Learning k-qubit Quantum Operators via Pauli Decomposition

Mohsen Heidari
Indiana University, Bloomington

Wojciech Szpankowski
Purdue University, West Lafayette

Abstract

Motivated by the limited qubit capacity of current quantum systems, we study the quantum sample complexity of k-qubit quantum operators, i.e., operations applicable on only k out of d qubits. The problem is studied according to the quantum probably approximately correct (QPAC) model abiding by quantum mechanical laws such as no-cloning, state collapse, and measurement incompatibility. With the delicacy of quantum samples and the richness of quantum operations, one expects a significantly larger quantum sample complexity.

This paper proves the contrary. We show that the quantum sample complexity of k-qubit quantum operations is comparable to the classical sample complexity of their counterparts (juntas), at least when $\frac{k}{d} \ll 1$. This is surprising, especially since sample duplication is prohibited, and measurement incompatibility would lead to an exponentially larger sample complexity with standard methods. Our approach is based on the Pauli decomposition of quantum operators and a technique called Quantum Shadow Sampling (QSS) to reduce the sample complexity exponentially. The results are proved by developing (i) a connection between the learning loss and the Pauli decomposition; (ii) a scalable QSS circuit for estimating the Pauli coefficients; and (iii) a quantum algorithm for learning k-qubit operators with sample complexity $O(\frac{k4^k}{\epsilon^2}\log d)$.

1 Introduction

Quantum-enhanced learning is one of the leading applications of quantum computers (QC) both for classical data (Giovannetti et al., 2008; Park et al., 2019; Lloyd et al., 2014; Schuld et al., 2020) and inherently quantum samples (Carleo and Troyer, 2017; Massoli et al., 2021; Lu et al., 2018). However, current state-of-the-art QCs have a limited qubit capacity of up to a few hundred qubits with infidelity. On the other hand, the dimension of quantum systems in typical applications far exceeds the qubit capacity of *near-term* QCs. Therefore, it is crucial to understand the fundamental limits of near-term QCs for learning applications. Motivated by this observation, we study the learning capability of k-qubit operations in d-qubit systems, where k is significantly smaller than d. Particularly, we characterize bounds on the quantum sample complexity of k-qubit systems and propose a quantum learning algorithm achieving the minimum learning loss.

There are several quantum learning models such as state discrimination, quantum property testing, and quantum state classification (see Section 1.1 for related works). For a comprehensive study, we consider a general formulation incorporating such models as special cases. In classical settings, *probably approximately correct* (PAC), developed by (Kearns et al., 1994; Valiant, 1984), is a concrete model to study fundamental limits such as sample complexity without any distributional or structural assumptions. In this work, we consider the quantum counterpart of this model, known as QPAC (Heidari et al., 2021).

This learning model consists of a set of n labeled qubits $(\rho_i, y_i)_{i=1}^n$ as the training samples. These samples might be classical or quantum originally. There is no structural assumption about the samples other than (1) being d-qubit states; and (2) being generated independently and identically distributed (i.i.d.) according to an unknown but fixed probability distribution. The samples are processed by a QC with a measurement at the end layer. We seek a procedure that takes the training set and tunes the quantum operations based on a library of choices (concept class). The objective is to minimize the loss in predicting labels of the next unseen quantum states. Quantum sample complexity is, then, the minimum number of required samples to obtain a minimal loss.

It is not difficult to see that this model subsumes several well-studied models. For example, state discrimination is a special case in which ρ_i's are identical and are equal to one of two known possible states. Classical learning is also

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, TBD, USA. PMLR: Volume XXX. Copyright 2023 by the author(s).

a special case in which ρ_i's are pure states $|x_i\rangle\langle x_i|$ with x_i representing the classical samples. Therefore, QPAC is a stronger requirement than these special cases, as it is a distribution-free and state-free condition. Whereas PAC is only distribution-free and state discrimination assumes certain prior structures. In addition, more difficulties arise from the quantum nature of the problem. The quantum samples are irreversibly disturbed by the algorithm due to state collapse. Further, sample duplication is prohibited abiding by the *no-cloning* principle.

With the delicacy of quantum samples and the richness of quantum operations, quantum sample complexity is expected to be significantly larger than the classical one. To see this, one natural approach for learning k-qubit operations in d-qubit systems is via *state tomography* with classical post proceeding. One first performs *state tomography* on each sample to arrive at an approximate description of the quantum states and then performs classical learning algorithms on the stored density matrices. This approach requires $O(\frac{1}{\epsilon^2}2^{2d})$ identical copies of the samples (Haah et al., 2016). As a comparison to the classical setting, one considers learning of k-juntas which are Boolean functions depending on k out of d inputs (Mossel et al., 2003). It is known that the sample complexity of k-juntas is $O(\frac{k}{\epsilon^2}\log\frac{d}{k})$ (Shalev-Shwartz and Ben-David, 2014). This observation implies that the quantum sample complexity might be exponentially larger.

Contributions: In this paper, we prove the contrary and show that the quantum sample complexity of k-qubit operations scales logarithmically with d. Although not equal, it is comparable to the classical sample complexity of k-juntas for small values of k. More precisely, we prove in Theorem 1 that the quantum sample complexity of k-qubit operations is $O(\frac{k4^k}{\epsilon^2}\log d)$. Furthermore, we strengthen this existential result by designing a quantum algorithm achieving this bound (see Algorithm 1).

Our approach is based on a Pauli decomposition of quantum operators and an estimation procedure called Quantum Shadow Sampling (QSS). We establish a connection between the learning loss and the Pauli decomposition of the induced operator of the training samples. We then develop a novel approach for estimating the Pauli coefficients of this induced operator. We argue that naive empirical estimations require $O(\frac{(4d)^k}{\epsilon^2})$ quantum samples. Hence, they lead to an exponentially larger sample complexity as compared to classical (See Section 3.2). We address this issue and propose QSS that reduces the quantum sample complexity to $O(\frac{k4^k}{\epsilon^2}\log d)$ that scales with the logarithm of the number of qubits (see Theorem 2 and Section 3.5). For that, we design an estimation circuit with $O(d)$ gate complexity. This design is scalable as it consists of a parallel set of completely independent sub-circuits each acting on a single qubit (see Figure 2). With this estimation, we develop our algorithm and prove that it learns the k-qubit operators without any distributional or structural assumption (*agnostic* QPAC). Lastly, in Section 4 we verify our results with a numerical experiment for detecting maximally entangled from separable qubits.

1.1 Related Works

The literature in this area is broad. We only can give pointers to a few of the best-known and most relevant works.

Quantum enhanced learning has been studied extensively for classical data (Schuld et al., 2014; Giovannetti et al., 2008; Park et al., 2019; Rebentrost et al., 2014; Lloyd et al., 2013, 2014) and for quantum data in recent literature in the context of diverse applications, including condensed matter for phase-of-matter detection (Carrasquilla and Melko, 2017; Broecker et al., 2017), ground-state search (Carleo and Troyer, 2017; Broughton et al., 2020; Biamonte et al., 2017), entanglement detection (Ma and Yung, 2018; Massoli et al., 2021; Lu et al., 2018; Hiesmayr, 2021; Chen et al., 2021; Deng et al., 2017), and other applications (Kassal et al., 2011; McArdle et al., 2020; Hempel et al., 2018; Cao et al., 2019; Heidari et al., 2022; Bauer et al., 2020).

There are several solutions and models for quantum learning. In *state tomography* the objective is to find an approximate description of an unknown quantum state ρ using measurements on multiple copies of the state. This problem has been studied under various distance/fidelity measures (O'Donnell and Wright, 2016, 2017; Haah et al., 2016). *State Certification* can be viewed as a quantum counterpart of property testing in which we would like to check where $\rho = \sigma$ or ϵ far away from it (Badescu et al., 2019; Bubeck et al., 2020). This is again done by measuring multiple identical copies of ρ. A survey on this topic is provided in (Montanaro and de Wolf, 2016). In *state discrimination* we want to tell whether $\rho = \sigma_1$ or σ_2 (Barnett and Croke, 2009; Gambs, 2008; Guta and Kotlowski, 2010). Another framework is *quantum hypothesis testing* as surveyed in (Audenaert et al., 2008). An operational view of learning quantum states is introduced by (Aaronson, 2007). In this work, the training samples are i.i.d. measurements. The objective is to approximate the acceptance probability $\mathrm{tr}\{E\rho\}$ for most measurement E. Another related work in this line is (Cheng et al., 2015), where there is an unknown measurement E to be learned from samples. The training samples are $\{(\rho_i, \mathrm{tr}\{E\rho_i\})\}_{i=1}^n$, where ρ_i's are i.i.d. random quantum states. At first glance, this formulation seems similar to our problem. However, as a careful reader will recognize, ρ_i's are pre-measured states. Contrary to this model, in our work, simultaneous access to pre-measured states and the measurement's outcomes are prohibited. Another distinction is that the probabilities $\mathrm{tr}\{E\rho_i\}$ are unknown in this paper. Another direction is based on the well-known work of Bshouty and Jackson

(1998). In this model, we measure identical copies of a *superposition* state to solve a classical PAC learning problem. This model is also different from QPAC in our paper, as the concept class in QPAC consists of quantum measurements rather than classical functions. Hence, QPAC is expected to subsume their model as well. Other related works in this area are (Arunachalam and de Wolf, 2017; Arunachalam and De Wolf, 2018; Kanade et al., 2019; Bernstein and Vazirani, 1997; Servedio and Gortler, 2004). Lastly, estimating the decomposition of an operator with respect to a set of elementary operators has been studied in (Crawford et al., 2020; Peruzzo et al., 2014).

2 Model Formulation

Notations: For shorthand, denote $[d]$ as $\{1, 2, ..., d\}$. Also, for any $\mathbf{s} \in \{0, 1, 2, 3\}^d$, define $\text{supp}(\mathbf{s}) := \{\ell \in [d] : s_\ell \neq 0\}$. For any $d \in \mathbb{N}$, let H_d be the Hilbert space of d-qubits. The identity operator on H_d is denoted by I_d. As usual, a quantum state is defined as a *density operator*; that is a Hermitian, unit-trace, and non-negative linear operator. A quantum measurement $\mathcal{M}$ is a positive operator-valued measure (POVM) represented by a set of operators $\mathcal{M} := \{M_v, v \in \mathcal{V}\}$, where $\mathcal{V}$ is theset of possible outcomes, $M_v \geq 0$ for any $v \in \mathcal{V}$, and $\sum_{v\in\mathcal{V}} M_v = I_d$. For an operator A, denote $\|A\|_1 = \text{tr}\{|A|\}$ as the trace norm, and $\|A\|_2 = \sqrt{\text{tr}\{A^\dagger A\}}$ as Hilbert–Schmidt norm.

2.1 Quantum Learning Model

Before presenting the main results, we formally define our quantum learning model. In this model (Heidari et al., 2021), the objective is to distinguish between multiple groups of unknown quantum states without prior knowledge about the states. Available is only a training set of quantum states with a classical label determining their group index. We seek an agnostic procedure that given enough samples learns the labeling law. The model in the binary case is defined more precisely as follows.

Let ρ_0 and ρ_1 be two unknown quantum states denoting each of the possible states of an unknown physical system. We associate to each state a label $y \in \{0, 1\}$. Let $p_0 = 1 - p_1 \in (0, 1)$ be an unknown probability distribution on $\{0, 1\}$. Each time, a sample ρ is randomly generated where $\rho = \rho_0$ with probability p_0 and $\rho = \rho_1$ with probability p_1. The objective is to tell which of the two states is generated without knowing what ρ_0, ρ_1 and/or (p_0, p_1) are. Available are only n training samples $\{(\rho_{y_i}, y_i)\}_{i=1}^n$, generated i.i.d. according to (p_0, p_1). We seek a procedure that given the training samples constructs a quantum measurement to distinguish between ρ_0 and ρ_1 with high accuracy.

A predictor is a quantum measurement that acts on the quantum state and outputs $\hat{y} \in \{0, 1\}$ as the predicted label.

The accuracy of a prediction measurement $\mathcal{M} := \{M_0, M_1\}$ is determined by randomly generating a test sample $(\rho_{y_{test}}, y_{test})$. Without revealing y_{test}, we measure $\rho_{y_{test}}$ to get $\hat{y}_{test}$. We use the 0-1 loss to measure the prediction error, that is $\mathbb{1}\{y_{test} \neq \hat{y}_{test}\}$. Hence, from Born's rule, the (expected) loss is calculated as

$$L_{0-1}(\mathcal{M}) = p_0 \,\text{tr}\{M_1\rho_0\} + p_1 \,\text{tr}\{M_0\rho_1\},$$

where the first and the second trace are the probability that erroneously $\hat{y}_{test} = 1$ and $\hat{y}_{test} = 0$, respectively. It is assumed that $\mathcal{M}$ belongs to a collection $\mathcal{C}$ of choices as the *concept class*. With this setup, a quantum learning algorithm is a process that selects a predictor $\mathcal{M}$ from $\mathcal{C}$, with the training samples as the input. We are interested in algorithms with guaranteed learning irrespective of ρ_0, ρ_1, p_0 and p_1.

Definition 1 (QPAC). *A quantum learning algorithm QPAC learns a measurement class $\mathcal{C}$ if there exists a function $n_\mathcal{C} : (0, 1)^2 \mapsto \mathbb{N}$ such that for every $\epsilon, \delta \in [0, 1]$ and given $n > n_\mathcal{C}(\epsilon, \delta)$ samples drawn i.i.d. according to any probability distributions (p_0, p_1) and from any unknown states (ρ_0, ρ_1), the algorithm outputs, with probability of at least $(1 - \delta)$, a measurement whose loss is less than $\inf_{\mathcal{M}\in\mathcal{C}} L_{0-1}(\mathcal{M}) + \epsilon$.*[1]

Consequently, the quantum sample complexity of a concept class $\mathcal{C}$ is the minimum of $n_\mathcal{C}$ for which there exists a QPAC learning algorithm. The focus of this study is on k-qubit operators that are formally defined below.

Definition 2 (k-qubit Operators). *An operator A on H_d is said to be a k-qubit operator, if there exists a coordinate subset $\mathcal{J} \subset [d]$ with $|\mathcal{J}| \leq k$ such that $A = \tilde{A}_\mathcal{J} \otimes I_{[d]\setminus\mathcal{J}}$, where $\tilde{A}$ is an operator on the subsystem corresponding to the coordinates $\mathcal{J}$ and $I_{[d]\setminus\mathcal{J}}$ is the identity operator on the residual sub-system.*

Classical counter parts of k-qubit operators are k-junta Boolean functions (Mossel et al., 2004). k-qubit operators subsumes k-juntas. They are significantly richer than their classical counterpart. While there are $\binom{d}{k}2^{2^k}$ juntas; k-qubit operators are infinite. The input dimension for a k-junta is k; while that of a k-qubit operator is 2^k. One can learn k-juntas by performing a brute-force exhaustive search over all k-juntas and finding the one minimizing the empirical loss. However, the learning task becomes more difficult in the quantum settings as there are infinitely many k-qubit circuits and sample duplication is prohibited. Therefore, with the richness of quantum concept classes and the the fragility of quantum samples, one wonders whether quantum learning is harder. In the next section, we show it is not, but it requires looking at the problem from a different angle.

[1] Naturally, we are interested in efficient learning with $n_\mathcal{C}$ being at most polynomial in ϵ, δ and d.

3 Main Results

Our first main contribution is the following theorem that is proved in Section 3.5.

Theorem 1. *There exists a quantum algorithm that QPAC learns k-qubit operators with an error up to*

$$\mathsf{opt}_k + \mathcal{O}\left(\sqrt{\frac{4^k}{n}\log\left(\frac{d^k 4^k}{\delta(k-1)!}\right)}\right),$$

where opt_k is the minimum loss of the concept class, and this is achieved by Algorithm 1.

With this result, for small $\frac{k}{d}$, the quantum sample complexity of k-qubit operators is simplified to $O(\frac{k4^k}{\epsilon^2}\log\frac{d}{\delta})$ which grows with logarithm of d, the number of qubits.

Next, we study a lower bound on the quantum sample complexity. Given that QPAC subsumes PAC and that k-juntas are special cases of k-qubit operations, the quantum sample complexity is lower bounded by the classical one. Hence, from Vapnik–Chervonenkis (VC) theory for k-juntas, (Shalev-Shwartz and Ben-David, 2014), we obtain the following lower bound.

Proposition 1. *The quantum sample complexity of k-qubit operations is $\Omega(\frac{1}{\epsilon^2}(k\log\frac{2d}{k}+\log\frac{1}{\delta}))$.*

This result and Theorem 1 suggest that quantum sample complexity is of the same order as the classical one at least for small values of k compared to d. Hence, though QPAC is a more difficult problem and low-width quantum circuits are much richer than classical juntas, yet the quantum sample complexity grows similarly for small k's. Whether the same holds for larger values of k is yet to be determined. In Section 3.2, we argue that primitive empirical estimation methods are not efficient in QPAC and that one needs a more sophisticated approach as in Algorithm 1. Before that, we present an overview of the Pauli decomposition and study its connection to learning loss.

3.1 Pauli Decomposition

Our approach relies on the Pauli decomposition of quantum operators (Montanaro and Osborne, 2010). We start with a brief overview of this decomposition. Then, we analyze the connection between the $0-1$ loss and the Pauli coefficients.

The Pauli operators with the identity are denoted as $\{\sigma^0, \sigma^1, \sigma^2, \sigma^3\}$ with $\sigma^0 = I_2$ and

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Define the Pauli tensor products as

$$\sigma^{\mathbf{s}} := \sigma^{s_1} \otimes \sigma^{s_2} \otimes \cdots \otimes \sigma^{s_d}, \qquad \forall \mathbf{s} \in \{0,1,2,3\}^d. \quad (1)$$

Fact 1 (Pauli Decomposition)**.** *Any bounded operator A on H_d is uniquely decomposed as*

$$A = \sum_{\mathbf{s}\in\{0,1,2,3\}^d} a_{\mathbf{s}}\, \sigma^{\mathbf{s}},$$

where $a_{\mathbf{s}} \in \mathbb{C}$ are the Pauli coefficients of A and are given by $a_{\mathbf{s}} = \frac{1}{2^d}\operatorname{tr}\{A\sigma^{\mathbf{s}}\}$.[2]

An immediate consequence of this decomposition is the following identity for any pair of operators on H_d:

$$\operatorname{tr}\{AB\} = 2^d \sum_{\mathbf{s}} a_{\mathbf{s}} b_{\mathbf{s}}, \qquad (2)$$

where $a_{\mathbf{s}}$ and $b_{\mathbf{s}}$ are the Pauli coefficients of A and B.

Next, we present the connection between the Pauli coefficients and the learning loss.

Lemma 1. *Let $\rho_{XY} = p_0\rho_0 \otimes |0\rangle\langle 0| + p_1\rho_1 \otimes |1\rangle\langle 1|$ denote the average state of the training samples. Then, the loss of any measurement $\mathcal{M} := \{M_0, M_1\}$ decomposes as*

$$L_{0-1}(\mathcal{M}) = \frac{1}{2} - 2^{d-1}\sum_{\mathbf{s}} g_{\mathbf{s}} f_{\mathbf{s}},$$

where $g_{\mathbf{s}}$ and $f_{\mathbf{s}}$ are the Pauli coefficients of $\mathrm{G}_M := M_1 - M_0$ and $\mathrm{F}_Y := -\sqrt{\rho_{XY}}(I_d \otimes \sigma^3)\sqrt{\rho_{XY}}$, respectively.

Proof. Given $M_1 = I_d - M_0$, the loss can be written as

$$\begin{aligned} L_{0-1}(\mathcal{M}) &= p_1\operatorname{tr}\{M_0\rho_1\} + p_0\operatorname{tr}\{M_1\rho_0\} \\ &= -\sum_y p_y(-1)^y \operatorname{tr}\{M_0\rho_y\} + p_0. \qquad (3) \end{aligned}$$

Observe that $(I_d \otimes \sigma^3)\rho_{XY} = \sum_y p_y(-1)^y \rho_y \otimes |y\rangle\langle y|$. Then, from the definition of F_Y and G_M, we have that

$$\begin{aligned} \operatorname{tr}\{G_M F_Y\} &= \operatorname{tr}\{F_Y\} - 2\operatorname{tr}\{M_0 F_Y\} \\ &= -\operatorname{tr}\big\{(I_d \otimes \sigma^3)\rho_{XY}\big\} - 2\operatorname{tr}\{M_0 F_Y\} \\ &= -\mathbb{E}_Y[(-1)^Y] + 2\sum_y p_y(-1)^y \operatorname{tr}\{M_0\rho_y\} \\ &\overset{(a)}{=} -\mathbb{E}[(-1)^Y] + 2\Big(p_0 - L_{0-1}(\mathcal{M})\Big) \\ &= 2p_1 - 1 + 2\Big(p_0 - L_{0-1}(\mathcal{M})\Big) \\ &= 1 - 2L_{0-1}(\mathcal{M}), \end{aligned}$$

where (a) follows from (3). Hence,

$$L_{0-1}(\mathcal{M}) = \frac{1}{2} - \frac{1}{2}\operatorname{tr}\{G_M F_Y\}.$$

Hence, the proof is complete, because from (2), the trace term above equals to $2^d \sum_{\mathbf{s}} g_{\mathbf{s}} f_{\mathbf{s}}$. □

We note that F_Y is viewed as the induced operator representing the labeled samples. In agnostic settings, F_Y is unknown as the states and the probabilities are unknown. We design our learning algorithm by estimating the Pauli coefficient of F_Y instead.

[2] The factor 2^d is because $\operatorname{tr}\{\sigma^{\mathbf{s}}\sigma^{\mathbf{s}}\} = \operatorname{tr}\{I_d\} = 2^d$.

3.2 Estimating the Pauli Coefficients

In light of the previous section, the main idea behind the proposed algorithm is to estimate a subset of the Pauli coefficients of the auxiliary operator F_Y. In the classical setting, estimating the Fourier coefficients is easily done by empirical averaging. In quantum, each coefficient $f_{\mathbf{s}}$ is indeed an observable acting on the samples' quantum state. The issue is that these observables are *incompatible* and, thus, are not simultaneously measurable.

Example 1. *In a single qubit system, the Pauli coefficients corresponding to σ^1 and σ^2 are incompatible as they do not commute with each other. Indeed, they are mutually unbiased observables. Hence, independent samples are needed for estimating each coefficient.*

The incompatibility and no-cloning make the estimation process more challenging than the classical one. In this section, we discuss the estimation process and derive bounds on the square loss. In the next section, we discuss the construction of a predictor from the estimated coefficients.

We start with estimating a single Pauli coefficient $f_{\mathbf{s}}$. For that we consider the POVM $M_{\mathbf{s}} := \{\Lambda_1^{\mathbf{s}}, \Lambda_{-1}^{\mathbf{s}}\}$ with outcomes in $\{-1, 1\}$ and operators

$$\Lambda_1^{\mathbf{s}} := \sigma_+^{\mathbf{s}}, \qquad \Lambda_{-1}^{\mathbf{s}} := \sigma_-^{\mathbf{s}}, \tag{4}$$

where $\sigma^{\mathbf{s}}$ is the Pauli operator corresponding to $\mathbf{s} \in \{0,1,2,3\}^d$ as in (1). Moreover, $\sigma_+^{\mathbf{s}}$ and $\sigma_-^{\mathbf{s}}$ are the positive and negative part of $\sigma^{\mathbf{s}}$ (such that $\sigma^{\mathbf{s}} = \sigma_+^{\mathbf{s}} - \sigma_-^{\mathbf{s}}$), constructed through the spectral decomposition of $\sigma^{\mathbf{s}}$. With these definitions, $f_{\mathbf{s}}$ is estimated by measuring each sample with $M_{\mathbf{s}}$. Note that we cannot use all the samples for estimating one coefficient; because the samples will be inaccessible as they collapse by the measurements. Suppose, we only use $m < n$ samples. Let $Z_i \in \{-1, 1\}$ be the output of $M_{\mathbf{s}}$ on the ith sample $(\rho_i, y_i), i = 1, 2, 3..., m$. Then, the estimation is computed as

$$\hat{f}_{\mathbf{s}} = \frac{-1}{m2^d} \sum_{i=1}^{m} (-1)^{y_i} Z_i. \tag{5}$$

From Born's rule, Z_i is a binary random variable with bias $\operatorname{tr}\{\sigma_+^{\mathbf{s}} \rho_i\}$. Hence, $\hat{f}_{\mathbf{s}}$ itself is random and it is not difficult to check that $\mathbb{E}[\hat{f}_{\mathbf{s}}] = f_{\mathbf{s}}$. Therefore, using standard concentration inequalities, we can show that, for any $\delta \in [0,1]$, with probability $(1-\delta)$, the estimation error is bounded as:

$$|\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}| = 2^{-d} \mathcal{O}\left(\sqrt{\frac{1}{m} \log \frac{1}{\delta}}\right). \tag{6}$$

Note that 2^{-d} is due to the normalization of $f_{\mathbf{s}}$ as in Fact 1 and $\hat{f}_{\mathbf{s}}$ as in (5).

For learning k-qubit measurements, all the Pauli coefficients $\sigma^{\mathbf{s}}$ with $|\operatorname{supp}(\mathbf{s})| \leq k$ need to be estimated. Let K be the number of such coefficients. Given that $k \leq d/2$, we bound K as

$$K \leq \sum_{\ell=0}^{k} \binom{d}{\ell} 4^\ell \leq 1 + k \binom{d}{k} 4^k = 1 + \frac{d^k}{(k-1)!} 4^k. \tag{7}$$

Given the incompatibility of the related observables, with a naive strategy, one would partition the total n samples into several equal-size groups one for each coefficient. Hence, with this approach and (6), the estimation loss satisfies

$$|\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}| = \mathcal{O}\left(2^{-d} \sqrt{\frac{K}{n} \log(1/\delta)}\right),$$

for all $\mathbf{s}$, with $|\operatorname{supp}(\mathbf{s})| \leq k$. In what follows, we propose an approach to exponentially reduce the estimation error.

Theorem 2. *Given any $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_K \in \{0,1,2,3\}^d$, there exists an algorithm that, given n training samples, estimates the corresponding Pauli coefficients of F_Y with an error bounded with probability at least $(1-\delta)$ as*

$$\sup_{j \in [K]} |\hat{f}_{\mathbf{s}_j} - f_{\mathbf{s}_j}| = \mathcal{O}\left(2^{-d} \sqrt{\frac{1}{n} \log\left(\frac{K}{\delta}\right)}\right).$$

For our case, $\mathbf{s}_j$'s are all $\mathbf{s}$ with $|\operatorname{supp}(\mathbf{s})| \leq k$, and K is as in (7). Hence, compared to the naive strategy with a fresh copy for each coefficient, we get exponential improvements.

3.2.1 Quantum Shadow Sampling

Our approach is inspired by Shadow Tomography (Aaronson, 2018; Huang et al., 2020), where repeated measurements obtain an approximate description of an unknown quantum state from its exact copies. In view of the no-cloning, in our work, we propose an alternate approach called QSS that takes a single quantum state and can generates multiple samples called shadows. This is a one-shot procedure that applies to each sample $(\rho_i, y_i), i \in [n]$ and is explained below:

First, we generate a unitary operator U_i randomly and uniformly from the space of all unitary operators on d qubits. We rotate ρ_i by applying U_i resulting the state $U_i^\dagger \rho_i U_i$. Then, we measure the rotated state along the computational basis $\left\{|b\rangle\langle b|, b \in \{0,1\}^d\right\}$. From Born's rule the probability of getting the output $b_i \in \{0,1\}^d$ is $P_{b_i} = \langle b_i | U_i^\dagger \rho_i U_i | b_i \rangle$. At the next step, given each output $b_i \in \{0,1\}^d$, the state $\omega_i = U_i |b_i\rangle\langle b_i| U_i^\dagger$ is prepared. Hence, with ρ_i we obtain the state ω_i with probability P_{b_i}.

Define the following mapping on any operator B on H_d:

$$\Gamma[B] := \mathbb{E}_U \Big[\sum_{b \in \{0,1\}^d} \langle b | U^\dagger B U | b \rangle \; U |b\rangle\langle b| U^\dagger \Big]. \tag{8}$$

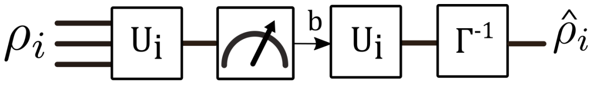

Figure 1: The process for estimating the Pauli coefficient $f_{\mathbf{s}}$. Once $\hat\rho_i$ is generated from the ith sample, we calculate $\mathrm{tr}\{\hat\rho_i \sigma^{\mathbf{s}}\}(-1)^{y_i}$. Then, the estimate $\hat f_{\mathbf{s}}$ is calculated by computing the empirical average over all the samples as in (9).

Note that Γ is a linear mapping on the space of density operators with its inverse denoted as Γ^{-1}. Moreover, observe that $\Gamma[\rho_i]$ equals to the expectation $\mathbb{E}[\omega_i]$ over the measurement randomness (P_b) and the choices of unitary U_i.

At our last step, we apply Γ^{-1} on ω_i resulting in the following state

$$\hat\rho_i := \Gamma^{-1}\big[U_i\,|b_i\rangle\langle b_i|\,U_i^\dagger\big].$$

Repeating this process for all samples, we obtain the shadow samples $\hat\rho_i, i \in [n]$. This process is demonstrated in Figure 1.

3.3 Proof of Theorem 2

After applying QSS on the entire training samples, we estimate each $f_{\mathbf{s}_j}$ by computing

$$\hat f_{\mathbf{s}_j} = \frac{1}{n} 2^{-d} \sum_{i=1}^{n} \mathrm{tr}\{\hat\rho_i \sigma^{\mathbf{s}_j}\}(-1)^{y_i}, \tag{9}$$

for all $j \in [K]$. We proceed with the following lemmas for the analysis.

Lemma 2. *$\hat\rho_i$ is an unbiased estimate of ρ_i, that is $\mathbb{E}_{U,b}[\hat\rho_i] = \rho_i$.*

Proof. By linearity of Γ^{-1}, taking the expectation of $\hat\rho_i$ over the choice of U_i and the randomness of b_i gives

$$\mathbb{E}_{\sim(U_i,b_i)}[\hat\rho_i] = \Gamma^{-1}\big[\mathbb{E}[U_i\,|b_i\rangle\langle b_i|\,U_i^\dagger]\big].$$

The expectation term equals to

$$\begin{aligned}\mathbb{E}\Big[U_i\,|b_i\rangle\langle b_i|\,U_i^\dagger\Big] &= \mathbb{E}_U\Big[\sum_{\mathbf{b}} \langle \mathbf{b}|U^\dagger \rho_i U|\mathbf{b}\rangle\ U\,|\mathbf{b}\rangle\langle \mathbf{b}|\,U^\dagger\Big] \\ &= \Gamma[\rho_i],\end{aligned}$$

where the last equality is from (8). □

Lemma 3. *The estimation $\hat f_{\mathbf{s}}$ is unbiased, that is $\mathbb{E}[\hat f_{\mathbf{s}}] = f_{\mathbf{s}}$, where the expectation is taken over all sources of randomness including the sample distribution.*

Proof. By taking the expectation, from Lemma 2 we obtain the following chain of equalities:

$$\begin{aligned}\mathbb{E}[\hat f_{\mathbf{s}}] &= 2^{-d}\mathbb{E}[\mathrm{tr}\{\hat\rho_1 \sigma^{\mathbf{s}}\}(-1)^{Y_1}] \\ &= 2^{-d}\mathbb{E}_{\sim(\rho_1,Y_1)}[\mathrm{tr}\{\mathbb{E}[\hat\rho_1|\rho_1]\sigma^{\mathbf{s}}\}(-1)^{Y_1}] \\ &= 2^{-d}\mathbb{E}_{\sim(\rho_1,Y_1)}[\mathrm{tr}\{\rho_1 \sigma^{\mathbf{s}}\}(-1)^{Y_1}] \\ &= 2^{-d}\,\mathrm{tr}\{F_Y \sigma^{\mathbf{s}}\} = f_{\mathbf{s}},\end{aligned}$$

where we used the definition of $f_{\mathbf{s}}$ in Lemma 1. □

Lastly, with Lemma 3, we apply the Chernoff inequality:

$$\mathbb{P}\Big\{\max_{j\in[K]} |\hat f_{\mathbf{s}_j} - f_{\mathbf{s}_j}| \geq \epsilon 2^{-d}\Big\} \leq 2K\,\exp\Big\{-\frac{n\epsilon^2}{2}\Big\}.$$

Equating the right-hand side to δ, we obtain the following bound on the estimation error:

$$\max_{j\in[K]} |\hat f_{\mathbf{s}_j} - f_{\mathbf{s}_j}| = \mathcal{O}\Bigg(2^{-d}\sqrt{\frac{1}{n}\log\Big(\frac{K}{\delta}\Big)}\Bigg). \tag{10}$$

With this inequality, we establish Theorem 2.

3.4 Creating the Predictor

Next, we describe the construction of a predictor using the estimated Pauli coefficients. Let $\mathcal{J} \subseteq [d]$ be the coordinate of a subsystem with k qubits. Define

$$F_Y^{\mathcal{J}} := \sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s})\subseteq\mathcal{J}} f_{\mathbf{s}}\sigma^{\mathbf{s}},$$

where $\mathrm{supp}(\mathbf{s}) := \{\ell \in [d] : s_\ell \neq 0\}$ for any $\mathbf{s} \in \{0,1,2,3\}^d$. Define the estimate of this operator as

$$\hat F_Y^{\mathcal{J}} := \sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s})\subseteq\mathcal{J}} \hat f_{\mathbf{s}}\sigma^{\mathbf{s}},$$

where $\hat f_{\mathbf{s}}$'s are the estimated Pauli coefficients. This operator has a spectral decomposition of the form

$$\hat F_Y^{\mathcal{J}} = \sum_i \lambda_i\,|\phi_i\rangle\langle\phi_i|\,.$$

Let $\hat\Pi^{\mathcal{J}}$ be the projection onto the subspace spanned by eigenstates with positive eigenvalues, i.e.,

$$\hat\Pi^{\mathcal{J}} := \sum_{i:\lambda_i>0} |\phi_i\rangle\langle\phi_i|\,. \tag{11}$$

Then, we create our predictor as the measurement $\hat{\mathcal{M}}_{\mathcal{J}} := \{\hat\Pi^{\mathcal{J}}, I_d - \hat\Pi^{\mathcal{J}}\}$. In what follows, we study the learning loss of $\hat{\mathcal{M}}_{\mathcal{J}}$. We show that if $\mathcal{J}$ is chosen appropriately, then the loss of $\hat{\mathcal{M}}_{\mathcal{J}}$ is close to the optimal value opt_k. For that, we present the following theorem.

Theorem 3. *Let $\mathcal{J}^*$ be the subset maximizing $\left\|\mathrm{F}_Y^{\mathcal{J}}\right\|_1$ among all k-element subsets. Let $\hat{\mathcal{M}}_{\mathcal{J}^*} = \{\hat{\Pi}^{\mathcal{J}^*}, I - \hat{\Pi}^{\mathcal{J}^*}\}$ be the measurement with the projection $\hat{\Pi}^{\mathcal{J}^*}$ given in (11) but with $\mathcal{J} = \mathcal{J}^*$. Then,*

$$L_{0-1}(\hat{\mathcal{M}}_{\mathcal{J}^*}) \leq \mathsf{opt}_k + 4\sqrt{2^d}\left\|\mathrm{F}_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2 \tag{12}$$

where opt_k is the minimum loss among all k-qubit operations, and $\hat{f}_{\mathbf{s}}$ is the estimation of $f_{\mathbf{s}}$.

This theorem implies an interesting connection between the QPAC learnability of a predictor and its Pauli decomposition. Moreover, it implies that the square loss is a suitable loss function for estimating the Pauli operators. We note that the factor 2^d is not problematic as it appears simply because of the way the Pauli coefficients are defined.

Proof Sketch of Theorem 3: The proof of this theorem is involved. Here, we only explain the sketch of its proof by presenting the following key lemmas with their proof in Appendix A. The first lemma characterizes opt_k and the second lemma is the key connection to Pauli estimations in our analysis.

Lemma 4. *If opt_k is the minimum loss among the class of all k-qubit measurements, then*

$$\mathsf{opt}_k = \frac{1}{2} - \frac{1}{2}\max_{\mathcal{J}\subset[d]:|\mathcal{J}|=k}\left\|\mathrm{F}_Y^{\mathcal{J}}\right\|_1,$$

where $\|\cdot\|_1$ is the trace norm.

Lemma 5. *Let $\mathcal{M}_{\mathcal{J}} = \{\hat{\Pi}^{\mathcal{J}}, I - \hat{\Pi}^{\mathcal{J}}\}$ be the measurement with the projection $\hat{\Pi}^{\mathcal{J}}$ given in (11). Then,*

$$L_{0-1}(\mathcal{M}_{\mathcal{J}}) \leq \frac{1}{2}\left(1 - \left\|\mathrm{F}_Y^{\mathcal{J}}\right\|_1\right) + U\left(\sqrt{2^d}\left\|\mathrm{F}_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2\right),$$

where $U(x) = x^3 + \frac{3}{2}x^2 + \frac{3}{2}x$, for all $x \geq 0$.

With these lemmas, Theorem 3 immediately follows. Let $\mathcal{J}^*$ be the coordinate as in Theorem 3. Then, with Lemma 4 and 5, and the fact that $U(x) \leq 4x$ for $x \leq 1$, we have that

$$L_{0-1}(\mathcal{M}_{\mathcal{J}^*}) \leq \mathsf{opt}_k + 4\sqrt{2^d}\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2.$$

With that, we establish the theorem.

3.5 Algorithm and Proof of the Main Theorem

So far, we discussed the estimation of the Pauli coefficients and the construction of the predictor. The estimation process in Section 3.2 in its current form may not be applicable when d is large. It is not clear how to create Γ^{-1} and U_i in Figure 1. In this section, we characterize a closed-form expression for Γ^{-1} and present an implementation of it with a scalable circuit.

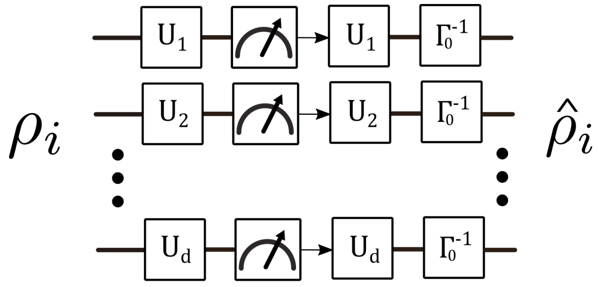

Figure 2: Scalable implementation of the estimation process in Figure 1. For each sample, $U_1, U_2, \cdots, U_d$ are selected randomly and independently from $\{I, H, S^\dagger H\}$. The circuits operate independently on each corresponding qubit implying a $O(d)$ gate complexity.

Consider $\Gamma[\rho_i]$ as in (8) for a single qubit system ($d = 1$). Instead of ranging over all unitary operators, we choose U from the following set with equal probabilities:

$$U \in \{I, H, S^\dagger H\},$$

where H is the Hadamard and $S = \sqrt{\sigma^3}$. With this set, the state is measured either along the computational basis, X-basis, or the Y-basis. Let

$$\Gamma_0[\rho] := \sum_{U\in\{I,H,S^\dagger H\}}\sum_{b\in\{0,1\}}\frac{1}{3}\langle b|U^\dagger\rho U|b\rangle\; U|b\rangle\langle b|U^\dagger.$$

It is not difficult to check that Γ_0 has an inverse and that Lemma 2 still holds as the above set is tomographically complete.

For general d-qubit systems, we apply the single-qubit process to each of the d qubits independently (see Figure 2). In the following, we show that this circuit gives an unbiased estimate of ρ, even though ρ could be an entangled state!

Lemma 6. *Let $\hat{\rho}_i$ be the operation's output in Figure 2. Then, $\mathbb{E}[\hat{\rho}_i] = \rho_i$.*

Proof. Consider the Pauli decomposition of $\rho_i = \sum_{\mathbf{s}}\alpha_s\sigma^{\mathbf{s}}$. As Γ_0^{-1} is a linear mapping, the circuit in Figure 2 is also linear. Let Ψ represent this operation. Then $\hat{\rho}_i = \Psi[\rho_i]$. The linearity implies that $\Psi[\rho_i] = \sum_{\mathbf{s}}\alpha_s\Psi[\sigma^{\mathbf{s}}]$. Since, $\sigma^{\mathbf{s}}$ is in tensor product and Ψ operates on each qubit independently, then $\Psi[\sigma^{\mathbf{s}}] = \otimes_{j=1}^d \Psi_j[\sigma^{s_j}]$, where Ψ_j is the jth wire on Figure 2. As a result,

$$\mathbb{E}[\hat{\rho}_i] = \sum_{\mathbf{s}}\alpha_s\bigotimes_{j=1}^{d}\mathbb{E}\big[\Psi_j[\sigma^{s_j}]\big].$$

Note that

$$\begin{aligned}\mathbb{E}\big[\Psi_j[\sigma^{s_j}]\big] &= \Gamma_0^{-1}\big[\mathbb{E}_{U_i,b_i}[U_i|b_i\rangle\langle b_i|U_i^\dagger]\big]\\ &= \Gamma_0^{-1}\Big[\sum_{U\in\{I,H,S^\dagger H\}}\sum_{b\in\{0,1\}}\frac{1}{3}\langle b|U^\dagger\sigma^{s_j}U|b\rangle\; U|b\rangle\langle b|U^\dagger]\Big]\\ &= \Gamma_0^{-1}[\Gamma[\sigma^{s_j}]] = \sigma^{s_j}.\end{aligned}$$

Algorithm 1: Algorithm for k-qubit Circuits

```
Input: $k \leq d$, and $n$ samples $(\rho_i, y_i)_{i=1}^n$.
Output: Predictor $\hat{\mathcal{M}}$
LearningAlgorithm:
    PauliEstimation:
        for $i = 1 : n$ do
            Choose $d$ unitary $U_j$ randomly from
              $\{I, H, S^\dagger H\}$.
            Apply the circuit in Figure 2 with the
              selected unitary operators.
        for $\mathbf{s} \in \{0,1,2,3\}^d$ with $|\mathrm{supp}(\mathbf{s})| \leq k$ do
            Compute $\hat{f}_{\mathbf{s}}$ using (9).
    for $\mathcal{J} \subset [d]$ with $|\mathcal{J}| = k$ do
        Compute $\hat{F}_Y^{\mathcal{J}} = \sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s}) \subseteq \mathcal{J}} \hat{f}_{\mathbf{s}} \sigma^{\mathbf{s}}$.
        Find $\hat{\mathcal{J}}$ that maximizes $\|\hat{F}_Y^{\mathcal{J}}\|_1$.
    Construct $\hat{\Pi}^{\mathcal{J}}$ as in (11) with $\mathcal{J} = \hat{\mathcal{J}}$.
    return POVM $\hat{\mathcal{M}} := \{\hat{\Pi}^{\hat{J}}, I - \hat{\Pi}^{\hat{J}}\}$.
```

As a result of this equation,

$$\mathbb{E}[\hat{\rho}_i] = \sum_{\mathbf{s}} \alpha_s \bigotimes_{j=1}^{d} \sigma^{s_j} = \rho_i.$$

Hence the proof is complete. □

It follows from large deviation analysis that we get the same error bound as in (10) with the circuit in Figure 2. Therefore, we obtain a scalable estimation circuit consisting of independent single-qubit quantum operations, resulting in a $O(d)$ gate complexity. With that in mind, we summarize our design and present Algorithm 1. It remains to complete the proof of Theorem 1.

Proof of Theorem 1: We show that $\hat{\mathcal{M}}$, the output of Algorithm 1 achieves the optimal loss opt_k. We use Theorem 3 followed by a Parseval-type identity. From (2), it follows that $\|A\|_2^2 = \mathrm{tr}\{A^\dagger A\} = 2^d \sum_{\mathbf{s}} |a_{\mathbf{s}}|^2$. Then,

$$L_{0-1}(\hat{\mathcal{M}}_{\mathcal{J}^*}) \leq \mathsf{opt}_k + 4\, 2^d \sqrt{\sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s}) \subseteq \mathcal{J}^*} (f_{\mathbf{s}} - \hat{f}_{\mathbf{s}})^2}.$$

Hence, with Theorem 2, we have that

$$\begin{aligned} L_{0-1}(\mathcal{M}_{\mathcal{J}^*}) &\leq \mathsf{opt}_k + 4\sqrt{\sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s}) \subseteq \mathcal{J}^*} \mathcal{O}\left(\frac{1}{n} \log\left(\frac{K}{\delta}\right)\right)} \\ &= \mathsf{opt}_k + \mathcal{O}\left(\sqrt{\frac{4^k}{n} \log\left(\frac{K}{\delta}\right)}\right), \end{aligned} \qquad (13)$$

where we used the fact that $|\mathcal{J}^*| = k$. Note that $\mathcal{J}^*$ is unknown as it is defined based on the true operator $F_Y^{\mathcal{J}}$. We need to show that $\hat{\mathcal{J}}$ in Algorithm 1 is "close" to $\mathcal{J}^*$. From Lemma 4, it suffices to show that $\|\hat{F}_Y^{\hat{\mathcal{J}}}\|_1$ is close to $\|{F_Y}^{\mathcal{J}^*}\|_1$ which gives opt_k. Since $\hat{\mathcal{J}}$ maximizes $\|\hat{F}_Y^{\mathcal{J}}\|_1$, then $\|\hat{F}_Y^{\hat{\mathcal{J}}}\|_1 \geq \|\hat{F}_Y^{\mathcal{J}^*}\|_1$. From the triangle inequality and the relation $\|\cdot\|_1 \leq \sqrt{dim}\|\cdot\|_2$, we have that

$$\left| \|\hat{F}_Y^{\mathcal{J}^*}\|_1 - \|{F_Y}^{\mathcal{J}^*}\|_1 \right| \leq \sqrt{2^d} \|\hat{F}_Y^{\mathcal{J}^*} - {F_Y}^{\mathcal{J}^*}\|_2 \leq \epsilon_n,$$

where ϵ_n is the second term in (13). The last inequality follows from Theorem 2. Combining this inequality with Lemma 5 and 4, we obtain that

$$L_{0-1}(\hat{\mathcal{M}}) \leq L_{0-1}(\mathcal{M}_{\mathcal{J}^*}) + \epsilon_n = \mathsf{opt}_k + 2\epsilon_n,$$

where $\hat{\mathcal{M}}$ is the output of Algorithm 1. The proof is complete by replacing the expression for ϵ_n in given (13) and that of K in (7).

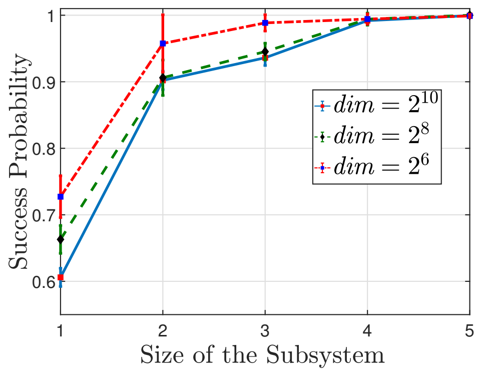

Figure 3: Success probability vs. the size of the sub-system (k) for classifying maximally entangled and separable d-qubit states with different values of d.

4 Numerical Validation

We test Algorithm 1 for classifying separable states from maximally entangled. For that, we generate a training data set by randomly generating d-qubit states. We generate two types of states: separable (with label $y = 0$) and maximally entangled (with label $y = 1$). For that, we use *RandomDensityMatrix* in (Johnston et al., 2016) to generate a separable random density matrix based on a Haar measure.

Figure 3 shows the success probability versus the size of the sub-system (k) with different dimensions (2^d). Our results indicate that accessing only a small subsystem is sufficient to obtain a reasonable accuracy. For instance, a success probability of 0.95 is possible using a 3-qubit subsystem inside the original 10-qubit system.

Moreover, we tested the accuracy of Algorithm 1 with different values of k versus various sample sizes ($n = 10^3$, 10^4, and 10^5). The dataset is the same as in the previous experiment but with a fixed dimension $dim = 2^6$. Figure

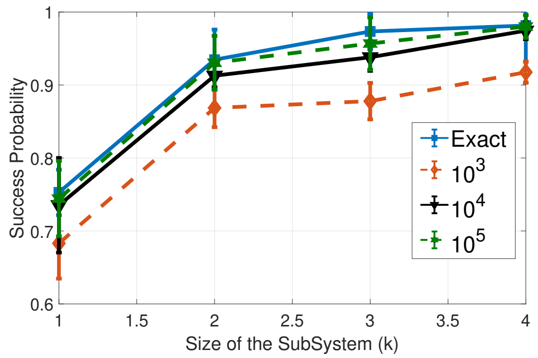

Figure 4: Success probability of Algorithm 1 for various k (subsystem size) and various samples $n = 10^3, 10^4, 10^5$ compared to theoretical bound (exact). The experiment is averaged over 5 runs with error bars showing the deviations.

4 demonstrates success probability as a function of k for various sample sizes. As observed, with more samples, the success probability converges to the theoretical values with exact computations.

Discussion and Future Directions

We prove that the quantum sample complexity of k-qubit quantum operations is $O(\frac{k4^k}{\epsilon^2} \log d)$ which grows logarithmically with the number of qubits and is comparable with the classical sample complexity of k-juntas. This is a surprising result due to the no-cloning principle, measurement incompatibility, richness of k-qubit operations, and the fact that QPAC is a stronger condition than classical PAC. We propose a quantum algorithm that provably QPAC learns k-qubit operations. We develop a new connection to Pauli decomposition with a new estimation method with an scalable circuit.

Our results indicate that shallow-width quantum circuits are learnable with a sample complexity growing logarithmically with the number of qubits (d). As future work, one can study learning of shallow-depth quantum circuits and compare the quantum sample complexity of such circuits with constant-depth classical circuits. Whether quantum sample complexity is comparable with the classical one is an important direction to pursue.

Acknowledgments

This work was partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and also by NSF Grants CCF-2006440, CCF-2007238, CCF-2211423, and Google Research Award.

References

S. Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, sep 2007. doi: 10.1098/rspa.2007.0113.

S. Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 325–338, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188802.

S. Arunachalam and R. de Wolf. A survey of quantum learning theory. *arXiv:1701.06806*, 2017.

S. Arunachalam and R. De Wolf. Optimal quantum sample complexity of learning algorithms. *J. Mach. Learn. Res.*, 19(1):2879–2878, Jan. 2018. ISSN 1532-4435.

K. M. R. Audenaert, M. Nussbaum, A. Szkoła, and F. Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, feb 2008. doi: 10.1007/s00220-008-0417-5.

C. Badescu, R. O'Donnell, and J. Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, jun 2019. doi: 10.1145/3313276.3316344.

S. M. Barnett and S. Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238, feb 2009. doi: 10.1364/aop.1.000238.

B. Bauer, S. Bravyi, M. Motta, and G. K.-L. Chan. Quantum algorithms for quantum chemistry and quantum materials science. *Chemical Reviews*, 120(22):12685–12717, oct 2020. doi: 10.1021/acs.chemrev.9b00829.

E. Bernstein and U. Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, oct 1997. doi: 10.1137/s0097539796300921.

J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd. Quantum machine learning. 549 (7671):195–202, sep 2017. doi: 10.1038/nature23474.

P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst. Machine learning quantum phases of matter beyond the fermion sign problem. 7(1), aug 2017. doi: 10.1038/s41598-017-09098-0.

M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. H. Yoo, S. V. Isakov, P. Massey, R. Halavati, M. Y. Niu, A. Zlokapa, E. Peters, O. Lockwood, A. Skolik, S. Jerbi, V. Dunjko, M. Leib, M. Streif, D. V. Dollen, H. Chen, S. Cao, R. Wiersema, H.-Y. Huang, J. R. McClean, R. Babbush, S. Boixo, D. Bacon, A. K. Ho, H. Neven, and M. Mohseni. Tensorflow quantum: A software framework for quantum machine learning. *arXiv:2003.02989*, Mar. 2020.

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, 1998.

S. Bubeck, S. Chen, and J. Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, nov 2020. doi: 10.1109/focs46700.2020.00070.

Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik. Quantum chemistry in the age of quantum computing. *Chemical Reviews*, 119(19):10856–10915, aug 2019. doi: 10.1021/acs.chemrev.8b00803.

G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. 355(6325):602–606, feb 2017. doi: 10.1126/science.aag2302.

J. Carrasquilla and R. G. Melko. Machine learning phases of matter. 13(5):431–434, feb 2017. doi: 10.1038/nphys4035.

C. Chen, C. Ren, H. Lin, and H. Lu. Entanglement structure detection via machine learning. *Quantum Science and Technology*, 2021.

H.-C. Cheng, M.-H. Hsieh, and P.-C. Yeh. The learnability of unknown quantum measurements. *QIC, Vol. 16, No. 7-8, 0615-0656 (2016)*, Jan. 2015.

O. Crawford, B. van Straaten, D. Wang, T. Parks, E. Campbell, and S. Brierley. Efficient quantum measurement of pauli operators in the presence of finite sampling error. *arXiv:1908.06942*, 2020.

D.-L. Deng, X. Li, and S. D. Sarma. Quantum entanglement in neural network states. *Physical Review X*, 7(2): 021021, 2017.

S. Gambs. Quantum classification. *0809.0444 [quant-ph]*, Sept. 2008.

V. Giovannetti, S. Lloyd, and L. Maccone. Quantum random access memory. *Physical Review Letters*, 100(16), apr 2008. doi: 10.1103/physrevlett.100.160501.

M. Guta and W. Kotlowski. Quantum learning: asymptotically optimal classification of qubit states. *New Journal of Physics*, 12(12):123032, dec 2010. doi: 10.1088/1367-2630/12/12/123032.

J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. Sample-optimal tomography of quantum states. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016. doi: 10.1145/2897518.2897585.

M. Heidari, A. Padakandla, and W. Szpankowski. A theoretical framework for learning from quantum data. In *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, jul 2021. doi: 10.1109/isit45174.2021.9517721.

M. Heidari, A. Grama, and W. Szpankowski. Toward physically realizable quantum neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6): 6902–6909, jun 2022. doi: 10.1609/aaai.v36i6.20647.

C. Hempel, C. Maier, J. Romero, J. McClean, T. Monz, H. Shen, P. Jurcevic, B. P. Lanyon, P. Love, R. Babbush, A. Aspuru-Guzik, R. Blatt, and C. F. Roos. Quantum chemistry calculations on a trapped-ion quantum simulator. *Physical Review X*, 8(3):031022, jul 2018. doi: 10.1103/physrevx.8.031022.

B. C. Hiesmayr. Free versus bound entanglement, a NP-hard problem tackled by machine learning. *Scientific Reports*, 11(1), oct 2021. doi: 10.1038/s41598-021-98523-6.

H.-Y. Huang, R. Kueng, and J. Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics 16, 1050–1057 (2020)*, Feb. 2020. doi: 10.1038/s41567-020-0932-7.

N. Johnston, A. Cosentino, and V. Russo. Qetlab: Qetlab v0.9, 2016.

V. Kanade, A. Rocchetto, and S. Severini. Learning dnfs under product distributions via μ-biased quantum fourier sampling. *arXiv:1802.05690v3*, 2019.

I. Kassal, J. D. Whitfield, A. Perdomo-Ortiz, M.-H. Yung, and A. Aspuru-Guzik. Simulating chemistry using quantum computers. *Annual Review of Physical Chemistry*, 62(1):185–207, may 2011. doi: 10.1146/annurev-physchem-032210-103512.

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3): 115–141, 1994. doi: 10.1007/bf00993468.

S. Lloyd, M. Mohseni, and P. Rebentrost. Quantum algorithms for supervised and unsupervised machine learning. *arXiv:1307.0411*, 2013.

S. Lloyd, M. Mohseni, and P. Rebentrost. Quantum principal component analysis. 10(9):631–633, jul 2014. doi: 10.1038/nphys3029.

S. Lu, S. Huang, K. Li, J. Li, J. Chen, D. Lu, Z. Ji, Y. Shen, D. Zhou, and B. Zeng. Separability-entanglement classifier via machine learning. *Physical Review A*, 98(1): 012315, 2018.

Y.-C. Ma and M.-H. Yung. Transforming bell's inequalities into state classifiers with machine learning. *npj Quantum Information*, 4(1), jul 2018. doi: 10.1038/s41534-018-0081-3.

F. V. Massoli, L. Vadicamo, G. Amato, and F. Falchi. A leap among entanglement and neural networks: A quantum survey. *arXiv:2107.03313*, July 2021.

S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1):015003, mar 2020. doi: 10.1103/revmodphys.92.015003.

A. Montanaro and R. de Wolf. A survey of quantum property testing. *Theory of Computing*, 1(1):1–81, 2016. doi: 10.4086/toc.gs.2016.007.

A. Montanaro and T. J. Osborne. Quantum boolean functions. *arXiv:0810.2435*, 2010.

E. Mossel, R. O'Donnell, and R. P. Servedio. Learning juntas. In *Proc. ACM Symp. on Theory of Computing*, pages 206–212, 2003.

E. Mossel, R. O'Donnell, and R. A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci*, 69(3):421–434, 2004.

R. O'Donnell and J. Wright. Efficient quantum tomography. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016. doi: 10.1145/2897518.2897544.

R. O'Donnell and J. Wright. Efficient quantum tomography II. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, jun 2017. doi: 10.1145/3055399.3055454.

D. K. Park, F. Petruccione, and J.-K. K. Rhee. Circuit-based quantum random access memory for classical data. *Scientific Reports*, 9(1), mar 2019. doi: 10.1038/s41598-019-40439-3.

A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1), jul 2014. doi: 10.1038/ncomms5213.

P. Rebentrost, M. Mohseni, and S. Lloyd. Quantum support vector machine for big data classification. *Physical Review Letters*, 113(13), sep 2014. doi: 10.1103/physrevlett.113.130503.

M. Schuld, I. Sinayskiy, and F. Petruccione. The quest for a quantum neural network. *Quantum Information Processing*, 13(11):2567–2586, Aug 2014. ISSN 1573-1332. doi: 10.1007/s11128-014-0809-8.

M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, mar 2020. doi: 10.1103/physreva.101.032308.

R. A. Servedio and S. J. Gortler. Equivalences and separations between quantum and classical learnability. *SIAM J. Comput.*, 33(5):1067–1092, May 2004. ISSN 0097-5397. doi: 10.1137/S0097539704412910. URL https://doi.org/10.1137/S0097539704412910.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, nov 1984. doi: 10.1145/1968.1972.

A Proof of The Technical Lemmas

A.1 Proof of Lemma 4

Proof. We start with proving a lower bound on opt_k. Fix a k element coordinate subset $\mathcal{I} \subset [d]$ and consider a k-qubit measurement that depends only on coordinates $\mathcal{I}$. From Lemma 1, the loss of $\mathcal{M}$ equals to

$$L_{0-1}(\mathcal{M}) = \frac{1}{2} - 2^{d-1} \sum_{\mathbf{s}} f_{\mathbf{s}} g_{\mathbf{s}}.$$

From Definition 2, as $\mathcal{M}$ depends only on coordinate $\mathcal{I}$, then one can show that $g_{\mathbf{s}} = 0$ for all $\mathbf{s}$ with $\mathrm{supp}(\mathbf{s}) \nsubseteq \mathcal{I}$. Therefore,

$$L_{0-1}(\mathcal{M}) = \frac{1}{2} - 2^{d-1} \sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s})\subseteq \mathcal{I}} f_{\mathbf{s}} g_{\mathbf{s}}.$$

Define the following operator:

$$F_Y^{\mathcal{I}} := \sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s})\subseteq \mathcal{I}} f_{\mathbf{s}} \sigma^{\mathbf{s}}. \tag{14}$$

Therefore, from (2), we have that

$$L_{0-1}(\mathcal{M}) = \frac{1}{2} - \frac{1}{2}\mathrm{tr}\{F_Y^{\mathcal{I}} G\} \geq \frac{1}{2} - \frac{1}{2}\mathrm{tr}\{|F_Y^{\mathcal{I}}||G|\} = \frac{1}{2} - \frac{1}{2}\mathrm{tr}\{|F_Y^{\mathcal{I}}|\},$$

where $|A| = \sqrt{A^\dagger A}$ and the first inequality follows as $\mathrm{tr}\{A\} \leq \mathrm{tr}\{|A|\}$. The second equality follows as the eigenvalues of G belong to $\{-1, 1\}$, implying that $|G| = I_d$. Therefore, we obtain that

$$L_{0-1}(\mathcal{M}) \geq \frac{1}{2} - \frac{1}{2}\left\|F_Y^{\mathcal{I}}\right\|_1 \geq \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} \left\|F_Y^{\mathcal{J}}\right\|_1,$$

where the last inequality holds by minimizing the lower bound over all k-element coordinates $\mathcal{J}$. Note that the above bound holds for all $\mathcal{M}$ depending on any k-element coordinate subset $\mathcal{I}$. Thus, we obtain the lower bound on opt_k:

$$\mathsf{opt}_k \geq \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} \left\|F_Y^{\mathcal{J}}\right\|_1. \tag{15}$$

Next, we establish the achievability of the lower bound. Again fix a k-element subset $\mathcal{J} \subset [d]$ and let $G_{M_{\mathcal{J}}} = \mathrm{sign}[F_Y^{\mathcal{J}}]$. Note that we can consider a valid measurement $\mathcal{M}_{\mathcal{J}}$ corresponding to $G_{M_{\mathcal{J}}}$. Moreover, $G_{M_{\mathcal{J}}}$ is a k-junta operator depending only on the coordinates $\mathcal{J}$. Therefore, its Fourier coefficients $g_{\mathbf{s}}$ are zero for any $\mathbf{s}$ with $\mathrm{supp}(\mathbf{s}) \nsubseteq \mathcal{J}$. As a result, from (14)

$$\begin{aligned} L_{0-1}(\mathcal{M}_{\mathcal{J}}) &= \frac{1}{2} - \frac{1}{2}\mathrm{tr}\{G_{M_{\mathcal{J}}} F_Y^{\mathcal{J}}\} \\ &= \frac{1}{2} - \frac{1}{2}\mathrm{tr}\{\mathrm{sign}[F_Y^{\mathcal{J}}] F_Y^{\mathcal{J}}\} \\ &= \frac{1}{2} - \frac{1}{2}\left\|F_Y^{\mathcal{J}}\right\|_1, \end{aligned}$$

where the last equality follows from the identity $\left\|A\right\|_1 = \mathrm{tr}\{A\,\mathrm{sign}[A]\}$ that holds for any Hermitian and bounded operator A. With the above inequality, optimizing over $\mathcal{J}$ gives

$$\min_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} L_{0-1}(\mathcal{M}_{\mathcal{J}}) = \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} \left\|F_Y^{\mathcal{J}}\right\|_1. \tag{16}$$

Note that opt_k is smaller than the left-hand side of (16), as it has an additional minimization over the choice of the measurements:

$$\mathsf{opt}_k = \min_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} \inf_{\mathcal{M}_{\mathcal{J}}} L_{0-1}(\mathcal{M}_{\mathcal{J}}).$$

Therefore, (16) is an upper bound for $\mathsf{opt}(k)$. As this upper bound matches with the lower bound in (15), then we obtain the equality in (16). This proves the expression for $\mathsf{opt}(k)$ and that $\mathcal{M}_{\mathcal{J}^*}$ is the best k-junta measurement.

□

A.2 Proof of Lemma 5

Let $G_{\mathcal{J}} = M_1 - M_0 = I - 2\hat{\Pi}^{\mathcal{J}}$. It is not difficult to check that $G_{\mathcal{J}} = \text{sign}[\hat{F}_Y^{\mathcal{J}}]$. From Lemma 1 in the main text, the loss of $M_{\mathcal{J}}$ can be written as $L_{0-1}(\mathcal{M}_{\mathcal{J}}) = \frac{1}{2} - 2^{d-1}\sum_{\mathbf{s}} f_{\mathbf{s}} g_{\mathbf{s}}$, where $f_{\mathbf{s}}$ and $g_{\mathbf{s}}$ are the Pauli coefficients of F_Y and $G_{\mathcal{J}}$. Note that $G_{\mathcal{J}}$ depends only on the coordinates of $\mathcal{J}$. More precisely, $G_{\mathcal{J}} = (I^{\mathcal{J}} - 2\hat{\Pi}^{\mathcal{J}}) \otimes I^{\mathcal{J}^c}$, where $I^{\mathcal{J}}$ and $I^{\mathcal{J}^c}$ are the identity operators on the corresponding systems. Hence, the Pauli coefficients of $g_{\mathbf{s}}$ of $G_{\mathcal{J}}$ are zero outside of $\mathcal{J}$. Therefore,

$$L_{0-1}(\mathcal{M}_{\mathcal{J}}) = \frac{1}{2} - 2^{d-1} \sum_{\mathbf{s}:\text{supp}(\mathbf{s})\subseteq\mathcal{J}} f_{\mathbf{s}} g_{\mathbf{s}}$$

Define the 2-norm of an operator A as $\|A\|_2 := \sqrt{\text{tr}\{A^\dagger A\}}$. Then, from (2), $\|A\|_2^2 = 2^d \sum_{\mathbf{s}} |a_{\mathbf{s}}|^2$. Moreover, for any pair of Hermitian operators A, B we have the identity $\|A - B\|_2^2 = \|A\|_2^2 + \|B\|_2^2 - 2\,\text{tr}\{AB\}$. Therefore, from (2), we obtain that

$$\sum_{\mathbf{s}} a_{\mathbf{s}} b_{\mathbf{s}} = 2^{-d}\,\text{tr}\{AB\} = 2^{-d-1}\Big(\|A\|_2^2 + \|B\|_2^2 - \|A - B\|_2^2\Big).$$

Therefore, from the definition of $F_Y^{\mathcal{J}}$, we have that

$$\sum_{\mathbf{s}:\text{supp}(\mathbf{s})\subseteq\mathcal{J}} f_{\mathbf{s}} g_{\mathbf{s}} = 2^{-d}\,\text{tr}\big\{F_Y^{\mathcal{J}} G_{\mathcal{J}}\big\}.$$

As a result, the loss of $\mathcal{M}_{\mathcal{J}}$ can be written as

$$\begin{aligned} L_{0-1}(\mathcal{M}_{\mathcal{J}}) &= \frac{1}{2} - \frac{1}{4}\Big(\|G_{\mathcal{J}}\|_2^2 + \|F_Y^{\mathcal{J}}\|_2^2 - \|F_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2^2\Big) \\ &= \frac{1}{4}\big(2 - 2^d - \|F_Y^{\mathcal{J}}\|_2^2 + \|F_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2^2\big), \end{aligned} \tag{17}$$

where we used the fact that $\|G_{\mathcal{J}}\|_2^2 = 2^d$ as the eigenvalues of $G_{\mathcal{J}}$ belong to $\{-1, 1\}$.

Next, we bound the last 2-norm quantity above. Recall that $\hat{F}_Y^{\mathcal{J}} := \sum_{\mathbf{s}:\text{supp}(\mathbf{s})\subseteq\mathcal{J}} \hat{f}_{\mathbf{s}} \sigma^{\mathbf{s}}$ is an approximation of $F_Y^{\mathcal{J}}$ using the estimated Pauli coefficients. By adding and subtracting $\hat{F}_Y^{\mathcal{J}}$, we have that

$$\begin{aligned} \|F_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2^2 &\overset{(a)}{\leq} \Big(\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2 + \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2\Big)^2, \\ &= \Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^2 + \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2^2 + 2\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2 \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2, \end{aligned} \tag{18}$$

where (a) follows from the Minkowski's Inequality inequality for 2-norm. Note that $G_{\mathcal{J}} = \text{sign}[\hat{F}_Y^{\mathcal{J}}]$. Moreover, note that for any function h the identity $|h - \text{sign}[h]| = |1 - |h||$ holds. Therefore,

$$\begin{aligned} \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2^2 &= \Big\|I_d - |\hat{F}_Y^{\mathcal{J}}|\Big\|_2^2 = \|I_d\|_2^2 + \Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_2^2 - 2\Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_1 \\ &= 2^d + \Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_2^2 - 2\Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_1. \end{aligned} \tag{19}$$

From this relation and equations (17), (18), we obtain the following upper bound

$$\begin{aligned} 4L_{0-1}(\mathcal{M}_{\mathcal{J}}) &\leq 2 - 2^d - \|F_Y^{\mathcal{J}}\|_2^2 + \Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^2 + \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2^2 + 2\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2 \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2 \\ &= 2 - 2^d - \|F_Y^{\mathcal{J}}\|_2^2 + \Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^2 + 2^d + \Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_2^2 - 2\Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_1 + 2\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2 \Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2 \\ &= 2 - 2\Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_1 + \underbrace{\Big\|\hat{F}_Y^{\mathcal{J}}\Big\|_2^2 - \|F_Y^{\mathcal{J}}\|_2^2}_{\text{(I)}} + \Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^2 + 2\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2 \underbrace{\Big\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\Big\|_2}_{\text{(II)}}. \end{aligned} \tag{20}$$

In what follows, we bound the terms denoted by (I) and (II).

Bounding (I): From the Minkowski's inequality for 2-norm, we have

$$\begin{aligned}
\left\|\hat{F}_Y^{\mathcal{J}}\right\|_2^2 &\leq \left(\|F_Y^{\mathcal{J}}\|_2 + \left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2\right)^2 \\
&= \|F_Y^{\mathcal{J}}\|_2^2 + \left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2^2 + 2\|F_Y^{\mathcal{J}}\|_2\left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2 \\
&\leq \|F_Y^{\mathcal{J}}\|_2^2 + \left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2^2 + 2\left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2,
\end{aligned}$$

where the second inequality is due Bessel's inequality and the following chain of inequalities

$$\|F_Y^{\mathcal{J}}\|_2 \leq \|F_Y\|_2 = \left\|\rho_{XY}(I_d \otimes \sigma^3)\right\|_2 = \|\rho_{XY}\|_2 \leq \|\rho_{XY}\|_1 = 1,$$

where we used the fact that $F_Y = -\sqrt{\rho_{XY}}(I_d \otimes \sigma^3)\sqrt{\rho_{XY}}$ which also equals to $\rho_{XY}(I_d \otimes \sigma^3)$, and that $\|\cdot\|_2 \leq \|\cdot\|_1$. Hence, the term (I) in (20) is upper bounded as

$$\text{(I)} \leq \lambda_1 := \left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2^2 + 2\left\|\hat{F}_Y^{\mathcal{J}} - F_Y^{\mathcal{J}}\right\|_2. \tag{21}$$

Bounding (II): From (19), we have

$$\begin{aligned}
\left\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\right\|_2^2 &= 1 + \left\|\hat{F}_Y^{\mathcal{J}}\right\|_2^2 - 2\left\|\hat{F}_Y^{\mathcal{J}}\right\|_1 \\
&\overset{(a)}{\leq} 1 + 2(\|F_Y^{\mathcal{J}}\|_2^2 + \left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2) - 2\left\|\hat{F}_Y^{\mathcal{J}}\right\|_1 \\
&\overset{(b)}{=} 1 + 2(\|F_Y^{\mathcal{J}}\|_2^2 + \left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2) - 2(\|F_Y^{\mathcal{J}}\|_1 + (\left\|\hat{F}_Y^{\mathcal{J}}\right\|_1 - \|F_Y^{\mathcal{J}}\|_1)) \\
&= 1 + 2(\|F_Y^{\mathcal{J}}\|_2^2 - \|F_Y^{\mathcal{J}}\|_1) + 2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 - 2(\left\|\hat{F}_Y^{\mathcal{J}}\right\|_1 - \|F_Y^{\mathcal{J}}\|_1) \\
&\overset{(c)}{\leq} 1 + 2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 - 2(\left\|\hat{F}_Y^{\mathcal{J}}\right\|_1 - \|F_Y^{\mathcal{J}}\|_1) \\
&\overset{(d)}{\leq} 1 + 2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 + 2\sqrt{2^d}\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2,
\end{aligned} \tag{22}$$

where (a) follows from the Minkowski's inequality for 2-norm and the inequality $(x+y)^2 \leq 2(x^2+y^2)$. Equality (b) follows by adding and subtracting $\|F_Y^{\mathcal{J}}\|_1$. Inequality (c) holds from the inequality $\|\cdot\|_2 \leq \|\cdot\|_1$ and the fact that $\|F_Y^{\mathcal{J}}\|_2 \leq 1$ which implies that $\|F_Y^{\mathcal{J}}\|_2^2 \leq \|F_Y^{\mathcal{J}}\|_2 \leq \|F_Y^{\mathcal{J}}\|_1$. Lastly, inequality (d) holds because of the following chain of inequalities

$$\left|\|F_Y^{\mathcal{J}}\|_1 - \left\|\hat{F}_Y^{\mathcal{J}}\right\|_1\right| \leq \left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_1 \leq \sqrt{2^d}\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2, \tag{23}$$

where the first inequality is due to the Minkowski's inequality for 1-norm and the second inequality is due to the inequality $\|\cdot\|_1 \leq \sqrt{dim}\|\cdot\|_2$ and the fact that $dim = 2^d$.

Next, we show that the quantity $\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2$ without the square is upper bounded by the same term as in the right-hand side of (22). That is

$$\text{(II)} = \|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2 \leq \lambda_2 := 1 + 2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 + 2\sqrt{2^d}\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2. \tag{24}$$

The argument is as follows: if $\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2 \leq 1$, then the upper bound holds trivially as $\lambda_2 \geq 1$; otherwise, if $\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2 > 1$, then quantity is less than its squared, i.e., $\|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2 \leq \|\hat{F}_Y^{\mathcal{J}} - G_{\mathcal{J}}\|_2^2$. In that case, we obtain an upper bound using (22).

As a result of the bounds in (20), (21), and (24) we obtain that

$$\begin{aligned}
4L_{0-1}(\mathcal{M}_{\mathcal{J}}) &\leq 2 - 2\left\|\hat{F}_Y^{\mathcal{J}}\right\|_1 + \lambda_1 + \left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 + 2\lambda_2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2 \\
&= 2 - 2\|F_Y^{\mathcal{J}}\|_1 + 2\left(\|F_Y^{\mathcal{J}}\|_1 - \left\|\hat{F}_Y^{\mathcal{J}}\right\|_1\right) + \lambda_1 + \left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 + 2\lambda_2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2 \\
&\leq 2 - 2\|F_Y^{\mathcal{J}}\|_1 + 2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2 + \lambda_1 + \left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2^2 + 2\lambda_2\left\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\right\|_2,
\end{aligned}$$

where the last inequality is due to (23). Next, from the definition of λ_1 and λ_2, the right hand side of the above inequality simplifies to the following

$$4L_{0-1}(\mathcal{M}_{\mathcal{J}}) \leq 2 - 2\big\|F_Y^{\mathcal{J}}\big\|_1 + 4\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^3 + 2\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^2 + 4\sqrt{2^d}\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2^2 + 6\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2,$$

We, further upper bound the right hand side by replacing $\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2$ with $\sqrt{2^d}\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2$ in the third, fourth and sixth terms above. As a result we get

$$4L_{0-1}(\mathcal{M}_{\mathcal{J}}) \leq 2 - 2\big\|F_Y^{\mathcal{J}}\big\|_1 + 4U(\sqrt{2^d}\Big\|F_Y^{\mathcal{J}} - \hat{F}_Y^{\mathcal{J}}\Big\|_2),$$

where $U(x) = x^3 + \frac{3}{2}x^2 + \frac{3}{2}x$ as in the statement of the lemma. Dividing both sides by 4 completes the proof.