH(m/n)}$ for all $m \le n/2$.*

The following facts are well-known in quantum information theory.

Fact 8. *Let binary random variable $\mathbf{b} \in \{0,1\}$ be uniformly distributed. Suppose an algorithm is given $|\psi_{\mathbf{b}}\rangle$ (for unknown b) and is required to guess whether $\mathbf{b} = 0$ or $\mathbf{b} = 1$. It will guess correctly with probability at most $\frac{1}{2} + \frac{1}{2}\sqrt{1 - |\langle\psi_0|\psi_1\rangle|^2}$.*

Note that if we could distinguish between the states $|\psi_0\rangle$ and $|\psi_1\rangle$ with probability $\ge 1 - \delta$, then $|\langle\psi_0|\psi_1\rangle| \le 2\sqrt{\delta(1-\delta)}$.

Fact 9. *(Subadditivity of quantum entropy): For an arbitrary bipartite state ρ_{AB} on the Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$, it holds that $S(\rho_{AB}) \le S(\rho_A) + S(\rho_B)$.*

3 Information-theoretic lower bounds

Upper bounds on sample complexity carry over from classical to quantum PAC learning, because a quantum example becomes a classical example if we just measure it. Our main goal is to show that the *lower* bounds also carry over. All our lower bounds will involve two terms, one that

is independent of $\mathscr{C}$ and one that is dependent on the VC dimension of $\mathscr{C}$. In Section 3.1 we prove the VC-independent part of the lower bounds for the *quantum* setting (which also is a lower bound for the classical setting), in Section 3.2 we present an information-theoretic lower bound on sample complexity for PAC learning and agnostic learning which yields optimal VC-dependent bounds in the classical case. Using similar ideas, in Section 3.3 we obtain near-optimal bounds in the quantum case.

3.1 VC-independent part of lower bounds

Lemma 10 ([AS05])**.** *Let $\mathscr{C}$ be a non-trivial concept class. For every $\delta \in (0,1/2)$, $\varepsilon \in (0,1/4)$, a (ε,δ)-PAC quantum learner for $\mathscr{C}$ has sample complexity $\Omega(\frac{1}{\varepsilon}\log\frac{1}{\delta})$.*

Proof. Since $\mathscr{C}$ is non-trivial, we may assume there are two concepts $c_1, c_2 \in \mathscr{C}$ defined on two inputs $\{x_1, x_2\}$ as follows $c_1(x_1) = c_2(x_1) = 0$ and $c_1(x_2) = 0, c_2(x_2) = 1$. Consider the distribution $D(x_1) = 1-\varepsilon$ and $D(x_2) = \varepsilon$. For $i \in \{1,2\}$, the state of the algorithm after T queries to $\mathrm{QPEX}(c_i, D)$ is $|\psi_i\rangle = (\sqrt{1-\varepsilon}|x_1, 0\rangle + \sqrt{\varepsilon}|x_2, c_i(x_2)\rangle)^{\otimes T}$. It follows that $\langle\psi_1|\psi_2\rangle = (1-\varepsilon)^T$. Since the success probability of an (ε,δ)-PAC quantum learner is $\geq 1-\delta$, Fact 8 implies $\langle\psi_1|\psi_2\rangle \leq 2\sqrt{\delta(1-\delta)}$. Hence $T = \Omega(\frac{1}{\varepsilon}\log\frac{1}{\delta})$. □

Lemma 11. *Let $\mathscr{C}$ be a non-trivial concept class. For every $\delta \in (0,1/2)$, $\varepsilon \in (0,1/4)$, a (ε,δ)-agnostic quantum learner for $\mathscr{C}$ has sample complexity $\Omega(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$.*

Proof. Since $\mathscr{C}$ is non-trivial, we may assume there are two concepts $c_1, c_2 \in \mathscr{C}$ and there exists an input $x \in \{0,1\}^n$ such that $c_1(x) \neq c_2(x)$. Consider the two distributions D_- and D_+ defined as follows: $D_\pm(x, c_1(x)) = (1 \pm \varepsilon)/2$ and $D_\pm(x, c_2(x)) = (1 \mp \varepsilon)/2$. Let $|\psi_\pm\rangle$ be the state after T queries to $\mathrm{QAEX}(D_\pm)$, i.e., $|\psi_\pm\rangle = (\sqrt{(1\pm\varepsilon)/2}|x, c_1(x)\rangle + \sqrt{(1\mp\varepsilon)/2}|x, c_2(x)\rangle)^{\otimes T}$. It follows that $\langle\psi_+|\psi_-\rangle = (1-\varepsilon^2)^{T/2}$. Since the success probability of an (ε,δ)-agnostic quantum learner is $\geq 1-\delta$, Fact 8 implies $\langle\psi_+|\psi_-\rangle \leq 2\sqrt{\delta(1-\delta)}$. Hence $T = \Omega(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ □

3.2 Information-theoretic lower bounds on sample complexity: classical case

3.2.1 Optimal lower bound for classical PAC learning

Theorem 12. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d+1$. Then for every $\delta \in (0,1/2)$ and $\varepsilon \in (0,1/4)$, every (ε,δ)-PAC learner for $\mathscr{C}$ has sample complexity $\Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$.*

Proof. Consider an (ε,δ)-PAC learner for $\mathscr{C}$ that uses T examples. The d-independent part of the lower bound, $T = \Omega(\log(1/\delta)/\varepsilon)$, even holds for quantum examples and was proven in Lemma 10. Hence it remains to prove $T = \Omega(d/\varepsilon)$. It suffices to show this for a specific distribution D, defined as follows. Let $\mathcal{S} = \{s_0, s_1, \ldots, s_d\} \subseteq \{0,1\}^n$ be some $(d+1)$-element set shattered by $\mathscr{C}$. Define $D(s_0) = 1-4\varepsilon$ and $D(s_i) = 4\varepsilon/d$ for all $i \in [d]$.

Because $\mathcal{S}$ is shattered by $\mathscr{C}$, for each string $a \in \{0,1\}^d$, there exists a concept $c_a \in \mathscr{C}$ such that $c_a(s_0) = 0$ and $c_a(s_i) = a_i$ for all $i \in [d]$. We define two correlated random variables $\mathbf{A}$ and $\mathbf{B}$ corresponding to the concept and to the examples, respectively. Let $\mathbf{A}$ be a random variable that is uniformly distributed over $\{0,1\}^d$; if $\mathbf{A} = a$, let $\mathbf{B} = \mathbf{B}_1 \ldots \mathbf{B}_T$ be T i.i.d. examples from c_a according to D. We give the following three-step analysis of these random variables:

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

   Proof. Let random variable $h(\mathbf{B}) \in \{0,1\}^d$ be the hypothesis that the learner produces (given the examples in $\mathbf{B}$) restricted to the elements $s_1, \ldots, s_d$. Note that the error of the hypothesis

$\mathrm{err}_D(h(\mathbf{B}), c_{\mathbf{A}})$ equals $d_H(\mathbf{A}, h(\mathbf{B})) \cdot 4\varepsilon/d$, because each s_i where $\mathbf{A}$ and $h(\mathbf{B})$ differ contributes $D(s_i) = 4\varepsilon/d$ to the error. Let $\mathbf{Z}$ be the indicator random variable for the event that the error is $\leq \varepsilon$. If $\mathbf{Z} = 1$, then $d_H(\mathbf{A}, h(\mathbf{B})) \leq d/4$. Since we are analyzing an (ε, δ)-PAC learner, we have $\Pr[\mathbf{Z} = 1] \geq 1 - \delta$, and $H(\mathbf{Z}) \leq H(\delta)$. Given a string $h(\mathbf{B})$ that is $d/4$-close to $\mathbf{A}$, $\mathbf{A}$ ranges over a set of only $\sum_{i=0}^{d/4} \binom{d}{i} \leq 2^{H(1/4)d}$ possible d-bit strings (using Fact 7), hence $H(\mathbf{A} \mid \mathbf{B}, \mathbf{Z} = 1) \leq H(\mathbf{A} \mid h(\mathbf{B}), \mathbf{Z} = 1) \leq H(1/4)d$. We now lower bound $I(\mathbf{A} : \mathbf{B})$ as follows:

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

2. $I(\mathbf{A} : \mathbf{B}) \leq T \cdot I(\mathbf{A} : \mathbf{B}_1)$.

   Proof. This inequality is essentially due to Jain and Zhang [JZ09, Lemma 5], we include the proof for completeness.

$$\begin{aligned} I(\mathbf{A} : \mathbf{B}) = H(\mathbf{B}) - H(\mathbf{B} \mid \mathbf{A}) &= H(\mathbf{B}) - \sum_{i=1}^{T} H(\mathbf{B}_i \mid \mathbf{A}) \\ &\leq \sum_{i=1}^{T} H(\mathbf{B}_i) - \sum_{i=1}^{T} H(\mathbf{B}_i \mid \mathbf{A}) = \sum_{i=1}^{T} I(\mathbf{A} : \mathbf{B}_i), \end{aligned}$$

   where the second equality used independence of the $\mathbf{B}_i$'s conditioned on $\mathbf{A}$, and the inequality uses Fact 9. Since $I(\mathbf{A} : \mathbf{B}_i) = I(\mathbf{A} : \mathbf{B}_1)$ for all i, we get the inequality.

3. $I(\mathbf{A} : \mathbf{B}_1) = 4\varepsilon$.

   Proof. View $\mathbf{B}_1 = (\mathbf{I}, \mathbf{L})$ as consisting of an index $\mathbf{I} \in \{0, 1, \ldots, d\}$ and a corresponding label $\mathbf{L} \in \{0, 1\}$. With probability $1 - 4\varepsilon$, $(\mathbf{I}, \mathbf{L}) = (0, 0)$. For each $i \in [d]$, with probability $4\varepsilon/d$, $(\mathbf{I}, \mathbf{L}) = (i, \mathbf{A}_i)$. Note that $I(\mathbf{A} : \mathbf{I}) = 0$ because $\mathbf{I}$ is independent of $\mathbf{A}$; $I(\mathbf{A} : \mathbf{L} \mid \mathbf{I} = 0) = 0$; and $I(\mathbf{A} : \mathbf{L} \mid \mathbf{I} = i) = I(\mathbf{A}_i : \mathbf{L} \mid \mathbf{I} = i) = H(\mathbf{A}_i \mid \mathbf{I} = i) - H(\mathbf{A}_i \mid \mathbf{L}, \mathbf{I} = i) = 1 - 0 = 1$ for all $i \in [d]$. We have

$$I(\mathbf{A} : \mathbf{B}_1) = I(\mathbf{A} : \mathbf{I}) + I(\mathbf{A} : \mathbf{L} \mid \mathbf{I}) = \sum_{i=1}^{d} \Pr[\mathbf{I} = i] \cdot I(\mathbf{A} : \mathbf{L} \mid \mathbf{I} = i) = 4\varepsilon.$$

Combining these three steps implies $T = \Omega(d/\varepsilon)$. □

3.2.2 Optimal lower bound for classical agnostic learning

Theorem 13. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d$. Then for every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1/4)$, every (ε, δ)-agnostic learner for $\mathscr{C}$ has sample complexity $\Omega\left(\frac{d}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$.*

Proof. The d-independent part of the lower bound, $T = \Omega(\log(1/\delta)/\varepsilon^2)$, even holds for quantum examples and was proven in Lemma 11. For the other part, the proof is similar to Theorem 12, as follows. Assume an (ε, δ)-agnostic learner for $\mathscr{C}$ that uses T examples. We need to prove $T = \Omega(d/\varepsilon^2)$. For shattered set $\mathcal{S} = \{s_1, \ldots, s_d\} \subseteq \{0, 1\}^n$ and $a \in \{0, 1\}^d$, define distribution D_a on $[d] \times \{0, 1\}$ by $D_a(i, \ell) = (1 + (-1)^{a_i + \ell} 4\varepsilon)/2d$.

Again let random variable $\mathbf{A} \in \{0,1\}^d$ be a uniformly distributed random variable, corresponding to the values of concept c_a on $\mathcal{S}$, and $\mathbf{B} = \mathbf{B}_1 \dots \mathbf{B}_T$ be T i.i.d. samples from D_a. Note that c_a is the minimal-error concept from $\mathscr{C}$ w.r.t. D_a, and concept $c_{\tilde{a}}$ has additional error $d_H(a,\tilde{a}) \cdot 4\varepsilon/d$. Accordingly, an (ε,δ)-agnostic learner has to produce (from $\mathbf{B}$) an $h(\mathbf{B}) \in \{0,1\}^d$, which, with probability at least $1-\delta$, is $d/4$-close to $\mathbf{A}$. Our three-step analysis is very similar to Theorem 12; only the third step changes:

1. $I(\mathbf{A}:\mathbf{B}) \geq (1-\delta)(1-H(1/4))d - H(\delta) = \Omega(d)$.
2. $I(\mathbf{A}:\mathbf{B}) \leq T \cdot I(\mathbf{A}:\mathbf{B}_1)$.
3. $I(\mathbf{A}:\mathbf{B}_1) = 1 - H(1/2+2\varepsilon) = O(\varepsilon^2)$.

   Proof. View the D_a-distributed random variable $\mathbf{B}_1 = (\mathbf{I},\mathbf{L})$ as index $\mathbf{I} \in [d]$ and label $\mathbf{L} \in \{0,1\}$. The marginal distribution of $\mathbf{I}$ is uniform; conditioned on $\mathbf{I} = i$, the bit $\mathbf{L}$ equals $\mathbf{A}_i$ with probability $1/2+2\varepsilon$. Hence

$$I(\mathbf{A}:\mathbf{L} \mid \mathbf{I}=i) = I(\mathbf{A}_i:\mathbf{L} \mid \mathbf{I}=i) = H(\mathbf{A}_i \mid \mathbf{I}=i) - H(\mathbf{A}_i \mid \mathbf{L},\mathbf{I}=i) = 1 - H(1/2+2\varepsilon).$$

   Using Fact 6, we have

$$\begin{aligned} I(\mathbf{A}:\mathbf{B}_1) = I(\mathbf{A}:\mathbf{I}) + I(\mathbf{A}:\mathbf{L} \mid \mathbf{I}) &= \sum_{i=1}^{d} \Pr[\mathbf{I}=i] \cdot I(\mathbf{A}:\mathbf{L} \mid \mathbf{I}=i) \\ &= 1 - H(1/2+2\varepsilon) = O(\varepsilon^2). \end{aligned}$$

Combining these three steps implies $T = \Omega(d/\varepsilon^2)$. □

In the theorem below, we optimize the constant in the lower bound of the sample complexity in Theorem 13. In learning theory such lower bounds are often stated slightly differently. In order to compare the lower bounds, we introduce the following. We first define an *ε-average agnostic learner* for a concept class $\mathscr{C}$ as a learner that, given access to T samples from an AEX(D) oracle (for some unknown distribution D), needs to output a hypothesis $h_{\mathbf{XY}}$ (where $(\mathbf{X},\mathbf{Y}) \sim D^T$) that satisfies

$$\mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\sim D^T}[\mathrm{err}_D(h_{\mathbf{XY}})] - \mathrm{opt}_D(\mathscr{C}) \leq \varepsilon.$$

Lower bounds on the quantity $(\mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim D^T}[\mathrm{err}_D(h_{\mathbf{XY}})] - \mathrm{opt}_D(\mathscr{C}))$ are generally referred to as *minimax lower bounds* in learning theory. For concept class $\mathscr{C}$, Audibert [Aud08, Aud09] showed that there exists a distribution D, such that if the agnostic learner uses T samples from AEX(D), then

$$\mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\sim D^T}[\mathrm{err}_D(h_{\mathbf{XY}})] - \mathrm{opt}_D(\mathscr{C}) \geq \frac{1}{6}\sqrt{\frac{d}{T}}.$$

Equivalently, this is a lower bound of $T \geq \frac{d}{36\varepsilon^2}$ on the sample complexity of an ε-average agnostic learner. We obtain a slightly weaker lower bound that is essentially $T \geq \frac{d}{62\varepsilon^2}$:

Theorem 14. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d$. Then for every $\varepsilon \in (0, 1/10]$, there exists a distribution for which every ε-average agnostic learner has sample complexity at least $\frac{d}{\varepsilon^2} \cdot \left(\frac{1}{62} - \frac{\log(2d+2)}{4d}\right)$.*

Proof. The proof is similar to Theorem 13. Assume an ε-average agnostic learner for $\mathscr{C}$ that uses T samples. For shattered set $\mathcal{S} = \{s_1, \ldots, s_d\} \subseteq \{0,1\}^n$ and $a \in \{0,1\}^d$, define distribution D_a on $[d] \times \{0,1\}$ by $D_a(i,\ell) = (1 + (-1)^{a_i + \ell}\beta\varepsilon)/2d$, for some constant $\beta \geq 2$ which we shall pick later.

Again let random variable $\mathbf{A} \in \{0,1\}^d$ be uniformly random, corresponding to the values of concept c_a on $\mathcal{S}$, and $\mathbf{B} = \mathbf{B}_1 \ldots \mathbf{B}_T$ be T i.i.d. samples from D_a. Note that c_a is the minimal-error concept from $\mathscr{C}$ w.r.t. D_a, and concept $c_{\tilde{a}}$ has additional error $d_H(a,\tilde{a}) \cdot \beta\varepsilon/d$. Accordingly, an ε-average agnostic learner has to produce (from $\mathbf{B}$) an $h(\mathbf{B}) \in \{0,1\}^d$, which satisfies $\mathbb{E}_{\mathbf{A},\mathbf{B}}[d_H(\mathbf{A}, h(\mathbf{B}))] \leq d/\beta$.

Our three-step analysis is very similar to Theorem 13; only the first step changes:

1. $I(\mathbf{A} : \mathbf{B}) \geq d(1 - H(1/\beta)) - \log(d+1)$.

   Proof. Define random variable $\mathbf{Z} = d_H(\mathbf{A}, h(\mathbf{B}))$, then $\mathbb{E}[\mathbf{Z}] \leq d/\beta$. Note that given a string $h(\mathbf{B})$ that is ℓ-close to $\mathbf{A}$, $\mathbf{A}$ ranges over a set of only $\binom{d}{\ell} \leq 2^{H(\ell/d)d}$ possible d-bit strings (using Fact 7), hence $H(\mathbf{A} \mid \mathbf{B}, \mathbf{Z} = \ell) \leq H(\mathbf{A} \mid h(\mathbf{B}), \mathbf{Z} = \ell) \leq H(\ell/d)d$. We now lower bound $I(\mathbf{A} : \mathbf{B})$

$$\begin{aligned}
I(\mathbf{A} : \mathbf{B}) &= H(\mathbf{A}) - H(\mathbf{A} \mid \mathbf{B}) \\
&\geq H(\mathbf{A}) - H(\mathbf{A} \mid \mathbf{B}, \mathbf{Z}) - H(\mathbf{Z}) \\
&= d - \sum_{\ell=0}^{d+1} \Pr[\mathbf{Z} = \ell] \cdot H(\mathbf{A} \mid \mathbf{B}, \mathbf{Z} = \ell) - H(\mathbf{Z}) \\
&\geq d - \mathop{\mathbb{E}}_{\ell \in \{0,\ldots,d\}} [H(\ell/d)d] - \log(d+1) && \text{(since } \mathbf{Z} \in \{0,\ldots,d\}\text{)} \\
&\geq d - dH\left(\frac{\mathbb{E}_\ell[\ell]}{d}\right) - \log(d+1) && \text{(using Jensen's inequality)} \\
&\geq d - dH(1/\beta) - \log(d+1), && \text{(using } \mathbb{E}[\mathbf{Z}] \leq d/\beta\text{)}
\end{aligned}$$

   where for the third inequality we used the concavity of the binary entropy function to conclude $\mathbb{E}_\ell[H(\ell/d)] \leq H(\mathbb{E}_\ell[\ell]/d)$, and for the fourth inequality we used that $\beta \geq 2$.

2. $I(\mathbf{A} : \mathbf{B}) \leq T \cdot I(\mathbf{A} : \mathbf{B}_1)$.

3. $I(\mathbf{A} : \mathbf{B}_1) = 1 - H(1/2 + \beta\varepsilon/2) \overset{\text{Fact } 6}{\leq} \beta^2\varepsilon^2/\ln 4 + O(\varepsilon^4)$.

Combining these three steps implies

$$T \geq \frac{d \ln 4}{\varepsilon^2} \cdot \left(\frac{1 - H(1/\beta)}{\beta^2 + O(\varepsilon^2)} - \frac{\log(d+1)}{\beta^2 d + O(d\varepsilon^2)} \right).$$

Using $\varepsilon \leq 1/10$, $\beta = 4$ to optimize this lower bound, we obtain $T \geq \frac{d}{\varepsilon^2} \cdot \left(\frac{1}{62} - \frac{\log(2d+2)}{4d} \right)$. □

3.3 Information-theoretic lower bounds on sample complexity: quantum case

Here we will "quantize" the above two classical information-theoretic proofs, yielding lower bounds for quantum sample complexity (in both the PAC and the agnostic setting) that are tight up to a logarithmic factor.

3.3.1 Near-optimal lower bound for quantum PAC learning

Theorem 15. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d + 1$. Then, for every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1/4)$, every (ε, δ)-PAC quantum learner for $\mathscr{C}$ has sample complexity $\Omega\left(\frac{d}{\varepsilon \log(d/\varepsilon)} + \frac{\log(1/\delta)}{\varepsilon} \right)$.*

Proof. The proof is analogous to Theorem 12. We use the same distribution D, with the $\mathbf{B}_i$ now being quantum samples: $|\psi_a\rangle = \sum_{i\in\{0,1,\ldots,d\}} \sqrt{D(s_i)}|i, c_a(s_i)\rangle$. The $\mathbf{AB}$-system is now in the following classical-quantum state:

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

The first two steps of our argument are identical to Theorem 12. We only need to re-analyze step 3:

1. $I(\mathbf{A}:\mathbf{B}) \geq (1-\delta)(1-H(1/4))d - H(\delta) = \Omega(d)$.
2. $I(\mathbf{A}:\mathbf{B}) \leq T \cdot I(\mathbf{A}:\mathbf{B}_1)$.
3. $I(\mathbf{A}:\mathbf{B}_1) \leq H(4\varepsilon) + 4\varepsilon\log(2d) = O(\varepsilon\log(d/\varepsilon))$.

   Proof. Since $\mathbf{AB}$ is a classical-quantum state, we have

   $$I(\mathbf{A}:\mathbf{B}_1) = S(\mathbf{A}) + S(\mathbf{B}_1) - S(\mathbf{AB}_1) = S(\mathbf{B}_1),$$

   where the first equality follows from definition and the second equality uses $S(\mathbf{A}) = d$ since $\mathbf{A}$ is uniformly distributed in $\{0,1\}^d$, and $S(\mathbf{AB}_1) = d$ 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 $\mathbf{B}_1$, 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 $D(s_0) = 1-4\varepsilon$, hence so is the upper-left entry of ρ. This implies $\sigma_0 \geq 1-4\varepsilon$. Consider sampling a number $\mathbf{N} \in \{0,1,\ldots,2d\}$ according to the σ-distribution. Let $\mathbf{Z}$ be the indicator random variable for the event $\mathbf{N} \neq 0$, which has probability $1-\sigma_0 \leq 4\varepsilon$. Note that $H(\mathbf{N} \mid \mathbf{Z} = 0) = 0$, because $\mathbf{Z} = 0$ implies $\mathbf{N} = 0$. Also, $H(\mathbf{N} \mid \mathbf{Z} = 1) \leq \log(2d)$, because if $\mathbf{Z} = 1$ then $\mathbf{N}$ ranges over $2d$ elements. We now have

   $$\begin{aligned} S(\rho) = H(\mathbf{N}) &= H(\mathbf{N},\mathbf{Z}) = H(\mathbf{Z}) + H(\mathbf{N} \mid \mathbf{Z}) \\ &= H(\mathbf{Z}) + \Pr[\mathbf{Z}=0]\cdot H(\mathbf{N} \mid \mathbf{Z}=0) + \Pr[\mathbf{Z}=1]\cdot H(\mathbf{N} \mid \mathbf{Z}=1) \\ &\leq H(4\varepsilon) + 4\varepsilon\log(2d) \\ &= O(\varepsilon\log(d/\varepsilon)). \qquad \text{(using Fact 6)} \end{aligned}$$

Combining these three steps implies $T = \Omega\left(\frac{d}{\varepsilon\log(d/\varepsilon)}\right)$. ∎

3.3.2 Near-optimal lower bound for quantum agnostic learning

Theorem 16. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d$. Then for every $\delta \in (0,1/2)$ and $\varepsilon \in (0,1/4)$, every (ε,δ)-agnostic quantum learner for $\mathscr{C}$ has sample complexity $\Omega\left(\frac{d}{\varepsilon^2\log(d/\varepsilon)} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$.*

Proof. The proof is analogous to Theorem 13, with the $\mathbf{B}_i$ now being quantum samples for D_a, $|\psi_a\rangle = \sum_{i\in[d],\ell\in\{0,1\}} \sqrt{D_a(i,\ell)}|i,\ell\rangle$. Again we only need to re-analyze step 3:

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

2. $I(\mathbf{A}:\mathbf{B}) \leq T \cdot I(\mathbf{A}:\mathbf{B}_1)$.

3. $I(\mathbf{A}:\mathbf{B}_1) = O(\varepsilon^2 \log(d/\varepsilon))$.
   Proof (of step 3). As in step 3 of the proof of Theorem 15, it suffices to upper bound the entropy of
   $$\rho = \frac{1}{2^d} \sum_{a\in\{0,1\}^d} |\psi_a\rangle\langle\psi_a|.$$
   We now lower bound the largest singular value of ρ. Consider $|\psi\rangle = \frac{1}{\sqrt{2d}} \sum_{i\in[d],\ell\in\{0,1\}} |i,\ell\rangle$.
   $$\langle\psi|\psi_a\rangle = \frac{1}{d} \sum_{i\in[d]} \frac{1}{2}\left(\sqrt{1+4\varepsilon} + \sqrt{1-4\varepsilon}\right) = \frac{1}{2}\left(\sqrt{1+4\varepsilon} + \sqrt{1-4\varepsilon}\right) \geq 1 - 2\varepsilon^2 - O(\varepsilon^4),$$
   where the last inequality used the Taylor series expansion of $\sqrt{1+x}$. This implies that the largest singular value of ρ is at least
   $$\langle\psi|\rho|\psi\rangle = \frac{1}{2^d} \sum_{a\in\{0,1\}^d} |\langle\psi|\psi_a\rangle|^2 \geq 1 - 4\varepsilon^2 - O(\varepsilon^4).$$
   We can now finish as in step 3 of the proof of Theorem 15:
   $$I(\mathbf{A}:\mathbf{B}_1) \leq S(\rho) \leq H(4\varepsilon^2) + 4\varepsilon^2 \log(2d) \overset{\text{Fact } 6}{=} O(\varepsilon^2 \log(d/\varepsilon)).$$

Combining these three steps implies $T = \Omega\left(\frac{d}{\varepsilon^2 \log(d/\varepsilon)}\right)$. □

4 A lower bound by analysis of state identification

In this section we present a tight lower bound on quantum sample complexity for both the PAC and the agnostic learning models, using ideas from Fourier analysis to analyze the performance of the Pretty Good Measurement. The core of both lower bounds is the following combinatorial theorem.

Theorem 17. *For $m \geq 10$, let $f : \{0,1\}^m \to \mathbb{R}$ be defined as $f(z) = (1 - \beta\frac{|z|}{m})^T$ for some $\beta \in (0,1]$ and $T \in [1, m/(e^3\beta)]$. For $k \leq m$, let $M \in \mathbb{F}_2^{m\times k}$ be a matrix with rank k. Suppose $A \in \mathbb{R}^{2^k\times 2^k}$ is defined as $A(x,y) = (f \circ M)(x+y)$ for $x, y \in \{0,1\}^k$, then*
$$\sqrt{A}(x,x) \leq \frac{2\sqrt{e}}{2^{k/2}}\left(1 - \frac{\beta}{2}\right)^{T/2} e^{11T^2\beta^2/m + \sqrt{Tm\beta}} \qquad \text{for all } x \in \{0,1\}^k.$$

Proof. The structure of the proof is to first diagonalize A, relating its eigenvalues to the Fourier coefficients of f. This allows to calculate the diagonal entries of $\sqrt{A}$ exactly in terms of those Fourier coefficients. We then upper bound those Fourier coefficients using a combinatorial argument.

We first observe the well-known relation between the eigenvalues of a matrix P defined as $P(x,y) = g(x+y)$ for $x, y \in \{0,1\}^k$, and the Fourier coefficients of g.

Claim 18. *Suppose $g : \{0,1\}^k \to \mathbb{R}$ and $P \in \mathbb{R}^{2^k\times 2^k}$ is defined as $P(x,y) = g(x+y)$, then the eigenvalues of P are $\{2^k\widehat{g}(Q) : Q \in \{0,1\}^k\}$.*

Proof. Let $H \in \mathbb{R}^{2^k \times 2^k}$ be the matrix defined as $H(x,y) = (-1)^{x\cdot y}$ for $x, y \in \{0,1\}^k$. It is easy to see that $H^{-1}(x,y) = (-1)^{x\cdot y}/2^k$. We now show that H diagonalizes P:

$$\begin{aligned}
(HPH^{-1})(x,y) &= \frac{1}{2^k} \sum_{z_1,z_2 \in \{0,1\}^k} (-1)^{z_1\cdot x + z_2 \cdot y} g(z_1+z_2) \\
&= \frac{1}{2^k} \sum_{z_1,z_2,Q \in \{0,1\}^k} (-1)^{z_1\cdot x + z_2\cdot y} \widehat{g}(Q)(-1)^{Q\cdot(z_1+z_2)} && \text{(Fourier expansion of } g\text{)} \\
&= \frac{1}{2^k} \sum_{Q\in\{0,1\}^k} \widehat{g}(Q) \sum_{z_1\in\{0,1\}^k} (-1)^{(x+Q)\cdot z_1} \sum_{z_2\in\{0,1\}^k} (-1)^{(y+Q)\cdot z_2} \\
&= 2^k \widehat{g}(x)\delta_{x,y} && \text{(using } \textstyle\sum_{z\in\{0,1\}^k}[(-1)^{(a+b)\cdot z}] = 2^k\delta_{a,b}\text{)}
\end{aligned}$$

The eigenvalues of P are the diagonal entries, $\{2^k\widehat{g}(Q) : Q \in \{0,1\}^k\}$. □

We now relate the diagonal entries of $\sqrt{A}$ to the Fourier coefficients of f:

Claim 19. *For all $x \in \{0,1\}^k$, we have*

$$\sqrt{A}(x,x) = \frac{1}{2^{k/2}} \sum_{Q\in\{0,1\}^k} \sqrt{\sum_{S\in\{0,1\}^m : M^tS = Q} \widehat{f}(S)}.$$

Proof. Since $A(x,y) = (f\circ M)(x+y)$, by Claim 18 it follows that H (as defined in the proof of Claim 18) diagonalizes A and the eigenvalues of A are $\{2^k\widehat{f\circ M}(Q) : Q \in \{0,1\}^k\}$. Hence, we have

$$\sqrt{A} = H^{-1}\cdot \mathrm{diag}\left(\left\{\sqrt{2^k\widehat{f\circ M}(Q)} : Q\in\{0,1\}^k\right\}\right)\cdot H,$$

and the diagonal entries of $\sqrt{A}$ are

$$\sqrt{A}(x,x) = \frac{1}{2^{k/2}} \sum_{Q\in\{0,1\}^k} \sqrt{\widehat{f\circ M}(Q)} \overset{\text{Claim 4}}{=} \frac{1}{2^{k/2}} \sum_{Q\in\{0,1\}^k} \sqrt{\sum_{S\in\{0,1\}^m : M^tS = Q} \widehat{f}(S)}.$$

□

In the following lemma, we give an upper bound on the Fourier coefficients of f, which in turn (from the claim above) gives an upper bound on the diagonal entries of $\sqrt{A}$.

Lemma 20. *For $\beta \in (0,1]$, the Fourier coefficients of $f : \{0,1\}^m \to \mathbb{R}$ defined as $f(z) = (1-\beta\frac{|z|}{m})^T$, satisfy*

$$0 \le \widehat{f}(S) \le 4e\left(1-\frac{\beta}{2}\right)^T \left(\frac{T\beta}{m}\right)^q e^{22T^2\beta^2/m}, \quad \text{for all } S \text{ such that } |S| = q.$$

Proof. In order to see why the Fourier coefficients of f are non-negative, we first define the set $U = \{u_x^{\otimes T}\}_{x\in\{0,1\}^m}$ where $u_x = \sqrt{1-\beta}|0,0\rangle + \sqrt{\beta/m}\sum_{i\in[m]}|i,x_i\rangle$. Let V be the $2^m \times 2^m$ Gram matrix for the set U. For $x, y \in \{0,1\}^m$, we have

$$\begin{aligned}
V(x,y) = (u_x^* u_y)^T &= \left(1-\beta+\frac{\beta}{m}\sum_{i=1}^m \langle x_i|y_i\rangle\right)^T \\
&= \left(1-\beta+\frac{\beta}{m}(m-|x+y|)\right)^T \\
&= \left(1-\beta\frac{|x+y|}{m}\right)^T = f(x+y).
\end{aligned}$$

By Claim 18, the eigenvalues of the Gram matrix V are $\{2^m \widehat{f}(S) : S \in \{0,1\}^m\}$. Since the Gram matrix is psd, its eigenvalues are non-negative, which implies that $\widehat{f}(S) \geq 0$ for all $S \in \{0,1\}^m$.

We now prove the upper bound in the lemma. By definition,

$$\begin{aligned}
\widehat{f}(S) &= \mathop{\mathbb{E}}_{z\in\{0,1\}^m}\left[\left(1-\beta\frac{|z|}{m}\right)^T(-1)^{S\cdot z}\right] \\
&= \mathop{\mathbb{E}}_{z\in\{0,1\}^m}\left[\left(1-\frac{\beta}{2}+\frac{\beta}{2m}\sum_{i=1}^m(-1)^{z_i}\right)^T(-1)^{S\cdot z}\right] && \left(\text{since } |z| = \textstyle\sum_{i\in[m]}\frac{1-(-1)^{z_i}}{2}\right) \\
&= \sum_{\ell=0}^T\binom{T}{\ell}\left(1-\frac{\beta}{2}\right)^{T-\ell}\left(\frac{\beta}{2m}\right)^\ell \mathop{\mathbb{E}}_{z\in\{0,1\}^m}\left[\sum_{i_1,\ldots,i_\ell=1}^m(-1)^{z\cdot(e_{i_1}+\cdots+e_{i_\ell}+S)}\right] \\
&= \sum_{\ell=0}^T\binom{T}{\ell}\left(1-\frac{\beta}{2}\right)^{T-\ell}\left(\frac{\beta}{2m}\right)^\ell \sum_{i_1,\ldots,i_\ell=1}^m 1_{[e_{i_1}+\cdots+e_{i_\ell}=S]} && \left(\text{using } \mathbb{E}_{z\in\{0,1\}^m}[(-1)^{(z_1+z_2)\cdot z}] = \delta_{z_1,z_2}\right)
\end{aligned}$$

We will use the following claim to upper bound the combinatorial sum in the quantity above.

Claim 21. *Fix $S \in \{0,1\}^m$ with Hamming weight $|S| = q$. For every $\ell \in \{q,\ldots,T\}$, we have*

$$\sum_{i_1,\ldots,i_\ell=1}^m 1_{[e_{i_1}+\cdots+e_{i_\ell}=S]} \leq \begin{cases} \ell!\cdot m^{(\ell-q)/2}\Big/\left(2^{(\ell-q)/2}((\ell-q)/2)!\right) & \text{if } (\ell-q) \text{ is even} \\ 0 & \text{otherwise} \end{cases}$$

Proof. Since $|S| = q$, we can write $S = e_{r_1}+\cdots+e_{r_q}$ for distinct $r_1,\ldots,r_q \in [m]$. There are $\binom{\ell}{q}$ ways to pick q indices in $(i_1,\ldots,i_\ell)$ (w.l.o.g. let them be $i_1,\ldots,i_q$) and there are $q!$ factorial ways to assign $(r_1,\ldots,r_q)$ to $(i_1,\ldots,i_q)$. It remains to count the number of ways that we can assign values to the remaining indices $i_{q+1},\ldots,i_\ell$ such that $e_{i_{q+1}}+\cdots+e_{i_\ell} = 0$. If $\ell-q$ is odd then this number is 0, so from now on assume $\ell-q$ is even. We upper bound the number of such assignments by partitioning the $\ell-q$ indices into pairs and assigning the same value to both indices in each pair.

We first count the number of ways to partition a set of $\ell-q$ indices into subsets of size 2. This number is exactly $(\ell-q)!\left(2^{(\ell-q)/2}((\ell-q)/2)!\right)^{-1}$. Furthermore, there are m possible values that can be assigned to the pair of indices in each of the $(\ell-q)/2$ subsets such that $e_i+e_j = 0$ within each subset. Note that assigning m possible values to each pair of indices in the $(\ell-q)/2$ subsets overcounts, but this rough upper bound is sufficient for our purposes.

Combining the three arguments, we conclude

$$\sum_{i_1,\ldots,i_\ell=1}^d 1_{[e_{i_1}+\cdots+e_{i_\ell}=S]} \leq \binom{\ell}{q}q!\cdot(\ell-q)!\cdot m^{(\ell-q)/2}\Big/\left(2^{(\ell-q)/2}((\ell-q)/2)!\right).$$

which yields the claim. ∎

Continuing with the evaluation of the Fourier coefficient and using the claim above, we have

$$\begin{aligned}
\widehat{f}(S) &= \sum_{\ell=0}^{T}\binom{T}{\ell}\Big(1-\frac{\beta}{2}\Big)^{T-\ell}\Big(\frac{\beta}{2m}\Big)^{\ell}\sum_{i_1,\dots,i_\ell=1}^{m} 1_{[e_{i_1}+\cdots+e_{i_\ell}=S]} \\
&\le \sum_{\ell=q}^{T}\binom{T}{\ell}\Big(1-\frac{\beta}{2}\Big)^{T-\ell}\Big(\frac{\beta}{2m}\Big)^{\ell}\ell!\cdot m^{(\ell-q)/2}\Big/\Big(2^{(\ell-q)/2}\Big(\frac{\ell-q}{2}\Big)!\Big) && \text{(by Claim 21)} \\
&= \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{2}{m}\Big)^{q/2}\sum_{\ell=q}^{T}\binom{T}{\ell}\ell!\Big(\frac{\beta}{m(2-\beta)}\Big)^{\ell}\Big(\frac{m}{2}\Big)^{\ell/2}\Big/\Big(\frac{\ell-q}{2}\Big)! \\
&\le \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{2}{m}\Big)^{q/2}\sum_{\ell=q}^{T}\Big(T\cdot\frac{\beta}{m}\cdot\sqrt{\frac{m}{2}}\Big)^{\ell}\Big/\Big(\frac{\ell-q}{2}\Big)! && \text{(since } \beta<1 \text{ and } \tbinom{T}{\ell}\ell!\le T^{\ell}\text{)} \\
&= \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}\sum_{r=0}^{T-q}\Big(\frac{T\beta}{\sqrt{2m}}\Big)^{r}\frac{1}{(r/2)!} && \text{(substituting } r\leftarrow(\ell-q)\text{)} \\
&\le \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}\sum_{r=0}^{T-q}\Big(\frac{T\beta}{\sqrt{2m}}\Big)^{r}\frac{e^{r/2}}{(r/2)^{r/2}} && \text{(using } n!\ge(n/e)^{n}\text{)} \\
&= \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}\sum_{r=0}^{T-q}\Big(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\Big)^{r} \\
&\le \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}\sum_{r=0}^{T}\Big(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\Big)^{r} && \text{(since the summands are } \ge 0\text{)} \\
&= \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}\Big(\sum_{r=0}^{\lceil e^3T^2\beta^2/m\rceil}\Big(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\Big)^{r}+\sum_{r=\lceil e^3T^2\beta^2/m\rceil+1}^{T}\Big(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\Big)^{r}\Big).
\end{aligned}$$

Note that by the assumptions of the theorem, $T^2e^3\beta^2/m \le T\beta \le T$, which allowed us to split the sum into two pieces in the last equality. At this point, we upper bound both pieces in the last equation separately. For the first piece, using Claim 5 it follows that $\left(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\right)^r$ is maximized at $r = \lceil T^2\beta^2/m\rceil$. Hence we get

$$\sum_{r=0}^{\lceil e^3T^2\beta^2/m\rceil}\Big(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\Big)^{r} \le \Big(2+\frac{e^3T^2\beta^2}{m}\Big)e^{\lceil T^2\beta^2/m\rceil/2} \le 2e^{22T^2\beta^2/m+1}, \tag{6}$$

where the first inequality uses Claim 5 and the second inequality uses $2+x \le 2e^x$ for $x \ge 0$ and $e^3 + 1/2 \le 22$. For the second piece, we use

$$\sum_{r=\lceil e^3T^2\beta^2/m\rceil+1}^{T}\Big(\frac{\sqrt{e}T\beta}{\sqrt{mr}}\Big)^{r} \le \sum_{r=\lceil e^3T^2\beta^2/m\rceil+1}^{T}\Big(\frac{1}{e}\Big)^{r} \le \sum_{r=1}^{T}\Big(\frac{1}{e}\Big)^{r} = \frac{1-e^{-T}}{e-1} \le 2/3. \tag{7}$$

So we finally get

$$\begin{aligned}
\widehat{f}(S) &\le \Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}\Big(2e^{22T^2\beta^2/m+1}+2/3\Big) && \text{(using Eq. (6), (7))} \\
&\le 4e\Big(1-\frac{\beta}{2}\Big)^{T}\Big(\frac{T\beta}{m}\Big)^{q}e^{22T^2\beta^2/m} && \text{(since } 22T^2\beta^2/m>0\text{)}
\end{aligned}$$

□

The theorem follows by putting together Claim 19 and Lemma 20:

$$\begin{aligned}
\sqrt{A}(x,x) &= \frac{1}{2^{k/2}} \sum_{Q\in\{0,1\}^k} \sqrt{\sum_{S\in\{0,1\}^m : M^t S = Q} \widehat{f}(S)} && \text{(using Claim 19)} \\
&\leq \frac{1}{2^{k/2}} \sum_{Q\in\{0,1\}^k} \sum_{S\in\{0,1\}^m : M^t S = Q} \sqrt{\widehat{f}(S)} && \text{(using lower bound from Lemma 20)} \\
&= \frac{1}{2^{k/2}} \sum_{S\in\{0,1\}^m} \sqrt{\widehat{f}(S)} && (\cup_Q \{S : M^t S = Q\} = \{0,1\}^m \text{ since rank}(M)=k) \\
&= \frac{1}{2^{k/2}} \sum_{q=0}^{m} \sum_{S\in\{0,1\}^m : |S|=q} \sqrt{\widehat{f}(S)} \\
&\leq \frac{2\sqrt{e}}{2^{k/2}} \left(1-\frac{\beta}{2}\right)^{T/2} e^{11T^2\beta^2/m} \sum_{q=0}^{m} \binom{m}{q} \left(\frac{T\beta}{m}\right)^{q/2} && \text{(using Lemma 20)} \\
&= \frac{2\sqrt{e}}{2^{k/2}} \left(1-\frac{\beta}{2}\right)^{T/2} e^{11T^2\beta^2/m} \left(1+\sqrt{\frac{T\beta}{m}}\right)^{m} && \text{(using binomial theorem)} \\
&\leq \frac{2\sqrt{e}}{2^{k/2}} \left(1-\frac{\beta}{2}\right)^{T/2} e^{11T^2\beta^2/m+\sqrt{Tm\beta}}. && \text{(using } (1+x)^t \leq e^{xt} \text{ for } x,t\geq 0)
\end{aligned}$$

□

4.1 Optimal lower bound for quantum PAC learning

We can now prove our tight lower bound on quantum sample complexity in the PAC model:

Theorem 22. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d+1$, for sufficiently large d. Then for every $\delta \in (0,1/2)$ and $\varepsilon \in (0,1/20)$, every (ε,δ)-PAC quantum learner for $\mathscr{C}$ has sample complexity $\Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$.*

Proof. The d-independent part of the lower bound is Lemma 10. To prove the d-dependent part, define a distribution D on a set $\mathcal{S} = \{s_0,\ldots,s_d\} \subseteq \{0,1\}^n$ that is shattered by $\mathscr{C}$ as follows: $D(s_0) = 1-20\varepsilon$ and $D(s_i) = 20\varepsilon/d$ for all $i \in [d]$.

Now consider a $[d,k,r]_2$ linear code (for $k \geq d/4$, distance $r \geq d/8$) as shown to exist in Theorem 3 with the generator matrix $M \in \mathbb{F}_2^{d\times k}$ of rank k. Let $\{Mx : x \in \{0,1\}^k\} \subseteq \{0,1\}^d$ be the set of codewords in this linear code; these satisfy $d_H(Mx,My) \geq d/8$ whenever $x \neq y$. For each $x \in \{0,1\}^k$, let c^x be a concept defined on the shattered set as: $c^x(s_0) = 0$ and $c^x(s_i) = (Mx)_i$ for all $i \in [d]$. The existence of such concepts in $\mathscr{C}$ follows from the fact that $\mathcal{S}$ is shattered by $\mathscr{C}$. From the distance property of the code, we have $\Pr_{s\sim D}[c^x(s) \neq c^y(s)] \geq \frac{20\varepsilon}{d}\frac{d}{8} = 5\varepsilon/2$. This in particular implies that an (ε,δ)-PAC quantum learner that tries to ε-approximate a concept from $\{c^x : x \in \{0,1\}^k\}$ should successfully *identify* that concept with probability at least $1-\delta$.

We now consider the following state identification problem: for $x \in \{0,1\}^k$, denote $|\psi_x\rangle = \sum_{i\in\{0,\ldots,d\}} \sqrt{D(s_i)}|s_i, c^x(s_i)\rangle$. Let the (ε,δ)-PAC quantum sample complexity be T. Assume $T \leq d/(20e^3\varepsilon)$, since otherwise $T \geq \Omega(d/\varepsilon)$ and the theorem follows. Suppose the learner has knowledge of the ensemble $\mathcal{E} = \{(2^{-k}, |\psi_x\rangle^{\otimes T}) : x \in \{0,1\}^k\}$, and is given $|\psi_x\rangle^{\otimes T} \in \mathcal{E}$ for a uniformly random x. The learner would like to maximize the average probability of success to identify the given

state. For this problem, we prove a lower bound on T using the PGM defined in Section 2.6. In particular, we show that using the PGM, if a learner successfully identifies the states in $\mathcal{E}$, then $T = \Omega(d/\varepsilon)$. Since the PGM is the optimal measurement[6] that the learner could have performed, the result follows. The following lemma makes this lower bound rigorous and will conclude the proof of the theorem.

Lemma 23. *For every $x \in \{0,1\}^k$, let $|\psi_x\rangle = \sum_{i\in\{0,\dots,d\}} \sqrt{D(s_i)}|s_i, c^x(s_i)\rangle$, and $\mathcal{E} = \{(2^{-k}, |\psi_x\rangle^{\otimes T}) : x \in \{0,1\}^k\}$. Then*[7]

$$P^{PGM}(\mathcal{E}) \leq \frac{4e}{2^{d/4+T\varepsilon}} e^{8800T^2\varepsilon^2/d + 4\sqrt{5Td\varepsilon}}.$$

Before we prove the lemma, we first show why it implies the theorem. Since we observed above that $P^{opt}(\mathcal{E}) = P^{PGM}(\mathcal{E})$, a good learner satisfies $P^{PGM}(\mathcal{E}) = \Omega(1)$ (say for $\delta = 1/4$), which in turn implies

$$\Omega(\max\{d, T\varepsilon\}) \leq O(\min\{T^2\varepsilon^2/d, \sqrt{Td\varepsilon}\}).$$

Note that if $T\varepsilon$ maximizes the left-hand side, then $d \leq T\varepsilon$ and hence $T \geq \Omega(d/\varepsilon)$. The remaining cases are $\Omega(d) \leq T^2\varepsilon^2/d$ and $\Omega(d) \leq \sqrt{Td\varepsilon}$. Both these statements give us $T \geq \Omega(d/\varepsilon)$. Hence the theorem follows, and it remains to prove Lemma 23:

Proof. Let $\mathcal{E}' = \{2^{-k/2}|\psi_x\rangle^{\otimes T} : x \in \{0,1\}^k\}$ and G be the $2^k \times 2^k$ Gram matrix for $\mathcal{E}'$. As we saw in Section 2.6, the success probability of identifying the states in the ensemble $\mathcal{E}$ using the PGM is

$$P^{PGM}(\mathcal{E}) = \sum_{x\in\{0,1\}^k} \sqrt{G}(x,x)^2.$$

For all $x, y \in \{0,1\}^k$, the entries of the Gram matrix G can be written as:

$$\begin{aligned} G(x,y) = \frac{1}{2^k}\langle\psi_x|\psi_y\rangle^T &= \frac{1}{2^k}\left((1-20\varepsilon) + \frac{20\varepsilon}{d}\sum_{i=1}^{d}\langle c^x(s_i)|c^y(s_i)\rangle\right)^T \\ &= \frac{1}{2^k}\left((1-20\varepsilon) + \frac{20\varepsilon}{d}(d - d_H(Mx, My))\right)^T \\ &= \frac{1}{2^k}\left(1 - \frac{20\varepsilon}{d}d_H(Mx, My)\right)^T, \end{aligned}$$

where Mx, $My \in \{0,1\}^d$ are codewords in the linear code defined earlier. Define $f : \{0,1\}^d \to \mathbb{R}$ as $f(z) = (1 - \frac{20\varepsilon}{d}|z|)^T$, and let $A(x,y) = (f \circ M)(x+y)$ for $x, y \in \{0,1\}^k$. Note that $G = A/2^k$. Since we assumed $T \leq d/(20e^3\varepsilon)$, we can use Theorem 17 (by choosing $m = d$ and $\beta = 20\varepsilon$) to upper bound

[6] For $x \in \{0,1\}^k$, define unitary $U_{c^x} : |s_i, b\rangle \to |s_i, b + c^x(s_i)\rangle$ for all $i \in \{0,\dots,d\}$. The ensemble $\mathcal{E}$ is generated by applying $\{U_{c^x}\}_{x\in\{0,1\}^k}$ to $|\varphi\rangle = \sum_{i\in\{0,\dots,d\}} \sqrt{D(s_i)}|s_i, 0\rangle$. View $c^x = (0, Mx) \in \{0,1\}^{d+1}$ as a concatenated string where Mx is a codeword of the $[d,k,r]_2$ code. Since the 2^k codewords of the $[d,k,r]_2$ code form a linear subspace, $\{U_{c^x}\}_{x\in\{0,1\}^k}$ is an Abelian group. From the discussion in Section 2.6, we conclude that the PGM is the optimal measurement for this state identification problem.

[7] We made no attempt to optimize the constants here.

the success probability of successfully identifying the states in the ensemble $\mathcal{E}$ using the PGM.

$$\begin{aligned}
P^{PGM}(\mathcal{E}) &= \sum_{x\in\{0,1\}^k} \sqrt{G}(x,x)^2 \\
&= \frac{1}{2^k} \sum_{x\in\{0,1\}^k} \sqrt{A}(x,x)^2 && \text{(since } G = A/2^k\text{)} \\
&\leq \frac{4e}{2^k}\left(1-\frac{\beta}{2}\right)^T e^{22T^2\beta^2/d+2\sqrt{Td}\beta} && \text{(using Theorem 17)} \\
&= \frac{4e}{2^k}\left(1-10\varepsilon\right)^T e^{8800T^2\varepsilon^2/d+4\sqrt{5Td}\varepsilon} && \text{(substituting } \beta = 20\varepsilon\text{)} \\
&\leq \frac{4e}{2^{k+T\varepsilon}} e^{8800T^2\varepsilon^2/d+4\sqrt{5Td}\varepsilon} && \text{(using } (1-10\varepsilon)^T \leq e^{-10\varepsilon T} \leq 2^{-\varepsilon T}\text{)}
\end{aligned}$$

The lemma follows by observing that $k \geq d/4$. □

□

4.2 Optimal lower bound for quantum agnostic learning

We now use the same approach to obtain a tight lower bound on quantum sample complexity in the *agnostic* setting.

Theorem 24. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d$, for sufficiently large d. Then for every $\delta \in (0,1/2)$ and $\varepsilon \in (0,1/10)$, every (ε,δ)-agnostic quantum learner for $\mathscr{C}$ has sample complexity $\Omega\left(\frac{d}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$.*

Proof. The d-independent part of the lower bound is Lemma 11. For the d-dependent term in the lower bound, consider a $[d,k,r]_2$ linear code (for $k \geq d/4$, distance $r \geq d/8$) as shown to exist in Theorem 3, with generator matrix $M \in \mathbb{F}_2^{d\times k}$ of rank k. Let $\{Mx : x \in \{0,1\}^k\} \subseteq \{0,1\}^d$ be the set of 2^k codewords in this linear code; these satisfy $d_H(Mx,My) \geq d/8$ whenever $x \neq y$. To each codeword $x \in \{0,1\}^k$ we associate a distribution D_x as follows:

$$D_x(s_i,b) = \frac{1}{d}\left(\frac{1}{2} + \frac{1}{2}(-1)^{(Mx)_i+b}\alpha\right), \qquad \text{for } (i,b) \in [d]\times\{0,1\},$$

where $\mathcal{S} = \{s_1,\ldots,s_d\}$ is a set that is shattered by $\mathscr{C}$, and α is a parameter which we shall pick later. Let $c^x \in \mathscr{C}$ be a concept that labels $\mathcal{S}$ according to $Mx \in \{0,1\}^d$. The existence of such $c^x \in \mathscr{C}$ follows from the fact that $\mathcal{S}$ is shattered by $\mathscr{C}$. Note that c^x is the minimal-error concept in $\mathscr{C}$ w.r.t. D_x. A learner that labels $\mathcal{S}$ according to some string $\ell \in \{0,1\}^d$ has additional error $d_H(Mx,\ell)\cdot\alpha/d$ compared to c^x. This in particular implies that an (ε,δ)-agnostic quantum learner has to find (with probability at least $1-\delta$) an ℓ such that $d_H(Mx,\ell) \leq d\varepsilon/\alpha$. We pick $\alpha = 20\varepsilon$ and we get $d_H(Mx,\ell) \leq d/20$. However, since Mx was a codeword of a $[d,k,r]_2$ code with distance $r \geq d/8$, finding an ℓ satisfying $d_H(Mx,\ell) \leq d/20$ is equivalent to *identifying* Mx, and hence x.

Now consider the following state identification problem: let $|\psi_x\rangle = \sum_{(i,b)\in[d]\times\{0,1\}} \sqrt{D_x(s_i,b)}|s_i,b\rangle$ for $x \in \{0,1\}^k$. Let the (ε,δ)-agnostic quantum sample complexity be T. Assume $T \leq d/(100e^3\varepsilon^2)$, since otherwise $T \geq \Omega(d/\varepsilon^2)$ and the theorem follows. Suppose the learner has knowledge of the ensemble $\mathcal{E} = \{(2^{-k}, |\psi_x\rangle^{\otimes T}) : x \in \{0,1\}^k\}$, and is given $|\psi_x\rangle^{\otimes T} \in \mathcal{E}$ for uniformly random x. The

learner would like to maximize the average probability of success to identify the given state. For this problem, we prove a lower bound on T using the PGM defined in Section 2.6. In particular, we show that using the PGM, if a learner successfully identifies the states in $\mathcal{E}$, then $T = \Omega(d/\varepsilon^2)$. Since the PGM is the optimal measurement[8] that the learner could have performed, the result follows. The following lemma makes this lower bound rigorous and will conclude the proof of the theorem.

Lemma 25. *For $x \in \{0,1\}^k$, let $|\psi_x\rangle = \sum_{(i,b)\in[d]\times\{0,1\}} \sqrt{D_x(s_i,b)}|s_i,b\rangle$, and $\mathcal{E} = \{(2^{-k}, |\psi_x\rangle^{\otimes T}) : x \in \{0,1\}^k\}$. Then*

$$P^{PGM}(\mathcal{E}) \leq \frac{4e}{e^{(d\ln 2)/4+25T\varepsilon^2}} e^{220000T^2\varepsilon^4/d+20\sqrt{Td\varepsilon^2}}.$$

Before we prove the lemma, we first show why it implies the theorem. Since we observed above that $P^{opt}(\mathcal{E}) = P^{PGM}(\mathcal{E})$, a good learner satisfies $P^{PGM}(\mathcal{E}) = \Omega(1)$ (say for $\delta = 1/4$), which in turn implies

$$\Omega(\max\{d, T\varepsilon^2\}) \leq O(\min\{T^2\varepsilon^4/d, \sqrt{Td\varepsilon^2}\}).$$

Like in the proof of Theorem 22, this implies a lower bound of $T = \Omega(d/\varepsilon^2)$ and proves the theorem. It remains to prove Lemma 25:

Proof. Let $\mathcal{E}' = \{2^{-k/2}|\psi_x\rangle^{\otimes T} : x \in \{0,1\}^k\}$ and G be the $2^k \times 2^k$ Gram matrix for the set $\mathcal{E}'$. As we saw in Section 2.6, the success probability of identifying the states in the ensemble $\mathcal{E}$ using the PGM is

$$P^{PGM}(\mathcal{E}) = \sum_{x\in\{0,1\}^k} \sqrt{G}(x,x)^2.$$

For all $x, y \in \{0,1\}^k$, the entries of G can be written as:

$$\begin{aligned}
2^k \cdot G(x,y) &= \langle\psi_x|\psi_y\rangle^T \\
&= \left(\sum_{(i,b)\in[d]\times\{0,1\}} \sqrt{D_x(i,b)D_y(i,b)}\right)^T \\
&= \left(\frac{1}{2d}\sum_{(i,b)\in[d]\times\{0,1\}} \sqrt{(1+10\varepsilon(-1)^{(Mx)_i+b})(1+10\varepsilon(-1)^{(My)_i+b})}\right)^T \\
&= \left(\frac{1}{2d}\sum_{\substack{(i,b):\\(Mx)_i=(My)_i}} (1+10\varepsilon(-1)^{(Mx)_i+b}) + \frac{1}{2d}\sum_{\substack{(i,b):\\(Mx)_i\neq(My)_i}} \sqrt{1-100\varepsilon^2}\right)^T \\
&= \left(\frac{d-d_H(Mx,My)}{d} + \frac{\sqrt{1-100\varepsilon^2}}{d} d_H(Mx,My)\right)^T \\
&= \left(1 - \frac{1-\sqrt{1-100\varepsilon^2}}{d} d_H(Mx,My)\right)^T.
\end{aligned}$$

where we used $\alpha = 20\varepsilon$ in the third equality.

Let $\beta = 1-\sqrt{1-100\varepsilon^2}$, which is at most 1 for $\varepsilon \leq 1/10$. Define $f : \{0,1\}^d \to \mathbb{R}$ as $f(z) = (1-\frac{\beta}{d}|z|)^T$, and let $A(x,y) = (f \circ M)(x+y)$ for $x, y \in \{0,1\}^k$. Then $G = A/2^k$. Note that $T \leq d/(100e^3\varepsilon^2) \leq$

[8] For $x \in \{0,1\}^k$, define unitary $U_{c^x} = \sum_{i\in[d]} |s_i\rangle\langle s_i| \otimes X^{(Mx)_i}$, where X is the NOT-gate, so $X^{(Mx)_i}|b\rangle = |b+(Mx)_i\rangle$ for $b \in \{0,1\}$. The ensemble $\mathcal{E}$ is generated by applying $\{U_{c^x}\}_{x\in\{0,1\}^k}$ to $|\varphi\rangle = \frac{1}{\sqrt{d}}\sum_{(i,b)\in[d]\times\{0,1\}} \sqrt{\frac{1}{2}+\frac{1}{2}(-1)^b\alpha}|s_i,b\rangle$. Since the 2^k codewords of the $[d,k,r]_2$ code form a linear subspace, $\{U_{c^x}\}_{x\in\{0,1\}^k}$ is an Abelian group. From the discussion in Section 2.6, we conclude that the PGM is the optimal measurement for this state identification problem.

$d/(e^3\beta)$ (the first inequality is by assumption and the second inequality follows for $\varepsilon \leq 1/10$ and $\beta \leq 1$). Since we assumed $T \leq d/(100e^3\varepsilon^2)$, we can use Theorem 17 (by choosing $m = d$ and $\beta = 1-\sqrt{1-100\varepsilon^2}$) to upper bound the success probability of identifying the states in the ensemble $\mathcal{E}$:

$$\begin{aligned}
P^{PGM}(\mathcal{E}) &= \sum_{x\in\{0,1\}^k} \sqrt{G}(x,x)^2 \\
&= \frac{1}{2^k} \sum_{x\in\{0,1\}^k} \sqrt{A}(x,x)^2 && \text{(since } G = A/2^k\text{)} \\
&\leq \frac{4e}{2^k}\left(1-\frac{\beta}{2}\right)^T e^{22T^2\beta^2/d+2\sqrt{Td\beta}} && \text{(using Theorem 17)} \\
&\leq \frac{4e}{2^k}\left(1-\frac{\beta}{2}\right)^T e^{220000T^2\varepsilon^4/d+20\sqrt{Td\varepsilon^2}} && \text{(using } \beta = 1-\sqrt{1-100\varepsilon^2} \leq 100\varepsilon^2\text{)} \\
&\leq \frac{4e}{2^k}\left(1-25\varepsilon^2\right)^T e^{220000T^2\varepsilon^4/d+20\sqrt{Td\varepsilon^2}} && \text{(using } \sqrt{1-100\varepsilon^2} \leq 1-50\varepsilon^2\text{)} \\
&\leq \frac{4e}{e^{k\ln 2+25T\varepsilon^2}} e^{220000T^2\varepsilon^4/d+20\sqrt{Td\varepsilon^2}}. && \text{(using } (1-x)^t \leq e^{-xt} \text{ for } x,t \geq 0\text{)}
\end{aligned}$$

The lemma follows by observing that $k \geq d/4$. □

□

4.3 Additional results

In this section we mention two additional results that can also be obtained using Theorem 17.

4.3.1 Quantum PAC sample complexity under random classification noise

In the theorem below, we show a lower bound on the quantum PAC sample complexity under the random classification noise model with noise rate η. Recall that in this model, for every $c \in \mathscr{C}$ and distribution D, $\varepsilon, \delta > 0$, given access to copies of the η-noisy state,

$$\sum_{x\in\{0,1\}^n} \sqrt{(1-\eta)D(x)}|x, c(x)\rangle + \sqrt{\eta D(x)}|x, 1-c(x)\rangle,$$

a (ε,δ)-PAC quantum learner is required to output an hypothesis h such that $\mathrm{err}_D(c,h) \leq \varepsilon$ with probability at least $1-\delta$.

Theorem 26. *Let $\mathscr{C}$ be a concept class with VC-dim$(\mathscr{C}) = d+1$, for sufficiently large d. Then for every $\delta \in (0,1/2)$, $\varepsilon \in (0,1/20)$ and $\eta \in (0,1/2)$, every (ε,δ)-PAC quantum learner for $\mathscr{C}$ in the PAC setting with random classification noise rate η, has sample complexity $\Omega\left(\frac{d}{(1-2\eta)^2\varepsilon} + \frac{\log(1/\delta)}{(1-2\eta)^2\varepsilon}\right)$.*

One can use exactly the same proof technique as in Lemma 10 and Theorem 22 to prove this, with only the additional inequality $1-2\sqrt{\eta(1-\eta)} \leq (1-2\eta)^2$, which holds for $\eta \leq 1/2$. We omit the details of the calculation.

4.3.2 Distinguishing codeword states

Ashley Montanaro (personal communication) alerted us to the following interesting special case of our PGM-based result.

Consider an $[n,k,d]_2$ linear code $\{Mx : x \in \{0,1\}^k\}$, where $M \in \mathbb{F}_2^{n\times k}$ is the rank-k generator matrix of the code, $k = \Omega(n)$, and distinct codewords have Hamming distance at least d.[9] For every $x \in \{0,1\}^k$, define a *codeword state* $|\psi_x\rangle = \frac{1}{\sqrt{n}}\sum_{i\in[n]}|i,(Mx)_i\rangle$. These states form an example of a *quantum fingerprinting* scheme [BCWW01]: 2^k states whose pairwise inner products are bounded away from 1. How many copies do we need to identify one such fingerprint?

Let $\mathcal{E} = \{(2^{-k},|\psi_x\rangle) : x \in \{0,1\}^k\}$ be an ensemble of codeword states. Consider the following task: given T copies of an unknown state drawn uniformly from $\mathcal{E}$, we are required to identify the state with probability $\geq 4/5$. From Holevo's theorem one can easily obtain a lower bound of $T = \Omega(k/\log n)$ copies, since the learner should obtain $\Omega(k)$ bits of information (i.e., identify k-bit string x with probability $\geq 4/5$), while each copy of the codeword state gives at most $\log n$ bits of information. In the theorem below, we improve that $\Omega(k/\log n)$ to the optimal $\Omega(k)$ for constant-rate codes.

Theorem 27. *Let $\mathcal{E} = \{|\psi_x\rangle = \frac{1}{\sqrt{n}}\sum_{i\in[n]}|i,(Mx)_i\rangle : x \in \{0,1\}^k\}$, where $M \in \mathbb{F}_2^{n\times k}$ is the generator matrix of an $[n,k,d]_2$ linear code with $k = \Omega(n)$. Then $\Omega(k)$ copies of an unknown state from $\mathcal{E}$ (drawn uniformly at random) are necessary to be able to identify that state with probability at least $4/5$.*

One can use exactly the proof technique of Theorem 22 to prove the theorem. Suppose we are given T copies of the unknown codeword state. Assume $T \leq n$, since otherwise $T \geq n \geq \sqrt{kn}$ and the theorem follows. Observe that the Gram matrix G for $\mathcal{E}' = \{2^{-k/2}|\psi_x\rangle^{\otimes T} : x \in \{0,1\}^k\}$ can be written as $G(x,y) = \frac{1}{2^k}\left(1 - \frac{|M(x+y)|}{n}\right)^T$ for $x,y \in \{0,1\}^k$. Using Theorem 17 (choosing $\beta = 1$ and $m = n$) to upper bound the success probability of successfully identifying the states in the ensemble $\mathcal{E}$ using the PGM, we obtain

$$P^{PGM}(\mathcal{E}) \leq \frac{4e}{2^{k+T}}e^{22T^2/n+2\sqrt{Tn}}.$$

As in the proof of Theorem 22, this implies the lower bound of Theorem 27. We omit the details of the calculation.

5 Conclusion

The main result of this paper is that quantum examples give no significant improvement over the usual random examples in passive, distribution-independent settings. Of course, these negative results do not mean that quantum machine learning is useless. In our introduction we already mentioned improvements from quantum examples for learning under the uniform distribution; improvements from using quantum membership queries; and improvements in time complexity based on quantum algorithms like Grover's and HHL. Quantum machine learning is still in its infancy, and we hope for many more positive results.

We end by identifying a number of open questions for future work:

- We gave lower bounds on sample complexity for the rather benign random classification noise. What about other noise models, such a *malicious* noise?

[9] Note that throughout this paper $\mathscr{C}$ was a concept class in $\{0,1\}^n$ and d was the VC dimension of $\mathscr{C}$. The use of n,d in this section has been changed to conform to the convention in coding theory.

- What is the quantum sample complexity for learning concepts whose range is $[k]$ rather than $\{0,1\}$, for some $k > 2$? Even the *classical* sample complexity is not fully determined yet [SB14, Section 29.2].
- Classically, it is still an open question whether the $\log(1/\varepsilon)$-factor in the upper bound of[BEHW89] for (ε,δ)-*proper* PAC learning is necessary. A weaker result (possibly easier to prove) would be to give a (ε,δ)-*quantum* proper PAC learner without this $\log(1/\varepsilon)$-factor.

- In the introduction we mentioned a few examples of learning under the *uniform* distribution where quantum examples are significantly more powerful than classical examples. Can we find more such examples of quantum improvements in sample complexity in fixed-distribution settings?
- Can we find more examples of quantum speed-up in *time* complexity of learning, for example for learning depth-3 or even constant-depth circuits?

Acknowledgments.

We thank Shalev Ben-David, Dmitry Gavinsky, Robin Kothari, Nishant Mehta, Ashley Montanaro, Henry Yuen for helpful comments and pointers to the literature. We also thank Ashley Montanaro for suggesting the additional remark in Section 4.3.2.

References

[Aar07] S. Aaronson. The learnability of quantum states. *Proceedings of the Royal Society of London*, 463(2088), 2007. quant-ph/0608142. 7

[Aar15] S. Aaronson. Quantum machine learning algorithms: Read the fine print. *Nature Physics*, 11(4):291–293, April 2015. 7

[AB09] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009. 5,9

[ABG06] E. Aïmeur, G. Brassard, and S. Gambs. Machine learning in a quantum world. In *Proceedings of Advances in Artificial Intelligence, 19th Conference of the Canadian Society for Computational Studies of Intelligence*, volume 4013, pages 431–442, 2006. 7

[ABG13] E. Aïmeur, G. Brassard, and S. Gambs. Quantum speed-up for unsupervised learning. *Machine Learning*, 90(2):261–287, 2013. 7

[AdW17] S. Arunachalam and R. de Wolf. A survey of quantum learning theory, 2017. To appear as Computational Complexity Column in SIGACT News, June 2017. Preprint at arxiv:1606.08920. 7

[AG98] B. Apolloni and C. Gentile. Sample size lower bounds in PAC learning by algorithmic complexity theory. *Theoretical Computer Science*, 209:141–162, 1998. 5

[AL88] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988. 6

[AM14] A. Ambainis and A. Montanaro. Quantum algorithms for search with wildcards and combinatorial group testing. *Quantum Information & Computation*, 14(5-6):439–453, 2014. arXiv:1210.1148. [8]

[AS05] A. Atıcı and R. Servedio. Improved bounds on quantum learning algorithms. *Quantum Information Processing*, 4(5):355–386, 2005. quant-ph/0411140. [1,4,7,14]

[AS09] A. Atıcı and R. Servedio. Quantum algorithms for learning and testing juntas. *Quantum Information Processing*, 6(5):323–348, 2009. arXiv:0707.3479. [4]

[Aud08] J. Audibert. Fast learning rates in statistical inference through aggregation, 2008. Research Report 06-20, CertisEcole des Ponts. math/0703854. [16]

[Aud09] J. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009. arXiv:0909.1468v1. [5,16]

[BCD06] D. Bacon, A. Childs, and W. van Dam. Optimal measurements for the dihedral hidden subgroup problem. *Chicago Journal of Theoretical Computer Science*, 2006. Earlier version in FOCS'05. quant-ph/0504083. [8]

[BCWW01] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf. Quantum fingerprinting. *Physical Review Letters*, 87(16), 2001. quant-ph/0102001. [28]

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. [2,12,29]

[BJ99] N. H. Bshouty and J. C. Jackson. Learning DNF over the uniform distribution using a quantum example oracle. *SIAM Journal on Computing*, 28(3):1136–1153, 1999. Earlier version in COLT'95. [1,3,4,10]

[BK02] H. Barnum and E. Knill. Reversing quantum dynamics with near-optimal quantum and classical fidelity. *Journal of Mathematical Physics*, 43:2097–2106, 2002. quant-ph/0004088. [11]

[BV97] E. Bernstein and U. Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, 1997. Earlier version in STOC'93. [4]

[DS16] A. Daniely and S. Shalev-Shwartz. Complexity theoretic limitations on learning DNF's. In *Proceedings of the 29th Conference on Learning Theory (COLT'16)*, 2016. [3]

[EF01] Y. C. Eldar and G. D. Forney Jr. On quantum detection and the square-root measurement. *IEEE Transactions and Information Theory*, 47(3):858–872, 2001. quant-ph/0005132. [12]

[EHKV89] A. Ehrenfeucht, D. Haussler, M. J. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989. Earlier version in COLT'98. [5]

[EMV03] Y. C. Eldar, A. Megretski, and G. C. Verghese. Designing optimal quantum detectors via semidefinite programming. *IEEE Transactions Information Theory*, 49(4):1007–1012, 2003. quant-ph/0205178. [12]

[Gav12] D. Gavinsky. Quantum predictive learning and communication complexity with single input. *Quantum Information and Computation*, 12(7-8):575–588, 2012. Earlier version in COLT'10. arXiv:0812.3429. 7

[GH01] C. Gentile and D. P. Helmbold. Improved lower bounds for learning from noisy examples: An information-theoretic approach. *Information and Computation*, 166:133–155, 2001. 5

[Gro96] L. K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of 28th ACM STOC*, pages 212–219, 1996. quant-ph/9605043. 7

[Han16] S. Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016. arXiv:1507.00473. 3,12

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. 3,9

[HHL09] A. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for solving linear systems of equations. 103(15):150502, 2009. arXiv:0811.3171. 7

[HJS+96] P. Hausladen, R. Jozsa, B. Schumacher, M. Westmoreland, and W. K. Wootters. Classical information capacity of a quantum channel. *Physical Review A*, 54:1869–1876, 1996. 6

[HMP+10] M. Hunziker, D. A. Meyer, J. Park, J. Pommersheim, and M. Rothstein. The geometry of quantum learning. *Quantum Information Processing*, 9(3):321–341, 2010. quant-ph/0309059. 7

[HW94] P. Hausladen and W. K. Wootters. A pretty good measurement for distinguishing quantum states. *Journal Of Modern Optics*, 41:2385–2390, 1994. 6

[Jac97] J. C. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, 1997. Earlier version in FOCS'94. 3,7

[JTY02] J. C. Jackson, C. Tamon, and T. Yamakami. Quantum DNF learnability revisited. In *Proceedings of 8th COCOON*, pages 595–604, 2002. quant-ph/0202066. 7

[JZ09] R. Jain and S. Zhang. New bounds on classical and quantum one-way communication complexity. *Theoretical Computer Science*, 410(26):2463–2477, 2009. arXiv:0802.4101. 15

[Kot14] R. Kothari. An optimal quantum algorithm for the oracle identification problem. In *31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014)*, pages 482–493, 2014. arXiv:1311.7685. 7

[KP16] A. Kontorovich and I. Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model, 2016. Preprint at arxiv:1606.08920. 5

[KSS94] M. J. Kearns, R. E. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994. Earlier version in COLT'92. 3,9

[KV94a] M. J. Kearns and L. G. Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM*, 41(1):67–95, 1994. [4]

[KV94b] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory*. MIT Press, 1994. [9]

[Mon07] A. Montanaro. On the distinguishability of random quantum states. *Communications in Mathematical Physics*, 273(3):619–636, 2007. quant-ph/0607011. [11]

[Mon12] A. Montanaro. The quantum query complexity of learning multilinear polynomials. *Information Processing Letters*, 112(11):438–442, 2012. arXiv:1105.3310. [7]

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. [8]

[SB14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014. [3,5,9,29]

[SG04] R. Servedio and S. Gortler. Equivalences and separations between quantum and classical learnability. *SIAM Journal on Computing*, 33(5):1067–1092, 2004. Combines earlier papers from ICALP'01 and CCC'01. quant-ph/0007036. [4,7]

[Sim96] H. U. Simon. General bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–254, 1996. Earlier version in COLT'93. [3,6,12]

[Sim15] H. U. Simon. An almost optimal PAC algorithm. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 1552–1563, 2015. [3]

[Tal94] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994. [3,12]

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. [2,9]

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. [2,9]

[VC74] V. Vapnik and A. Chervonenkis. Theory of pattern recognition. 1974. In Russian. [3,12]

[Ver90] K. A. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages 314–326, 1990. [3]

[WKS14] N. Wiebe, A. Kapoor, and K. M. Svore. Quantum deep learning, 2014. Preprint at arXiv:1412.3489. [7]

[WKS16] N. Wiebe, A. Kapoor, and K. M. Svore. Quantum perceptron models, 2016. Preprint at arXiv:1602.04799. [7]

[Zha10] C. Zhang. An improved lower bound on query complexity for quantum PAC learning. *Information Processing Letters*, 111(1):40–45, 2010. [1,4]