Quantum Coupon Collector

Srinivasan Arunachalam
IBM Research, Yorktown Heights, NY, USA
Srinivasan.Arunachalam@ibm.com

Aleksandrs Belovs
Faculty of Computing, University of Latvia, Riga, Latvia
aleksandrs.belovs@lu.lv

Andrew M. Childs
Department of Computer Science, Institute for Advanced Computer Studies,
University of Maryland, College Park, MD, USA
Joint Center for Quantum Information and Computer Science,
University of Maryland, College Park, MD, USA
amchilds@umd.edu

Robin Kothari
Microsoft Quantum, Redmond, WA, USA
Microsoft Research, Redmond, WA, USA
robin.kothari@microsoft.com

Ansis Rosmanis
Graduate School of Mathematics, Nagoya University, Japan
ansis.rosmanis@math.nagoya-u.ac.jp

Ronald de Wolf
QuSoft, Amsterdam, The Netherlands
CWI, Amsterdam, The Netherlands
University of Amsterdam, The Netherlands
rdewolf@cwi.nl

──── **Abstract** ────

We study how efficiently a k-element set $S \subseteq [n]$ can be learned from a uniform superposition $|S\rangle$ of its elements. One can think of $|S\rangle = \sum_{i \in S} |i\rangle / \sqrt{|S|}$ as the quantum version of a uniformly random sample over S, as in the classical analysis of the "coupon collector problem." We show that if k is close to n, then we can learn S using asymptotically fewer quantum samples than random samples. In particular, if there are $n - k = O(1)$ missing elements then $O(k)$ copies of $|S\rangle$ suffice, in contrast to the $\Theta(k \log k)$ random samples needed by a classical coupon collector. On the other hand, if $n - k = \Omega(k)$, then $\Omega(k \log k)$ quantum samples are necessary.

More generally, we give tight bounds on the number of quantum samples needed for every k and n, and we give efficient quantum learning algorithms. We also give tight bounds in the model where we can additionally reflect through $|S\rangle$. Finally, we relate coupon collection to a known example separating proper and improper PAC learning that turns out to show no separation in the quantum case.

1   Introduction

Learning from quantum states is a major topic in quantum machine learning. While this task has been studied extensively [15, 27, 7, 8, 6, 17, 4], many fundamental questions about the power of quantum learning remain. Determining properties of quantum states has potential applications not only in the context of machine learning, but also as a basic primitive for other types of quantum algorithms and for quantum information processing more generally.

In this paper we study a very simple and natural quantum learning problem. We are given copies of the uniform superposition

$$|S\rangle := \frac{1}{\sqrt{|S|}} \sum_{i \in S} |i\rangle$$

over the elements of an unknown set $S \subseteq [n] := \{1, \ldots, n\}$ (sometimes referred to as a uniform *quantum sample* from S [2]). Assume we know the size $k := |S| < n$. Our goal is to learn S exactly. How many copies of $|S\rangle$ do we need for this? And given the information-theoretically minimal number of copies needed, can we learn S gate-efficiently (i.e., using a quantum circuit with gate count polynomial in k and $\log n$)?

As a warm-up, first consider what happens if we just measure our copies of $|S\rangle$ in the computational basis, giving uniform samples from S. How many such samples do we need before we learn S? As long as there is some element of S that we have not seen, we cannot even guess S with constant success probability, so we need to sample until we see all k distinct elements. This is known as the "coupon collector problem." Analyzing the required number of samples is easy to do in expectation, as follows. Suppose we have already seen $i < k$ distinct elements from S. Then the probability that we see a new element in the next sample is $(k - i)/k$, and the expected number of samples to see an $(i + 1)$st element is the reciprocal of that probability, $k/(k - i)$. By linearity of expectation we can add this up over all i from 0 to $k - 1$, obtaining the expected number of samples to see all k elements:

$$\sum_{i=0}^{k-1} \frac{k}{k-i} = k \sum_{j=1}^{k} \frac{1}{j} \sim k \ln k.$$

With a bit more work one can show that $\Theta(k \log k)$ samples are necessary and sufficient to identify S with high probability [25, Chapter 3.6]:

▶ **Proposition 1** (Classical coupon collector). *Given uniformly random samples from a set $S \subseteq [n]$ of size $k < n$, the number of samples needed to identify S with high probability is $\Theta(k \log k)$.*

The relationship between the probability of seeing all elements of S and the number of samples is extremely well understood. In particular, we can achieve probability arbitrarily close to 1 using only $k \ln k + O(k)$ samples [25, Theorem 3.8].

Of course, measuring $|S\rangle$ in the computational basis is not the only approach a quantum computer could take. The goal of this paper is to identify when and how we can do better, reducing the number of copies of $|S\rangle$ that are used to solve this "quantum coupon collector problem." It turns out that we can asymptotically beat the classical threshold of $\Theta(k \log k)$ if and only if the number $m = n - k$ of "missing elements" is small (whereas classically the parameter m is irrelevant). Specifically, we give a simple, gate-efficient quantum algorithm that learns S from $O(n \log(m + 1))$ copies of $|S\rangle$. For small m this is significantly more efficient than classical coupon collection. In particular, for $m = O(1)$ we only need $O(k)$ quantum samples, saving a factor of $O(\log k)$.

As we explain in Section 5, this result is relevant for the comparison of *proper* and *improper* learning in the PAC model. A "proper" learner is one that only outputs hypotheses from the same concept class that its target function comes from. The coupon collector problem can be viewed as a learning task where the sample complexity of proper learners from classical random examples is asymptotically higher than that of proper learners from quantum examples.

We also prove *lower* bounds on the number T of copies needed, using the general (i.e., negative-weights) adversary bound of quantum query complexity [20]. This approach may be surprising, since no queries are involved when trying to learn S from copies of $|S\rangle$.[1] However, the adversary bound also characterizes the quantum query complexity of "state conversion" [21] and "state discrimination." Our learning problem may be viewed as the problem of converting the state $|S\rangle^{\otimes T}$ to a basis state that gives a classical description of the k-set S. To employ the general adversary bound, we exploit the underlying symmetries of the problem using the mathematical machinery of association schemes (see also [3, 23] for prior uses of association schemes in proving adversary lower bounds). Using this, we show that, unless the number of missing elements $m = n - k$ is very small, the $O(k \log k)$ classical coupon collector algorithm is essentially optimal even in the quantum case. This means that the quantum coupon collector might as well just measure the copies of the state in the computational basis, unless m is very small.

We also study the situation where, in addition to copies of $|S\rangle$, we can also apply a unitary operation $R_S = 2|S\rangle\langle S| - \mathsf{Id}$ that reflects through the state $|S\rangle$ (i.e., $R_S |S\rangle = |S\rangle$ and $R_S |\phi\rangle = -|\phi\rangle$ for all states $|\phi\rangle$ orthogonal to $|S\rangle$). This model is reasonable to consider because if we had a unitary that prepared $|S\rangle$, or even $|S\rangle |\psi\rangle$ for some garbage state $|\psi\rangle$, starting from some canonical state $|0\rangle$, then we could use this unitary to create the unitary R_S in a black-box manner. For example, if $U |0\rangle = |S\rangle$, then $R_S = U(2|0\rangle\langle 0| - \mathsf{Id})U^\dagger$.

This model gives us extra power and enables more efficient learning of the set S: $\Theta(\sqrt{km})$ states and reflections are necessary and sufficient to learn S for large k (i.e., small m), and $\Theta(k)$ states and reflections are necessary and sufficient for small k.

The following table summarizes our main results. Sections 2 and 3 prove the upper and lower bounds in the first row, respectively, while Section 4 proves the results in the second row.

---

[1] A natural approach is to analyze the success probability of the "pretty good measurement" (PGM) for discriminating the states $|S\rangle^{\otimes T}$. The PGM is an explicit measurement whose average success probability is no more than quadratically worse than that of the optimal measurement [10] (in fact, one can show that the PGM is optimal in our case because our set of states is "geometrically uniform"). One can write down the average success probability of the PGM explicitly, and upper bounding it would establish a lower bound on the required number of copies of $|S\rangle$. However, we have been unable to suitably bound this expression despite considerable effort.

■ **Table 1** Main results about the complexity of learning the set S with $m = n - k$ missing elements.

	$k \geq n/2$	$k \leq n/2$
Number of copies of $|S\rangle$:	$\Theta(k \log(m + 1))$ Theorem 2 and Theorem 6	$\Theta(k \log k)$ Proposition 1 and Theorem 6
Number of copies and reflections:	$\Theta(\sqrt{km})$ Theorem 10 and Theorem 11	$\Theta(k)$ Theorem 12 and Theorem 13

We contrast our work with recent results on the quantum query complexity of approximate counting by Aaronson, Kothari, Kretschmer, and Thaler [1]. They consider a similar model, given copies of the state $|S\rangle$, the ability to reflect through $|S\rangle$, and also the ability to query membership in S. However, in their work the size of S is unknown and the goal is to approximately *count* this set up to small multiplicative error. They obtain tight upper and lower bounds on the complexity of this approximate-counting task using techniques quite different from ours (specifically, Laurent polynomials for the lower bounds). This allows them to give an oracle separation between the complexity classes SBP and QMA. In contrast, in our case the size k of the set S is already known to the learner from the start, and the goal is to *identify S* exactly.

2   Upper bound on quantum samples

In this section we prove upper bounds on the number of copies of $|S\rangle$ that suffice to identify the k-element set $S \subseteq [n]$ with high probability.

The easiest way to recover S is by measuring $O(k \log k)$ copies of $|S\rangle$ in the computational basis. By the classical coupon collector problem (Proposition 1), we will (with high probability) see all elements of S at least once. As we will show later, this number of copies of $|S\rangle$ turns out to be asymptotically optimal if the number of missing elements $m = n - k$ is large (at least polynomial in n). However, here we show that something better is possible for very small m.

▶ **Theorem 2** (Upper bound for small m). *Let $S \subseteq [n]$ be a set of size $k < n$ and let $m = n - k$. We can identify S with high probability using $O(k \log(m + 1))$ copies of $|S\rangle$ by a gate-efficient quantum algorithm.*

Proof. This bound is trivial when m is polynomial in n, since an upper bound of $O(k \log k)$ follows from Proposition 1. So let us now assume that $m \leq n^{1/4}$ and hence $k \geq n - n^{1/4}$.

Consider the uniform superposition over all elements of the universe $[n]$:

$$|[n]\rangle = \frac{1}{\sqrt{n}} \sum_{i \in [n]} |i\rangle .$$

Performing the 2-outcome projective measurement with operators $|[n]\rangle \langle [n]|$ and $\mathsf{Id} - |[n]\rangle\langle[n]|$ is no harder than preparing $|[n]\rangle$, so it can be implemented gate-efficiently. If we apply this measurement to a copy of $|S\rangle$, then we get the first outcome with probability $|\langle S|[n]\rangle|^2 = k/n$ and the second outcome with probability m/n. In the latter case, the post-measurement state is

$$|\psi\rangle = \sqrt{\frac{m}{n}} |S\rangle - \sqrt{\frac{k}{n}} |\overline{S}\rangle$$

which is close to $-|\overline{S}\rangle$ if $m \ll n$.

We use an expected number of $O\left(\frac{n}{m} \cdot m \log(m+1)\right) = O(n \log(m+1))$ copies of $|S\rangle$ to prepare $O(m \log(m+1))$ copies of $|\psi\rangle$. If $|\psi\rangle$ were exactly equal to $-|\overline{S}\rangle$, then measuring in the computational basis would sample uniformly over the set $\overline{S}$ of m missing elements, and $O(m \log(m+1))$ such samples suffice to recover $\overline{S}$ by the classical coupon collector problem (Proposition 1). Instead, $|\psi\rangle$ only approximately equals $-|\overline{S}\rangle$: if we measure it then each $i \in \overline{S}$ has probability $\frac{k}{nm}$, while each $i \in S$ has (much smaller but nonzero) probability $\frac{m}{nk}$. Suppose we prepare and measure $T = 10m \log(m+1)$ copies of $|\psi\rangle$. Then the expected number of occurrences of each $i \in \overline{S}$ is $T \cdot \frac{k}{nm} \geq 5 \log(m+1)$ since $k \geq n/2$, while the expected number of occurrences of each $i \in S$ is $T \cdot \frac{m}{nk} = O(\log(n)/n^{3/2})$. In both cases the number of occurrences is tightly concentrated.[2] Hence if we keep only the elements that appear, say, at least $\log(m+1)$ times among the T outcomes, then with high probability we will have found $\overline{S}$, and hence learned $S = [n] \setminus \overline{S}$. ◀

3   Lower bound on quantum samples

In this section we prove lower bounds on the number of copies of $|S\rangle$ needed to identify S with high probability. Before establishing the lower bounds claimed in Table 1, we introduce some preliminary concepts, namely the γ_2-norm (Section 3.1), association schemes (Section 3.2), the Johnson scheme (Section 3.3), and the adversary bound for state discrimination (Section 3.4). The lower bound itself is established in Section 3.5.

3.1 γ_2-norm

The γ_2-norm of a $D_1 \times D_2$ matrix A with entries $A(x,y)$ for $x \in [D_1]$ and $y \in [D_2]$ can be defined in two equivalent ways [12, Section 3]. The primal definition is

$$
\begin{aligned}
&\text{minimise} &&\max\left\{\max_{x \in [D_1]} \|u_x\|^2, \max_{y \in [D_2]} \|v_y\|^2\right\} \\
&\text{subject to} &&A(x,y) = \langle u_x, v_y \rangle \qquad \text{for all } x \in [D_1] \text{ and } y \in [D_2],
\end{aligned}
\tag{1}
$$

where $\{u_x : x \in [D_1]\}$ and $\{v_y : y \in [D_2]\}$ are vectors of the same dimension. The dual definition is

$$
\begin{aligned}
&\text{maximise} &&\|\Gamma \circ A\| \\
&\text{subject to} &&\|\Gamma\| \leq 1
\end{aligned}
\tag{2}
$$

where Γ ranges over $D_1 \times D_2$-matrices, $\circ$ denotes the Hadamard (entrywise) product of matrices, and $\|\cdot\|$ is the spectral norm of a matrix. Note that $\gamma_2(A \circ B) \leq \gamma_2(A)\gamma_2(B)$: consider the vectors obtained from the optimal feasible solutions of $\gamma_2(A)$ and $\gamma_2(B)$ in Eq. (1) and observe that the tensor product of these vectors forms a feasible solution for the primal problem for $\gamma_2(A \circ B)$ with (not necessarily minimal) value $\gamma_2(A)\gamma_2(B)$.

---

[2] Suppose we flip T 0/1-valued coins, each taking value 1 with probability p. Let X be their sum (i.e., the number of 1s), which has expectation $\mu = pT$. The Chernoff bound implies $\Pr[X \leq (1-\delta)\mu] \leq \exp(-\delta^2\mu/2)$. To get concentration for the number of occurrences of a specific $i \in \overline{S}$, apply this tail bound with $p = k/(nm)$, $\mu = Tp \geq 5 \log(m+1)$, $\delta = 4/5$ to obtain $\Pr[X \leq \log(m+1)] \ll 1/m$. Hence, by a union bound, the probability that among the m elements $i \in \overline{S}$ there is one of which we see fewer than $\log(m+1)$ occurrences, is $\ll 1$. For an $i \in S$, by Markov's inequality the probability to see at least $\log(m+1)$ occurrences of this i among the T samples is $\ll 1/n$, and we can use a union bound over all $i \in S$.

3.2 Association schemes

Here we present a quick introduction to association schemes (see, for example, [16, Chapter 1] for a more thorough treatment).

▶ **Definition 3.** *An* association scheme *on the set U is a finite set of real symmetric $U \times U$ matrices $\{A_0, A_1, \ldots, A_s\}$ satisfying all the following properties:*

— *each A_j only has entries 0 and 1;*

— *A_0 is the identity matrix* Id*;*

— *$\sum_{j=0}^{s} A_j$ is the all-1 matrix J; and*

— *for every i and j, the product $A_i A_j$ is a linear combination of the matrices $\{A_0, \ldots, A_s\}$.*
The space spanned by the set $\{A_0, A_1, \ldots, A_s\}$ forms an algebra, which is called the Bose–Mesner algebra *corresponding to the scheme. By abuse of terminology, we may also refer to this algebra as the association scheme.*

We now state a few properties of $\{A_0, \ldots, A_s\}$. First, observe that A_j has zero diagonal for $j > 0$. Additionally, $\{A_0, \ldots, A_s\}$ form a basis of the corresponding Bose–Mesner algebra, since for every (x, y), there is exactly one j for which $A_j(x, y) \neq 0$. Also, the basis $\{A_0, \ldots, A_s\}$ satisfies $A_i \circ A_j = \mathbf{1}_{[i=j]} A_i$, where $\mathbf{1}_{[P]}$ is the indicator function of predicate P (i.e., 1 if P is true and 0 if P is false). It is possible to find another basis $\{E_0, \ldots, E_s\}$ consisting of *idempotent* matrices for span$\{A_0, \ldots, A_s\}$ that satisfy $E_i E_j = \mathbf{1}_{[i=j]} E_i$, with respect to the usual product of matrices. The operators E_i are orthogonal projectors onto the *eigenspaces* of the association scheme. We have

$$E_0 = J/N \qquad \text{and} \qquad \sum_{j=0}^{s} E_j = \text{Id},$$

where $N = |U|$. Since both $\{A_i\}$ and $\{E_j\}$ are bases for the space of $N \times N$ matrices, it is possible to write

$$A_i = \sum_{j=0}^{s} p_i(j) E_j \quad \text{and} \quad E_j = \sum_{i=0}^{s} \frac{q_j(i)}{N} A_i, \tag{3}$$

where $p_i(j)$ and $q_j(i)$ are called the *eigenvalues* and *dual eigenvalues* of the association scheme, respectively.

It is easy to show that the Hadamard product and the usual product of any two elements of the association scheme also belong to the association scheme. Clearly for every i, j we know that $A_i \circ A_j$ and $E_i \cdot E_j$ are elements of the basis of the scheme. Also observe that $A_i \cdot A_j$ and $E_i \circ E_j$ are elements of the scheme by writing out these products using Eq. (3) and observing that $A_i \cdot A_j$ (resp. $E_i \circ E_j$) is a linear combination of elements of $\{A_0, \ldots, A_s\}$ (resp. $\{E_0, \ldots, E_s\}$). In particular, we can write

$$E_i \circ E_j = \frac{1}{N} \sum_{\ell=0}^{s} q_{i,j}(\ell) E_\ell. \tag{4}$$

The real numbers $q_{i,j}(\ell)$ are called the *Krein parameters* of the association scheme.

3.3 Johnson scheme

In the Johnson association scheme $\mathcal{J}(n, k)$, the set U is the set of all k-subsets of $[n]$. Therefore, $N = |U| = \binom{n}{k}$. Let $m = \min\{k, n - k\}$. For $j = 0, 1, \ldots, m$, define $A_j(x, y) := \mathbf{1}_{[|x \cap y| = k - j]}$.

The idempotent E_j is defined as follows: for $x \in U$, let $e_x \in \mathbb{R}^U$ be the indicator vector defined as $e_x(y) = \mathbf{1}_{[x=y]}$ for $y \in U$, and let

$$\mathcal{V}_j := \begin{cases} \text{span}\{\sum_{x \supseteq z} e_x \colon z \subseteq [n] \text{ with } |z| = j\} & \text{if } k \leq n/2, \\ \text{span}\{\sum_{x \subseteq z} e_x \colon z \subseteq [n] \text{ with } |z| = n - j\} & \text{if } k > n/2, \end{cases}$$

where the sums are over $x \in U$. These spaces satisfy $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_m = \mathbb{R}^U$ and the dimension of $\mathcal{V}_j$ is $\binom{n}{j}$. For $j \in \{1, \ldots, m\}$, the idempotent E_j is defined as the orthogonal projector on $\mathcal{V}_j \cap \mathcal{V}_{j-1}^\perp$, and E_0 is the orthogonal projector on $\mathcal{V}_0$. Hence, for $j \in \{0, 1, \ldots, m\}$, the dimension of the jth eigenspace is

$$d_j := \text{Tr}[E_j] = \binom{n}{j} - \binom{n}{j-1}. \tag{5}$$

We do not require explicit expressions for most eigenvalues and valencies of $\mathcal{J}(n,k)$, the only exceptions being the dual eigenvalues

$$q_0(i) = 1 \quad \text{and} \quad q_1(i) = \frac{n(n-1)}{n-k}\left(\frac{k-i}{k} - \frac{k}{n}\right). \tag{6}$$

See [26, Eq. 1.24] for the latter. We are only interested in the following Krein parameters of this association scheme. When one idempotent is E_0, we have

$$q_{i,0}(j) = \mathbf{1}_{[i=j]}. \tag{7}$$

When one idempotent is E_1, we have

$$q_{j-1,1}(j) = \frac{j(n-1)n(k-j+1)(m-j+1)}{mk(n-2j+1)(n-2j+2)}, \tag{8a}$$

$$q_{j,1}(j) = \frac{j(n-1)(n-j+1)(m-k)^2}{mk(n-2j)(n-2j+2)}, \tag{8b}$$

$$q_{j+1,1}(j) = \frac{n(n-1)(n-j+1)(k-j)(m-j)}{mk(n-2j)(n-2j+1)}, \tag{8c}$$

and $q_{i,1}(j) = 0$ whenever $|i - j| > 1$ (see [9, Section 3.2]).

3.4   Adversary lower bound for state discrimination

Consider the following state-discrimination problem.

(∗) Let $f \colon D \to R$ be a function for some finite sets D and R. Let $\{|\psi_x\rangle : x \in D\}$ be a family of quantum states of the same dimension. Given a copy of $|\psi_x\rangle$ for an arbitrary $x \in D$, the goal is to determine $f(x)$ with high success probability.

Let A be the Gram matrix of the states, namely

$$A(x,y) = \langle \psi_x \mid \psi_y \rangle,$$

and let F be the $D \times D$ matrix with

$$F(x,y) = \mathbf{1}_{[f(x) \neq f(y)]}.$$

Informally, the main result of this section is that the above state-discrimination problem can be solved with small error probability if and only if

$$\gamma_2(A \circ F)$$

is small. We start with the proof of the lower bound. With constants refined, it reads as follows:

▶ **Proposition 4.** *If the above state-discrimination problem (∗) can be solved with success probability $1 - \varepsilon$, then $\gamma_2(A \circ F) \leq 4\sqrt{\varepsilon}$.*

Proof. This is essentially the result of [12, Claim 3.27], which is also closely related to [20]. For completeness we repeat the proof, with slight modifications.

Without loss of generality we may assume the measurement is projective (this follows from Neumark's theorem). Thus, there exist orthogonal projectors $\{\Pi_a\}_{a \in R}$ such that $\left\|\Pi_{f(x)} \left|\psi_x\right\rangle\right\|^2 \geq 1 - \varepsilon$ for all $x \in D$. Denote $\Pi_a^\perp = \mathsf{Id} - \Pi_a$, so that $\|\Pi_{f(x)}^\perp \left|\psi_x\right\rangle\|^2 \leq \varepsilon$ for all $x \in D$. We first write

$$A(x,y) = \langle \psi_x \mid \psi_y \rangle = \langle \psi_x | \Pi_{f(y)} |\psi_y\rangle + \langle \psi_x | \Pi_{f(y)}^\perp |\psi_y\rangle$$
$$= \langle \psi_x | \Pi_{f(x)} \Pi_{f(y)} |\psi_y\rangle + \langle \psi_x | \Pi_{f(x)}^\perp \Pi_{f(y)} |\psi_y\rangle + \langle \psi_x | \Pi_{f(y)}^\perp |\psi_y\rangle .$$

Note that if $f(x) \neq f(y)$, then the first term is 0 because $\Pi_{f(x)}$ and $\Pi_{f(y)}$ project onto orthogonal subspaces. This motivates us to define the $D \times D$ matrix

$$B(x,y) = \langle \psi_x | \Pi_{f(x)}^\perp \Pi_{f(y)} |\psi_y\rangle + \langle \psi_x | \Pi_{f(y)}^\perp |\psi_y\rangle .$$

We have $A(x,y) = B(x,y)$ whenever $f(x) \neq f(y)$, and hence $A \circ F = B \circ F$. Note that $\gamma_2(B) \leq 2\sqrt{\varepsilon}$ by taking the vectors $u_x = \left(\varepsilon^{-1/4} \Pi_{f(x)}^\perp |\psi_x\rangle \ , \ \varepsilon^{1/4} |\psi_x\rangle\right)$ and $v_y = \left(\varepsilon^{1/4} \Pi_{f(y)} |\psi_y\rangle \ , \ \varepsilon^{-1/4} \Pi_{f(y)}^\perp |\psi_y\rangle\right)$. Now we have

$$\gamma_2(A \circ F) = \gamma_2(B \circ F) \leq \gamma_2(B)\gamma_2(F) \leq 4\sqrt{\varepsilon},$$

where we used the composition property of the γ_2-norm in the first inequality and in the second inequality we used $\gamma_2(F) \leq 2$, which follows by considering the vectors $u_x, v_y \in \{0,1\}^{|R|+1}$ whose last coordinate is always 1, and where u_x has a 1 at coordinate $f(x)$ and v_y has a -1 at coordinate $f(y)$ (identifying R with $\{1, \ldots, |R|\}$ for the purposes of indexing these vectors), and whose remaining entries are all 0.   ◀

▶ **Proposition 5.** *The above state-discrimination problem (∗) can be solved with success probability at least $1 - \gamma_2(A \circ F)$.*

Proof. If B is the Gram matrix of the collection of states $\{|\psi_x\rangle \otimes |f(x)\rangle\}_{x \in D}$, then

$$A - B = A \circ F.$$

Using [22, Claim 3.10], there exists a unitary U such that

$$\left(\langle \psi_x | \otimes \langle f(x)|\right) U \left(|\psi_x\rangle \otimes |0\rangle\right) \geq 1 - \varepsilon/2$$

where $\varepsilon := \gamma_2(A \circ F)$. Thus, if we measure the second register of $U(|\psi_x\rangle \otimes |0\rangle)$, we get $f(x)$ with probability at least $(1 - \varepsilon/2)^2 \geq 1 - \varepsilon$.   ◀

3.5   Lower bound

For $x \subseteq [n]$ of size k, let

$$|\psi_x\rangle = \frac{1}{\sqrt{k}} \sum_{i \in x} |i\rangle .$$

This is what we denoted by $|S\rangle$ earlier ($x = S$); we use $|\psi_x\rangle$ here for consistency with the common notation in lower bounds. The task is to identify the subset x using as few copies of the state $|\psi_x\rangle$ as possible. We prove the following lower bound.

▶ **Theorem 6.** *To find x with success probability $\Omega(1)$, it is necessary to have $\Omega(k \log(\min\{k, n - k\}))$ copies of the state $|\psi_x\rangle$.*

Let $m = n - k$. Since we could add more elements to the ambient space artificially, the problem becomes no easier as n grows with k fixed. Thus, it suffices to prove the lower bound of $\Omega(k \log(m + 1))$ under the assumption $m \ll k$.

Define the Gram matrix Ψ by $\Psi(x, y) = \langle \psi_x \mid \psi_y \rangle$. The Gram matrix corresponding to $|\psi_x\rangle^{\otimes \ell}$ is $\Psi^{\circ \ell}$ (where $\Psi^{\circ \ell}$ is the Hadamard product of Ψ with itself ℓ times). The function we want to compute is $f : x \mapsto x$, so we have $F(x, y) = \mathbf{1}_{[f(x) \neq f(y)]} = \mathbf{1}_{[x \neq y]}$, i.e., $F = J - \mathsf{Id}$. By Proposition 4, it thus suffices to prove that for some $\ell = \Omega(k \log(m + 1))$ we have

$$\gamma_2\big(\Psi^{\circ \ell} \circ (J - \mathsf{Id})\big) = \Omega(1).$$

To that end, we use the dual formulation of the γ_2-norm (in Eq. (2)) and construct a matrix Γ such that

$$\|\Gamma\| = 1, \qquad \Gamma \circ \mathsf{Id} = 0, \qquad \text{and} \qquad \|\Gamma \circ \Psi^{\circ \ell}\| = \Omega(1).$$

We now construct a Γ that satisfies the constraints above. To do so, we first write Γ in terms of the idempotents $\{E_j\}_{j=0}^m$ of the Johnson association scheme (as defined above Eq. (5)): for $\{\gamma_j\}_j$ which we define shortly, let

$$\Gamma = \sum_{j=0}^m \gamma_j E_j. \tag{9}$$

To satisfy $\Gamma \circ \mathsf{Id} = 0$, we would like Γ to have zero diagonal. Note that Γ has zero diagonal if and only if $\mathrm{Tr}[\Gamma] = \sum_{j=0}^m \gamma_j \mathrm{Tr}[E_j] = \sum_{j=0}^m \gamma_j d_j = 0$, where d_j was defined in Eq. (5). We now fix $\{\gamma_j\}_j$ as follows: since $d_m = \binom{n}{m} - \binom{n}{m-1}$ is larger than the sum of the remaining d_js, we let

$$\gamma_0 = \gamma_1 = \cdots = \gamma_{m-1} = 1, \qquad \gamma_m \in [-1, 0] \tag{10}$$

so that $\mathrm{Tr}[\Gamma] = 0$ and $\|\Gamma\| = 1$. Thus, it remains to show that

$$\|\Gamma \circ \Psi^{\circ \ell}\| = \Omega(1). \tag{11}$$

For that, we use the following technical result.

▶ **Lemma 7.** *For each $j = 0, 1, \ldots, m$, we have*

$$E_j \circ \Psi = p_{j+1,-1} E_{j+1} + p_{j,0} E_j + p_{j-1,+1} E_{j-1},$$

where

$$p_{j,-1} = \frac{j(k - j + 1)(m - j + 1)}{(n - 2j + 1)(n - 2j + 2)k},$$

$$p_{j,0} = \frac{k}{n} + \frac{j(n - j + 1)(m - k)^2}{nk(n - 2j)(n - 2j + 2)},$$

$$p_{j,+1} = \frac{(n - j + 1)(k - j)(m - j)}{(n - 2j)(n - 2j + 1)k}.$$

Before we proceed with the proof of this lemma, let us state a simple consequence.

▶ **Corollary 8.** *For each $j \in \{0, \ldots, m\}$, the numbers $p_{j,-1}$, $p_{j,0}$, and $p_{j,+1}$ are non-negative, and satisfy $p_{j,-1} + p_{j,0} + p_{j,+1} = 1$.*

Proof. The non-negativity is obvious. For the last property note that

$$\sum_{j=0}^{m} E_j = \mathsf{Id} = \Psi \circ \mathsf{Id} = \Psi \circ \left(\sum_{j=0}^{m} E_j \right) = \sum_{j=0}^{m} \left(p_{j,-1} + p_{j,0} + p_{j,1} \right) E_j,$$

where the first equality uses the definition of an association scheme, the second equality follows because $\Psi(x, x) = 1$ by definition, and the last equality is by the assumption of Lemma 7.   ◀

Proof of Lemma 7. It suffices to write out Ψ in the basis $\{E_j\}_{j=0}^{m}$ and use the Krein parameters. By definition of $|\psi_x\rangle = \frac{1}{\sqrt{k}} \sum_{i \in x} |i\rangle$, we have that $\Psi(x, y) = \langle \psi_x \mid \psi_y \rangle$ equals $\frac{1}{k}$ times the intersection of x and y, and

$$\Psi = \sum_{i=0}^{m} \left(1 - \frac{i}{k} \right) A_i,$$

where A_i was defined at the beginning of Section 3.3 as $A_i(x, y) := \mathbf{1}_{[|x \cap y| = k - i]}$. We now rewrite Ψ as follows: using Eq. (6), we have

$$\frac{k}{n} E_0 + \frac{n-k}{n(n-1)} E_1 = \frac{1}{N} \sum_{i=0}^{m} \left(\frac{k}{n} q_0(i) + \frac{n-k}{n(n-1)} q_1(i) \right) A_i = \frac{1}{N} \sum_{i=0}^{m} \frac{k-i}{k} A_i = \frac{1}{N} \Psi,$$

where the first equality used Eq. (3). Additionally observe that

$$N E_j \circ E_0 = q_{j,0}(j) E_j \qquad \text{and} \qquad N E_j \circ E_1 = q_{j,1}(j-1) E_{j-1} + q_{j,1}(j) E_j + q_{j,1}(j+1) E_{j+1}.$$

Plugging in the values of $q_{j,\cdot}$ from Eq. (8), we get the required equality.   ◀

We are now ready to prove our main lower bound in Theorem 6.

Proof of Theorem 6. We prove this by induction on the number of copies of the state $|\psi_x\rangle$, which we denote by s. Let us define $\gamma_j^{(s)}$ via

$$\Gamma \circ \Psi^{\circ s} = \sum_{j=0}^{m} \gamma_j^{(s)} E_j.$$

Since the E_j are pairwise-orthogonal projections, the norm of $\Gamma \circ \Psi^{\circ s}$ equals $\max_j |\gamma_j^{(s)}|$. Hence to lower bound $\| \Gamma \circ \Psi^{\circ s} \|$, it suffices to lower bound $\gamma_0^{(s)}$.

We have

$$\Gamma \circ \Psi^{\circ (s+1)} = \sum_{j=0}^{m} \gamma_j^{(s)} E_j \circ \Psi$$

and using Lemma 7 we get

$$\gamma_j^{(s+1)} = p_{j,-1} \gamma_{j-1}^{(s)} + p_{j,0} \gamma_j^{(s)} + p_{j,+1} \gamma_{j+1}^{(s)}. \tag{12}$$

For every $j \in \{0, \ldots, m\}$, we now consider the following probabilistic sequence $\{B_j^{(s)}\}$. For $s = 0$, we let $B_j^{(0)} = \gamma_j$ and

$$B_j^{(s+1)} = \begin{cases} B_{j-1}^{(s)} & \text{with probability } p_{j,-1}, \\ B_j^{(s)} & \text{with probability } p_{j,0}, \\ B_{j+1}^{(s)} & \text{with probability } p_{j,+1}, \end{cases}$$

using the fact that $p_{j,-1} + p_{j,0} + p_{j,+1} = 1$. Note that $B_j^{(s)}$ only takes values from $\{\gamma_0, \ldots, \gamma_m\}$ and there are only two distinct such values, namely 1 and γ_m (since $\gamma_0 = \gamma_1 = \cdots = \gamma_{m-1} = 1$ as defined in Eq. (10)). Also note that $p_{0,-1} = p_{m,+1} = 0$, so we do not have to explicitly handle the boundaries. Induction on s using Eq. (12) shows that $\mathbb{E}[B_j^{(s)}] = \gamma_j^{(s)}$, which is the motivation behind defining these variables.

Define similarly $C_j^{(s)}$ as $C_j^{(0)} = \gamma_j$ and

$$C_j^{(s+1)} = \begin{cases} C_j^{(s)} & \text{with probability } p_{j,-1} + p_{j,0}, \\ C_{j+1}^{(s)} & \text{with probability } p_{j,+1}. \end{cases}$$

Let us give an intuitive description of how the random variables $C_j^{(s)}$ behave. For each s, the head of the sequence $C_0^{(s)}, C_1^{(s)}, \ldots$, up to some $C_\ell^{(s)}$ consists purely of 1s, and the tail $C_{\ell+1}^{(s)}, \ldots, C_m^{(s)}$ consists purely of γ_m. Initially, for $s = 0$, the tail consists of one element $C_m^{(s)}$ only, but the tail gradually extends as s grows (and the head, respectively, shrinks). The probability of growing the length of the tail from $m - j$ to $m - j + 1$ in one step is $p_{j,+1}$.

The random variables $B_j^{(s)}$ behave similarly, but are slightly more complicated, since the tail can also shrink and 1s can get into the tail. This is the reason why we replace $B_j^{(s)}$ with $C_j^{(s)}$ in our analysis: $C_j^{(s)}$ is easier to analyze, and it suffices to lower bound its expectation because $B_j^{(s)}$ dominates $C_j^{(s)}$, i.e., for each s and j and real t we have $\Pr[B_j^{(s)} \geq t] \geq \Pr[C_j^{(s)} \geq t]$. The latter is proven by induction, as follows. The base case $s = 0$ is trivial, and the inductive step is

$$\begin{aligned} \Pr[B_j^{(s+1)} \geq t] &= p_{j,-1} \Pr[B_{j-1}^{(s)} \geq t] + p_{j,0} \Pr[B_j^{(s)} \geq t] + p_{j,+1} \Pr[B_{j+1}^{(s)} \geq t] \\ &\geq p_{j,-1} \Pr[C_{j-1}^{(s)} \geq t] + p_{j,0} \Pr[C_j^{(s)} \geq t] + p_{j,+1} \Pr[C_{j+1}^{(s)} \geq t] \\ &\geq (p_{j,-1} + p_{j,0}) \Pr[C_j^{(s)} \geq t] + p_{j,+1} \Pr[C_{j+1}^{(s)} \geq t] = \Pr[C_j^{(s+1)} \geq t], \end{aligned}$$

since $C_{j-1}^{(s)} \geq C_j^{(s)}$ by our above analysis.

The analysis of $C_j^{(s)}$ is very similar to the classical coupon collector problem if we interpret the length of the tail as the number of acquired coupons. We briefly repeat the argument. For each j, define random variable T_j as the first value of s such that $C_j^{(s)} = \gamma_m$. Obviously, $T_m = 0$. We can interpret T_j as the first value of s such that the length of the tail becomes $m - j + 1$. The random variable $T_j - T_{j+1}$ is the number of steps required to grow the length of the tail from $m - j$ to $m - j + 1$. Clearly, these variables are independent for different j. Also, each of them is distributed according to a geometric distribution and standard probability theory gives us that $\mathbb{E}[T_j - T_{j+1}] = 1/p_{j,+1}$ and $\text{Var}[T_j - T_{j+1}] = (1 - p_{j,+1})/p_{j,+1}^2$. We have $p_{j,+1} = \Theta((m - j)/k)$ from Lemma 7, so

$$\mathbb{E}[T_0] = \sum_{j=0}^{m-1} \frac{1}{p_{j,+1}} = \Theta(k) \left(\sum_{j=0}^{m-1} \frac{1}{m - j} \right) = \Theta(k \log(m + 1)).$$

Similarly,

$$\mathrm{Var}[T_0] = \sum_{j=0}^{m-1} \frac{1 - p_{j+1}}{p_{j,+1}^2} = \Theta(k^2)\left(\sum_{j=0}^{m-1} \frac{1}{(m-j)^2}\right) = \Theta(k^2).$$

Hence, using Chebyshev's inequality, there exists $\ell = \Theta(k \log(m+1))$ such that

$$\Pr[T_0 > \ell] \geq 3/4.$$

Since $C_0^{(\ell)}$ can take only two values (1 and $\gamma_m \in [-1, 0]$), we have that

$$\gamma_0^{(\ell)} = \mathbb{E}[C_0^{(\ell)}] \geq 3/4 \cdot 1 + 1/4 \cdot \gamma_m \geq 1/2.$$

Finally, since $B_0^{(\ell)}$ dominates $C_0^{(\ell)}$, we get

$$\gamma_0^{(\ell)} = \mathbb{E}[B_0^{(\ell)}] \geq \mathbb{E}[C_0^{(\ell)}] \geq 1/2,$$

implying Eq. (11). This shows the existence of $\ell = \Theta(k \log(m+1))$ such that the error probability of any measurement on ℓ copies of $|\psi_x\rangle$ has error probability $\Omega(1)$ in identifying x. ◄

4   Learning from quantum samples and reflections

In the previous sections we assumed we were given a number of copies of the unknown state $|S\rangle$. In this section we assume a stronger model: in addition to a number of copies of the state $|S\rangle$, we are also given the ability to apply the reflection $R_S = 2|S\rangle\langle S| - \mathsf{Id}$ through $|S\rangle$. The key additional tool we will use is (exact) amplitude amplification, encapsulated by the next theorem, which follows from [14]:

▶ **Theorem 9** (Exact amplitude amplification). *Let $|\phi\rangle$ and $|\psi\rangle$ be states such that $\langle\phi|\psi\rangle = \alpha > 0$. Suppose we know α exactly, and we can implement reflections through $|\phi\rangle$ and $|\psi\rangle$. Then we can convert $|\phi\rangle$ into $|\psi\rangle$ (exactly) using $O(1/\alpha)$ reflections and $\widetilde{O}(1/\alpha)$ other gates.*

We distinguish the two regimes of $k \geq n/2$ and $k < n/2$.

4.1   Tight bound if $k \geq n/2$

▶ **Theorem 10** (Upper bound for small m). *Let $S \subseteq [n]$ be a set of size $k \geq n/2$ and let $m = n - k$. We can identify S with probability 1 using $O\left(\sqrt{km}\right)$ uses of $R_S = 2|S\rangle\langle S| - \mathsf{Id}$.*

Proof. Our algorithm sequentially finds all m missing elements. We would like to use amplitude amplification to prepare a copy of $|\overline{S}\rangle$, which is the uniform superposition over the m missing elements. Consider the uniform state over the n-element universe:

$$|[n]\rangle = \sqrt{\frac{k}{n}}\,|S\rangle + \sqrt{\frac{m}{n}}\,|\overline{S}\rangle.$$

This state is easy to prepare, and hence also easy to reflect through. Note that in the 2-dimensional plane spanned by $|S\rangle$ and $|\overline{S}\rangle$, reflection through $|\overline{S}\rangle$ is the same as a reflection through $|S\rangle$ up to an irrelevant global phase. The inner product between $|[n]\rangle$ and $|\overline{S}\rangle$ equals $\sqrt{m/n}$. Accordingly, using $O(\sqrt{n/m})$ rounds of exact amplitude amplification (which only rotates in the 2-dimensional space spanned by $|S\rangle$ and $|\overline{S}\rangle$; each round "costs" one application of R_S) we can turn $|[n]\rangle$ into $|\overline{S}\rangle$, up to a global phase.

Measuring $|\overline{S}\rangle$ gives us one of the missing elements, uniformly at random. Now we remove this element from the universe. Note that $|S\rangle$ does not change since we removed an element of the universe that was missing from S. We then repeat the above algorithm on a universe of size $n-1$ with $m-1$ missing elements in order to find another missing element at the cost of $O(\sqrt{(n-1)/(m-1)})$ rounds of amplitude amplification, and so on. This finds all missing elements (and hence S) with probability 1, using

$$\sum_{i=0}^{m-1} O\left(\sqrt{\frac{n-i}{m-i}}\right) = O(\sqrt{n})\sum_{j=1}^{m}\frac{1}{\sqrt{j}} = O(\sqrt{nm}) = O(\sqrt{km})$$

applications of R_S, where we used $k \geq n/2$. Note that in this regime we do not need any copies of $|S\rangle$, just reflections R_S.   ◀

▶ **Theorem 11** (Lower bound for small m). *Let $S \subseteq [n]$ be a set of size $k < n$ and let $m = n - k$. Any quantum algorithm that identifies S with high probability using a total of T copies of $|S\rangle$ and uses of R_S, must satisfy $T = \Omega\left(\sqrt{km}\right)$. The lower bound holds even if we allow T copies of $|S\rangle$, uses of R_S, and membership queries to S.*

Proof. We prove a matching lower bound in a stronger model, namely in a model where we can make queries to the n-bit characteristic vector x for S. That is, we now assume we have a unitary U_S that maps

$$U_S : |i, b\rangle \mapsto |i, b \oplus x_i\rangle \quad \text{for all } i \in [n], b \in \{0, 1\},$$

where $x_i = 1$ iff $i \in S$.

We first argue that this is indeed a stronger model, by showing how we can unitarily prepare a copy of $|S\rangle$ using $O(1)$ applications of U_S. Note that $\langle[n]|S\rangle = \sqrt{k/n} \geq 1/\sqrt{2}$ under the current assumption that $k \geq n/2$. Also note that, in the 2-dimensional space spanned by $|S\rangle$ and $|\overline{S}\rangle$, a reflection through $|S\rangle$ corresponds to a "phase query" to x, which can be implemented by one query to U_S (setting the target qubit to $(|0\rangle - |1\rangle)/\sqrt{2}$). Hence using $O(1)$ rounds of exact amplitude amplification suffices to prepare a copy of $|S\rangle$ starting from the state $|[n]\rangle$, which is easy to prepare and reflect through. Thus we can implement the state-preparation map $G_S : |0\rangle \mapsto |S\rangle$ using $O(1)$ applications of U_S. Note that one application of G_S^{-1}, followed by a reflection through $|0\rangle$ and an application of G_S, implements a reflection through $|S\rangle$. Thus preparing a copy of $|S\rangle$ and reflecting through $|S\rangle$ each "cost" only $O(1)$ queries to x (i.e., applications of U_S).

Accordingly, an algorithm that learns S using at most T copies of $|S\rangle$ and at most T applications of R_S implies a quantum algorithm that can learn an n-bit string x of weight $k \geq n/2$ using $O(T)$ queries to x. But it is known that this requires $\Omega(\sqrt{nm}) = \Omega(\sqrt{km})$ queries to x, even when allowing bounded error probability. This follows, for instance, from [11, Theorem 4.10]. Hence we obtain the same lower bound on the number of copies of $|S\rangle$ plus the number of reflections through $|S\rangle$.   ◀

4.2   Tight bound if $k < n/2$

▶ **Theorem 12** (Upper bound for small k). *Let $S \subseteq [n]$ be a set of size $k < n$. We can identify S with probability 1 using $O(k)$ copies of $|S\rangle$ and uses of $R_S = 2|S\rangle\langle S| - \mathsf{Id}$.*

Proof. Our algorithm sequentially finds all elements of S. We start with a copy of $|S\rangle$ and measure to find one $i_1 \in S$. Then we use exact amplification to convert a fresh copy of $|S\rangle$ into $|S \setminus \{i_1\}\rangle$. This requires being able to reflect through $|S\rangle$ (i.e., apply R_S), and reflect

through $|S \setminus \{i_1\}\rangle$. In the 2-dimensional plane spanned by $|S\rangle$ and $|S \setminus \{i_1\}\rangle$, the latter reflection is equivalent to putting a minus in front of $|i_1\rangle$, which is easy to do. We measure $|S \setminus \{i_1\}\rangle$ and learn (with probability 1) another element $i_2 \in S \setminus \{i_1\}$. Then we change a fresh copy of $|S\rangle$ into $|S \setminus \{i_1, i_2\}\rangle$, measure, and learn some $i_3 \in S \setminus \{i_1, i_2\}$. We repeat this until we have seen all k elements.

The amplitude amplifications get more costly as we find more elements of S: If we have already found a set $I \subseteq S$, then changing a fresh copy of $|S\rangle$ to $|S \setminus I\rangle$ uses $O(\frac{1}{\langle S|S \setminus I \rangle}) = O(\sqrt{k/(k - |I|)})$ reflections, and hence $O(\sqrt{k/(k - |I|)})$ applications of R_S. Overall, this procedure finds S using $k = |S|$ copies of $|S\rangle$, and

$$\sum_{i=0}^{k-1} O\left(\sqrt{\frac{k}{k-i}}\right) = O(\sqrt{k}) \sum_{j=1}^{k} \frac{1}{\sqrt{j}} = O(k)$$

applications of R_S. ◀

▶ **Theorem 13** (Lower bound for small k). *Let $S \subseteq [n]$ be a set of size $k < n$. Any quantum algorithm that identifies S with high probability using a total of T copies of $|S\rangle$, and uses of R_S must satisfy $T = \Omega(k)$. The lower bound holds even if we allow T copies of $|S\rangle$, uses of R_S, and membership queries to S.*

Proof. To prove a matching lower bound, suppose our algorithm receives advice in the form of $n - 2k$ of the missing elements. This advice reduces the problem to one with universe size $n' = n - (n - 2k) = 2k$ and $m' = m - (n - 2k) = k$ missing elements. Importantly, note that $|S\rangle$, and hence R_S, do not change after learning these missing elements. But in Theorem 11 we already proved an $\Omega(\sqrt{n'm'}) = \Omega(k)$ lower bound on the number of copies of $|S\rangle$, reflections, and queries to S needed to solve this special case. Since the extra advice cannot have made the original problem harder, the same lower bound applies to our original problem. ◀

5 Proper PAC learning

As mentioned briefly in the introduction, one of the motivations for this research is the question whether the sample complexity of *proper* quantum PAC learning is higher than that of improper PAC learning. Let us precisely define Valiant's PAC model [28]. We are trying to learn an unknown element f from a *concept class* $\mathcal{C}$. For simplicity we only consider fs that are Boolean-valued functions on $[n]$. Our access to f is through random examples, which are pairs of the form $(x, f(x))$, where x is distributed according to a distribution $D: [n] \to [0, 1]$ that is unknown to the learner. A learning algorithm takes a number T of such i.i.d. examples as input, and produces a hypothesis $h: [n] \to \{0, 1\}$ that is supposed to be close to the target function f. The error of the hypothesis h (with respect to the target f, under distribution D) is defined as

$$\mathrm{err}_D(f, h) := \Pr_{x \sim D}[f(x) \neq h(x)].$$

We say that a learning algorithm is an (ε, δ)-*PAC learner for* $\mathcal{C}$, if it probably (i.e., with probability at least $1 - \delta$) outputs an approximately correct (i.e., with error at most ε) hypothesis h:

$$\forall f \in \mathcal{C}, \forall D : \Pr[\mathrm{err}_D(f, h) > \varepsilon] \leq \delta,$$

where the probability is taken over the sequence of T D-distributed examples that the learner receives, as well as over its internal randomness. The (ε, δ)-*PAC sample complexity* of $\mathcal{C}$ is the minimal T for which such a learning algorithm exists.[3]

The PAC sample complexity of $\mathcal{C}$ is essentially determined by its VC-dimension d as[4]

$$\Theta\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right). \tag{13}$$

See Blumer et al. [13] for the lower bound and Hanneke [18] for the upper bound.

The above upper bound on sample complexity allows the learner to be improper, i.e., to sometimes output hypotheses $h \notin \mathcal{C}$. The following folklore example, which we learned from Steve Hanneke [19], shows that the sample complexity of *proper* learning can be asymptotically larger.[5] Consider the concept class $\mathcal{C} = \{f \colon [n] \to \{0, 1\} \mid \exists! \ i \ \text{s.t.} \ f(i) = 0\}$ of functions that are all-1 except on one "missing element" i. The VC-dimension of this class is 1, hence $\Theta\left(\frac{\log(1/\delta)}{\varepsilon}\right)$ classical examples are necessary and sufficient for PAC learning $\mathcal{C}$ by (13). With $\varepsilon = 1/n$ and $\delta = 1/3$, this bound becomes $\Theta(n)$. Now fix an (ε, δ)-PAC *proper* learner for this class that uses some T examples; we will show that $T = \Omega(n \log n)$, exhibiting an asymptotic separation between the sample complexities of proper and improper PAC learning.

For every $i \in [n]$, consider a distribution D_i that is uniform over $[n] \setminus \{i\}$. If the target concept f has i as its missing element then the learner has to output that f, since any other $g \in \mathcal{C}$ will make an error on its own missing element and hence would have error at least $1/(n-1) > \varepsilon$ under D_i. In other words, when sampling from D_i the learner has to identify the one missing element i with success probability $\geq 2/3$. But we know from the coupon collector argument that this requires $\Omega(n \log n)$ samples. Note that a D_i-distributed $(x, f(x))$ is equivalent to sampling uniformly from $[n] \setminus \{i\}$, since the label $f(x)$ is always 1 under D_i.

What about *quantum* PAC learning? Bshouty and Jackson [15] generalized the PAC model by considering superposition states

$$|\psi_{D,f}\rangle = \sum_x \sqrt{D(x)} |x, f(x)\rangle$$

instead of random samples. The learner now receives T copies of this "quantum example" state, and has to output a probably approximately correct hypothesis. Measuring a quantum example gives a classical example, so quantum examples are at least as useful as classical examples, but one of the questions in quantum learning theory is in what situations they are significantly more useful. Two of us [6] have shown that the bound of (13) also applies to learning from quantum examples, so for improper learning the quantum and classical sample complexities are equal up to constant factors. However, quantum examples *are* beneficial for learning $\mathcal{C}$ under the D_i distributions. Note that $|\psi_{D_i, f_i}\rangle$ is just the uniform superposition over the set $S = [n] \setminus \{i\}$, tensored with an irrelevant extra $|1\rangle$. As we showed in Section 2, given $O(n)$ copies of $|\psi_{D_i, f_i}\rangle$ we can identify the one missing element i with probability $\geq 2/3$. So the example that separates the sample complexities of *classical* proper and improper

---

[3] This definition uses the information-theoretic notion of *sample* complexity. We do not consider the *time* complexity of learning here. For more on sample and time complexity of quantum learning, we refer the reader to [5].

[4] The VC-dimension of $\mathcal{C}$ is the maximum size among all sets $T \subseteq [n]$ that are "shattered" by $\mathcal{C}$. A set T is shattered by $\mathcal{C}$ if for all $2^{|T|}$ labelings $\ell \colon T \to \{0, 1\}$ of the elements of T, there is an $f \in \mathcal{C}$ that has that labeling (i.e., where $f_{|T} = \ell$).

[5] In a recent result, Montasser et al. [24] proved another separation between proper and improper learning.

learning, does not separate *quantum* proper and improper learning. This naturally raises the question of whether the quantum sample complexities of proper and improper PAC learning are asymptotically equal (which, as mentioned, they provably are not in the classical case).

─── **References** ───

1   S. Aaronson, R. Kothari, W. Kretschmer, and J. Thaler. Quantum lower bounds for approximate counting via Laurent polynomials. arXiv:1904.08914, supersedes arXiv:1808.02420 and arXiv:1902.02398, 2019.

2   D. Aharonov and A. Ta-Shma. Adiabatic quantum state generation and statistical zero knowledge. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '03, page 20–29, 2003. `doi:10.1145/780542.780546`.

3   A. Ambainis, L. Magnin, M. Roetteler, and J. Roland. Symmetry-assisted adversaries for quantum state generation. In *2011 IEEE 26th Annual Conference on Computational Complexity*, pages 167–177, June 2011. `doi:10.1109/CCC.2011.24`.

4   S. Arunachalam, S. Chakraborty, T. Lee, M. Paraashar, and R. de Wolf. Two new results about quantum exact learning. In *46th International Colloquium on Automata, Languages, and Programming, ICALP*, pages 16:1–16:15, 2019. arXiv:1810.00481.

5   S. Arunachalam and R. de Wolf. Guest column: A survey of quantum learning theory. *SIGACT News*, 48(2):41–67, 2017. arXiv:1701.06806.

6   S. Arunachalam and R. de Wolf. Optimal quantum sample complexity of learning algorithms. *Journal of Machine Learning Research*, 19, 2018. Earlier version in CCC'17. arXiv:1607.00932.

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

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

9   E. Bannai and T. Itō. *Algebraic Combinatorics I: Association Schemes*. Mathematics lecture note series. Benjamin/Cummings Pub. Co., 1984.

10   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   R. Beals, H. Buhrman, R. Cleve, M. Mosca, and R. de Wolf. Quantum lower bounds by polynomials. *jacm*, 48(4):778–797, 2001. Earlier version in FOCS'98. quant-ph/9802049.

12   A. Belovs. *Applications of the Adversary Method in Quantum Query Algorithms*. PhD thesis, University of Latvia, 2014. 1402.3858.

13   A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

14   G. Brassard, P. Høyer, M. Mosca, and A. Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Quantum Information: A Millennium Volume*, volume 305 of *AMS Contemporary Mathematics Series*, pages 53–74. AMS, 2002. quant-ph/0005055.

15   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.

16   C. Godsil. Association schemes. Available at `https://www.math.uwaterloo.ca/~cgodsil/assocs/pdfs/Assoc.pdf`, 2018.

17   A. B. Grilo, I. Kerenidis, and T. Zijlstra. Learning with Errors is easy with quantum samples. *Physical Review Letters A*, 99:032314, 2019. arXiv: 1702.08255.

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

19   S. Hanneke. Personal communication with Srinivasan Arunachalam, February 2018. See also `https://cstheory.stackexchange.com/questions/40161/proper-pac-learning-vc-dimension-bounds`.

20 P. Høyer, T. Lee, and R. Špalek. Negative weights make adversaries stronger. In *Proceedings of 39th ACM STOC*, pages 526–535, 2007. quant-ph/0611054.

21 T. Lee, R. Mittal, B. Reichardt, R. Špalek, and M. Szegedy. Quantum query complexity of state conversion. In *Proceedings of 52nd IEEE FOCS*, pages 344–353, 2011. arXiv:1011.3020.

22 T. Lee and J. Roland. A strong direct product theorem for quantum query complexity. *Computational Complexity*, 22(2):429–462, 2013. Earlier version in CCC'12. arXiv:1104.4468.

23 N. Lindzey and A. Rosmanis. A Tight Lower Bound For Non-Coherent Index Erasure. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 59:1–59:37, 2020. `doi:10.4230/LIPIcs.ITCS.2020.59`.

24 O. Montasser, S. Hanneke, and N. Srebro. VC classes are adversarially robustly learnable, but only improperly. *Journal of Machine Learning Research*, 99:1–19, 2019.

25 R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge International Series on Parallel Computation. Cambridge University Press, 1995. `doi:10.1017/CBO9780511814075`.

26 A. Rosmanis. *Lower Bounds on Quantum Query and Learning Graph Complexities*. PhD thesis, University of Waterloo, July 2014. Available at `http://hdl.handle.net/10012/8577`.

27 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.

28 L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134—-1142, 1984.