Quantum State Classification via Quantum Fourier

Mohsen Heidari MHEIDARI@PURDUE.EDU
Department of Computer Sciences, Purdue University
West Lafayette, IN, USA

Wojciech Szpankowski SZPAN@PURDUE.EDU
Department of Computer Sciences, Purdue University
West Lafayette, IN, USA

Abstract

We study *learning* from quantum data, in particular quantum state classification which has applications, among others, in classifying the *separability* of quantum states. In this learning model, there are n quantum states with classical labels as the training samples. Predictors are quantum measurements that when applied to the next unseen quantum state predict its classical label. By integrating learning theory with quantum information, we introduce a quantum counterpart of the PAC framework for learning with respect to classes of measurements. We argue that major challenges arising from the quantum nature of the problem are measurement *incompatibility* and the *no-cloning* principle — prohibiting sample reuse. Then, after introducing a Fourier expansion through Pauli's operators, we study learning with respect to an infinite class of quantum measurements whose operator's Fourier spectrum is concentrated on low degree terms. We propose a quantum learning algorithm and show that the quantum sample complexity depends on the "compatibility structure" of such measurement classes – the more compatible the class is, the lower the quantum sample complexity will be. We further introduce k-junta measurements as a special class of low-depth quantum circuits whose Fourier spectrum is concentrated on low degrees.

1. Introduction

Classification of quantum states has been studied under various specifications such as *separability* of quantum states [GQJ+18,MY18], to tell if an unknown quantum state is entangled or separable, and integrated quantum photonics [KBI+20], to classify if a quantum emitter is "single" or "not-single". In this paper, we study this problem and seek to integrate learning theory with quantum information. Following our earlier work [HPS21], we introduce a quantum counterpart of well-studied *probably approximately correct* (PAC) developed by [KSS94, Val84].

The focus of this paper is on the binary classification of qubit states. In this learning model, similar to its classical counterpart, there are a set of quantum states living on the Hilbert space of d qubits. Associated to each state is a classical label $y \in \{0, 1\}$. Let ρ_0 and ρ_1 denote the overall state of the system when the label is 0 and 1, respectively. The density operators of the states and the underlying law governing the labeling are unknown. Available are n labeled quantum samples $(\rho_{y_i}, y_i), i = 1, 2, ..., n$, generated independent and identically distributed (i.i.d.) according to an unknown, but fixed, probability distribution. The predictors in this model are quantum measurements (positive operator-valued measure (POVM)). The objective is to process the training set and select a quantum measurement from a library of choices (concept class) that accurately predicts the label of the next unseen quantum states.

Suppose $\mathcal{M} := \{M_0, M_1\}$ is the selected POVM with outcomes in $\{0, 1\}$. To test $\mathcal{M}$, a new sample $(\rho_{Y_{test}}, Y_{test})$ is drawn. Then without revealing random variable Y_{test}, we measure $\rho_{Y_{test}}$ with $\mathcal{M}$. Conditioned on this sample, the outcome of this predictor is $\hat{y}$ with probability equal to the trace $\mathrm{tr}\{M_{\hat{y}}\rho_{y_{test}}\}, \hat{y} \in \{0, 1\}$. Note that due to the quantum nature of the problem, even though the test sample is fixed, the predictor's outcome is random. This is different from the classical settings, where the output of the predictor is a deterministic function of the samples.

We use the $0 - 1$ loss to measure the accuracy of the predictors. Hence, the predictor's risk is $L_{0-1}(\mathcal{M}) = \mathbb{P}\{Y_{test} \neq \hat{Y}\}$. If $\mathcal{C}$ is a collection of POVM's to choose from, we seek a quantum learning algorithm that selects the most accurate predictor $\mathcal{M} \in \mathcal{C}$. To complete this formulation, we present a quantum counterpart of PAC, called QPAC.

Definition 1.1 (QPAC) *A measurement concept class $\mathcal{C}$ is QPAC learnable, if there exist a function $n_{\mathcal{C}} : (0, 1)^2 \mapsto \mathbb{N}$ and a quantum learning algorithm 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 $\inf_{\mathcal{M}\in\mathcal{C}} L_{0-1}(\mathcal{M}) + \epsilon$.*

We note that QPAC is a natural extension of the classical PAC. It subsumes the classical learning model when the quantum samples are orthogonal and the predictors in $\mathcal{C}$ are functions. We note, however, that some difficulties arise from the quantum nature of the problem. For instance, from the axioms of quantum measurement, the quantum samples are disturbed by the algorithm, irreversibly. Further, due to the *no-cloning* principle, we cannot clone or re-use the samples for other processes. Whereas, in the classical framework, a learning algorithm can process a copy version of the training samples without losing them.

That being said, one might ask the following question:

- *What is the quantum sample complexity of a given measurement concept class $\mathcal{C}$?*

For that, one natural approach for QPAC learning, which we call approximate-then-learn, is via *state tomography*. Here, one first performs *state tomography* to arrive at an approximate description of

the quantum states and then performs classical learning algorithms on the stored density matrices. From the copy complexity bound of [HHJ+16], this approach requires $\mathcal{O}\big(\frac{1}{\epsilon^2}2^{2d}\big)$ identical copies of ρ_0 and ρ_1. Hence, the exponentially large sample complexity of this approach is prohibitive in large dimensional settings.

Toward answering the above question, bounds on the quantum sample complexity of finite classes have been provided in our earlier work [HPS21] and using quantum Empirical Risk Minimization (ERM). In this paper, we investigate the sample complexity of two concept classes containing uncountably many quantum measurements. These classes are called ϵ-*concentrated* and k-*juntas* measurements, where the second class is a special case of the first with $\epsilon = 0$. The classical counterpart of these classes, which are finite, capture several concept classes such as half-spaces, constant-depth boolean circuits, and decision trees [GR06, Val12, BOW10, MOS03, LMN93, KKYS05].

In the classical setting, the Fourier expansion on the Boolean cube has been proved powerful in learning such concept classes [DW08, O'D14]. Inspired by this, we study learning with a quantum Fourier expansion adapted for operators on the space of qubits which is based on Pauli operators. We rigorously show that by applying tools from statistics, operator theory, and quantum information such as operator Chernoff's inequality, and Hilbert-Schmidt inner product, one can get a quantum sample complexity for such measurement classes. We observe that quantum sample complexity depends on the "compatibility structure" of the class — a property unique to quantum measurements. The more compatible the class is, the lower the quantum sample complexity will be. To the best of our knowledge, this is the first of such result that obtains a non-trivial sample complexity for infinite quantum concept classes. We believe our work opens the doors to further study of theoretical quantum learning algorithms with practical application.

1.1 Overview of the Main Results

To establish our results, we take an operator perspective. We first associate a joint feature-label quantum state for each sample. This is done by adding an auxiliary qubit representing the label. With that, the classical label 0 and 1 are identified with the qubit $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$, respectively. Further, we define an average density operator for the samples, that is $\rho_{XY} = p_0\rho_0 \otimes |0\rangle\langle 0| + p_1\rho_1 \otimes |1\rangle\langle 1|$. In order to analyze the prediction loss of measurements, we prove the following result.

Theorem 3.1 (Abbreviated) *There is an operator* F_Y*, such that the expected loss of any measurement* $\mathcal{M} := \{M_0, M_1\}$ *equals*

$$L_{0\text{-}1}(\mathcal{M}) = \frac{1}{2} - \frac{1}{2}\operatorname{tr}\Big\{\mathrm{G}_M F_Y \rho_{XY}\Big\}, \tag{1}$$

where $\mathrm{G}_M = M_1 - M_0$.

We call F_Y and G_M the labeling operator and prediction operator, respectively. The above result resembles the a similar equation in classical setting. If U and V are two binary random variables taking values from $\{-1, 1\}$, then $\mathbb{P}\{U \neq V\} = \frac{1}{2} - \frac{1}{2}\mathbb{E}[UV]$. Therefore, the trace quantity in (1) can be viewed as a operator analog of correlation between random variables. With this result and the copy complexity bound of [HHJ+16] in state tomography, we immediately conclude that at most $\mathcal{O}\big(2^{2d}/\epsilon^2\big)$ samples are needed to learn a measurement concept class $\mathcal{C}$ on the space of d-qubits.

Next, we built upon the *quantum Fourier expansion*, define our concept classes, and derive our sample complexity results. It is well-known that Pauli operators together with the identity form an orthonormal basis for the Hilbert-Schmidt operators on one qubit. Lets denote such operators by σ^s, $s \in \{0,1,2,3\}$. As discussed in [MO10], tensor products of these operators form an orthonormal basis for the space of operators on d qubits, the Hilbert space of $\dim = 2^d$. In other words, any bounded operator A on this Hilbert space is uniquely decomposed as a linear combination of tensor operators $\sigma^{\mathbf{s}} = \otimes_{\ell=1}^{d} \sigma^{s_\ell}, \mathbf{s} \in \{0,1,2,3\}^d$. The coefficients of this decomposition, denoted by $a_{\mathbf{s}}, \mathbf{s} \in \{0,1,2,3\}^d$, are called the Fourier coefficients of A (for more details see Section 2.2).

Concentrated measurement classes: Based on the Fourier coefficients, we can associate a "power spectrum" in the Fourier domain. We are interested in operators whose Fourier spectrum is dominated by a smalls set of coefficients, say $a_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ for a subset $\mathcal{A} \subseteq \{0,1,2,3\}^d$. More precisely, $\sum_{\mathbf{s} \notin \mathcal{A}} a_{\mathbf{s}}^2 \leq \epsilon$ for a small parameter $\epsilon \in (0,1)$. That said, a concept class is said to be *concentrated* around a set $\mathcal{A}$, if all predictors in the class have almost all the "power spectrum" on $\mathcal{A}$ (see Definition 2.2).

Next, we present our results on the learnability of measurements *concentrated* around a degree set $\mathcal{A}$. We assume that the X-system in ρ_{XY} is maximally mixed, i.e., $\mathrm{tr}_Y\{\rho_{XY}\} = I_d$. In classical settings, this condition translates to the uniform distribution on the inputs as in [LMN93]. However, here, the probabilities p_0, p_1 can be non-uniform, as long as the average state $\mathrm{tr}_Y\{\rho_{XY}\}$ is maximally mixed. Let $\mathsf{opt}_{\mathcal{C}} := \inf_{\mathcal{M} \in \mathcal{C}} L_{0-1}(\mathcal{M})$. With that, our second result is presented in the following.

Theorem 3.5 *Let $\mathcal{C}$ be a concept class that is ϵ-concentrated around a degree set $\mathcal{A}$, where $\epsilon \in [0, 1]$. Then, for any $\varepsilon, \delta \in (0,1)$, under the maximally mixed X-system, there exists a quantum learning algorithm that with probability $(1-\delta)$ learns $\mathcal{C}$ with error up to $2\mathsf{opt}_{\mathcal{C}} + 2\epsilon + \varepsilon$, provided that $n > \frac{1}{\varepsilon^2} S(\mathcal{A}, \delta)$ samples are available, where*

$$S(\mathcal{A}, \delta) = \mathcal{O}\left(\min_{\mathcal{B}_j \text{ covering of } \mathcal{A}} \sum_{j=1}^{m} \sqrt{|\mathcal{B}_j| \log \frac{|\mathcal{B}_j|}{\delta}} \right)^2.$$

Here, $S(\mathcal{A}, \delta)$ is a quantity incorporating the compatibility structure of the class and is characterized by finding "compatibility covers" of $\mathcal{A}$. A set $\mathcal{B} \subseteq \{0,1,2,3\}^d$ is said to be compatible, if all $\sigma^{\mathbf{s}}, \mathbf{s} \in \mathcal{B}$ are *commuting* with each other. A compatibility cover for $\mathcal{A}$ is a collection of compatible subset $\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_m$ such that $\mathcal{A} = \bigcup_j \mathcal{B}_j$. Therefore, if the class is fully compatible then $S(\mathcal{A}, \delta) = \mathcal{O}\big(|\mathcal{A}| \log \frac{|\mathcal{A}|}{\delta}\big)$. On the other hand, if $\mathcal{C}$ is fully incompatible, then $S(\mathcal{A}, \delta) = \mathcal{O}\big(|\mathcal{A}|^2 \log \frac{1}{\delta}\big)$.

The classical version of these classes containing Boolean functions with concentrated Boolean Fourier expansions has been studied in the literature. The well-known low-degree algorithm [LMN93] has been used to learn such classes under the uniform distribution on inputs. Using this algorithm, any classical concentrated concept $\mathcal{C}_{\text{cl}}$ is learnable with error up to $8\mathsf{opt}_{\mathcal{C}_{\text{cl}}}$ and with sample complexity $\mathcal{O}\big(|\mathcal{A}| \log \frac{|\mathcal{A}|}{\delta}\big)$, where $\mathcal{A}$ is the degree set of the class. It is not difficult to see that Theorem 3.5 subsumes the learnability of classical concentrated concepts.

We prove Theorem 3.5 by proposing quantum low-degree algorithm (QLD), with running time of $\mathcal{O}\big(n4^{|\mathcal{A}|}\big)$ (see Algorithm 1). The heart of our learning algorithm is a method for estimating quantum Fourier coefficients from the training samples. However, this estimation process in our quantum setting is not straightforward as the observables may not be compatible.

Measurement incompatibility: Since our interactions with quantum samples are through quantum measurements, we face the issue of measurement *incomatibility* (or joint measurability). Measurement compatibility concerns whether there exists a fine-grained "reference" measurement which can be measured to determine the outcomes of the individual measurements [Hol12]. When the measurements are incompatible, the uncertainty principle limits the accuracy of the joint observations. Each Fourier coefficient is modeled by a *quantum observable* and in general, such observables might be *incompatible*. In Section 3.2.2, we address this issue by proposing a compatibility covering. The idea is to "cover" $\mathcal{A}$ with subsets $\mathcal{B}_j \subseteq \mathcal{A}$ such that $\mathcal{A} = \bigcup_j \mathcal{B}_j$ and that the observables inside $\mathcal{B}_j$ are compatible with each other.

Junta measurements: Another family of quantum concept classes studied in this paper is k-junta measurements. Junta measurements are examples of low-depth quantum circuits that are also ϵ-concentrated with $\epsilon = 0$. Classical k-juntas have been studied extensively in the literature [MOS03, Val12, Bla09, FKR+04]. Given $k \in \mathbb{N}$, a k-junta is a Boolean function that depends on at most k input coordinates. Following a similar path, we say that a two-valued quantum measurement is a k-junta POVM if its outcomes depend on at most k coordinates of the underlying d-qubits system. In practice, classifying quantum states likely requires processing only a small subsystem of dimension $2^k \ll 2^d$. Therefore, the natural question is: *Which subsystem of* $\dim = 2^k$ *leads to the lowest prediction loss?*

Answering this question motivates us to study learning k-juntas. For that, we adopt our QLD algorithm and propose a method to find such subsystems (see Algorithm 2). We show in the following theorem that this algorithm provably finds the best subsystem and the best k-junta measurement.

Theorem 3.16 *Under the maximally mixed X-system and given $k \leq d$, Algorithm 2 learns the class of k-junta measurements with error up to* $\mathsf{opt}_{\mathcal{C}_k} + \sqrt{\frac{1}{n} S(\mathcal{A}_k, \delta)}$ *where $\mathcal{C}_k$ is the class of all k-junta measurements and $\mathcal{A}_k$ the set of all* $\mathbf{s} \in \{0,1,2,3\}^d$ *with at most k non-zero elements. Furthermore, the output of the algorithm is a k-junta measurement on a subsystem of* $\dim = 2^k$ *and its running time is* $\mathcal{O}\big(n(4d)^k\big)$.

Open problem: In the classical setting, it is known that several classes such as constant-depth circuits are special cases of concentrated classes. One asks for a quantum analogous of this result: *are constant-depth quantum circuits have almost all of their "power spectrum" concentrated on a small set of quantum Fourier coefficients?* Answering this question with the results of this work gives fundamental insights into the complexity and learning capability of constant-depth quantum circuits. For that one may require to develop a quantum variant of well-known Håstad's switching lemma [Hås01] as it was the case in classical logical circuits.

1.2 Related Works

The literature in this area is very broad. Therefore, we only can give pointers to some of the best known and most relevant results.

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 [OW16, OW17, HHJ+16]. In particular, for trace distance as the measure, [HHJ+16] showed that $\mathcal{O}\big(d^2/\epsilon\big)$ copies are needed to approximate a full rank state ρ on a d dimensional space. *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 [BOW19, BCL20]. This is again done by measuring multiple identical copies of ρ. A survey on this topic is provided in [MdW16]. Another problem in this category is *state discrimination*, in which we want to tell whether $\rho = \sigma_1$ or σ_2 [BC09, Gam08, GK10]. In the asymptotic sense, another framework is *quantum hypothesis testing* which is surveyed in [ANSV08]. An operational view of learning quantum states is introduced by [Aar07]. 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 measurements E. Another related work in this line is [CHY15]. Here, there is an unknown two-outcome measurement E. 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. Our training samples consist of post-measurement states and with labels in $\{0, 1\}$. Another direction is based on the well-known work of [BJ98]. In this model, we measure identical copies of a *superposition* state to solve a classical PAC learning problem. Other related works in this area are [AdW17, ADW18, KRS19, BV97, SG04]. Lastly, estimating the decomposition of an operator with respect to a set of elementary operators has been studied in [CvSW$^+$20, PMS$^+$14].

2. Preliminaries

Notations: For any $d \in \mathbb{N}$, let H_d be a the Hilbert space of d-qubits. By $\mathcal{B}(H_d)$ denote the space of all bounded (linear) operators acting on H_d. The identity operator is denoted by I_d. An operator Π is a projection if $\Pi^2 = \Pi$. As usual, a quantum state, in its most general form, is denoted by a *density operator*; that is a Hermitian, unit-trace, and non-negative linear operator. A quantum measurement/observable is modeled as a POVM. A POVM $\mathcal{M}$ is represented by a set of operators $\mathcal{M} := \{M_v, v \in \mathcal{V}\}$, where $\mathcal{V}$ is the (finite) set of possible outcomes. The operators M_v are non-negative and form a resolution of identity, that is $M_v = M_v^\dagger \geq 0, \sum_v M_v = I_d$. The measurement $\mathcal{M}$ is *sharp/projective* if the operators M_v are projections. For any sequence $\mathbf{s} \in \{0, 1, 2, 3\}^d$, define $\mathrm{supp}(\mathbf{s}) \triangleq \{\ell \in [d] : s_\ell \neq 0\}$.

2.1 Quantum Learning Model

Before presenting the main results, let us carefully define our quantum learning model. We follow the model presented in [HPS21] and focus on binary classification of quantum states on H_d for some $d \in \mathbb{N}$. Therefore, $\{0, 1\}$ is the set of possible classical labels. Further, ρ_0 and ρ_1 denote the overall state of the system when the label is 0 and 1, respectively. Let $p_0 = 1 - p_1 \in (0, 1)$ be the probability distribution on $\{0, 1\}$. The states ρ_0, ρ_1 and the probabilities p_0, p_1 are unknown. Available are only n training samples $(\rho_{y_i}, y_i), i = 1, 2, ..., n$, generated i.i.d. according to (p_0, p_1).

We seek a procedure that given the training samples, construct a predictor for the task of classification. Therefore, a predictor is a *quantum measurement* which acts on the feature state H_d and outputs $\hat{y} \in \{0, 1\}$ as the predicted label. To test a predictor $\mathcal{M} := \{M_0, M_1\}$, a new sample is drawn randomly according to (p_0, p_1). Suppose that the test sample is $(\rho_{y_{test}}, y_{test})$. Then without revealing y_{test}, we measure $\rho_{y_{test}}$ with $\mathcal{M}$. The outcome of this predictor is $\hat{y}$ with probability $\mathrm{tr}\{M_{\hat{y}}\rho_{y_{test}}\}, \hat{y} \in \{0, 1\}$. Hence, the joint distribution of $(Y, \hat{Y})$ is given by $\mathbb{P}\{Y = y,$

$\hat{Y} = \hat{y}\} = p_y \operatorname{tr}\{M_{\hat{y}}\rho_y\}$. We use the $0-1$ loss to measure the accuracy of the predictors. Hence, the generalization loss of a predictor $\mathcal{M}$ is $L_{0-1}(\mathcal{M}) = \mathbb{P}\{Y \neq \hat{Y}\}$. We also consider a *concept class* that is a collection $\mathcal{C}$ of quantum measurements from which the predictor is to be selected. With this setup, a quantum learning algorithm is a process that with the training samples as the input, selects a predictor from the given concept class.

For the task of binary classification which is the focus of this work, the predictors are two-valued POVMs denoted by $\mathcal{M} := \{M_0, M_1\}$ with outcome in $\{0, 1\}$. For each measurement $\mathcal{M}$, we define $G_M := M_1 - M_0$ as its operator representation. With G_M one can construct $\mathcal{M}$. This is because $M_0 = \frac{1}{2}(I - G_M)$ and $M_1 = \frac{1}{2}(I + G_M)$. Hence, there is a one-to-one correspondence between any valid G_M and two-valued measurement $\mathcal{M}$.

2.2 Review of the Quantum Fourier Expansion on Qubits

Before presenting our main results in Section 3, we give a more detailed overview of the quantum Fourier expansion on qubits.

The standard Fourier expansion on the Boolean cube states that any real-valued and bounded function on $\{-1, 1\}^d$ can be written as $f(\mathbf{x}) = \sum_{S\subseteq[d]} f_{\mathcal{S}}\psi_{\mathcal{S}}(\mathbf{x})$, where $\psi_{\mathcal{S}}(\mathbf{x}) = \prod_{j\in\mathcal{S}} x_j$ and $f_{\mathcal{S}} = \mathbb{E}_{\sim Unif}[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})]$. The quantum analogous of the Boolean Fourier expansion is based on *Pauli* operators together with the identity. These operators are denoted as $\{\sigma^0, \sigma^1, \sigma^2, \sigma^3\}$, and characterized in the standard basis as

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

It is known that any Hermitian operator on H_1 can be uniquely written as a linear combination of the Pauli operators with real coefficients [Hol12]. Therefore, based on Pauli operators, one can develop a Fourier expansion for bounded operators on H_d, the Hilbert space of d qubits. In what follows, we give an overview of this expansion as presented in [MO10].

We first present a *noncommutative* analog of correlation between random variables, which is an important concept in ordinary probability theory. Let ρ be a fixed density operator acting on the Hilbert space H_d. Consider the following inner product between any bounded operators A, B:

$$\langle A, B\rangle_\rho \triangleq \operatorname{tr}\{A^\dagger B\rho\}. \tag{2}$$

It is easy to see that $\langle A, B\rangle_\rho$ is a valid *complex inner product*. Further, when ρ is maximally mixed (i.e., $\rho = \frac{1}{2^d} I_d$), then this inner product is equivalent to the normalized Hilbert-Schmidt inner product [Wil13].

Going back to quantum Fourier development, it is not difficult to check that for $d = 1$, the Pauli operators are orthonormal. For $d > 1$, we create an orthonormal basis by *tensoring* the Pauli operators. For that define,

$$\sigma^{\mathbf{s}} \equiv \sigma^{s_1} \otimes \sigma^{s_2} \otimes \cdots \otimes \sigma^{s_d}, \qquad \forall \mathbf{s} \in \{0, 1, 2, 3\}^d. \tag{3}$$

When ρ is maximally mixed, by construction, these operators are orthonormal, that is for any $\mathbf{s}$, $\mathbf{t} \in \{0, 1, 2, 3\}^d$, $\mathbf{s} \neq \mathbf{t}$ the following equalities hold: $\langle \sigma^{\mathbf{s}}, \sigma^{\mathbf{t}}\rangle_\rho = \prod_{j=1}^d \operatorname{tr}\{\sigma^{s_j}\sigma^{t_j}\} = 0$, and $\langle \sigma^{\mathbf{s}}, \sigma^{\mathbf{s}}\rangle_\rho = 1$. Further, the set of all $\sigma^{\mathbf{s}}$ for $\mathbf{s} \in \{0, 1, 2, 3\}^d$ forms an orthonormal basis for bounded operators on H_d. More formally, we obtain the Fourier expansion below.

Proposition 2.1 (Quantum Fourier on qubits) *Any bounded operator A on the Hilbert space of d qubits can be uniquely written 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 Fourier coefficients of A and are given as $a_{\mathbf{s}} = \frac{1}{2^d} \operatorname{tr}\left\{A\sigma^{\mathbf{s}}\right\}$.*

For completeness, the proof is given in Appendix B. We end this section by formally describing the concept classes in our paper.

Definition 2.2 (Concentrated measurement classes) *Given $\epsilon \in [0,1]$ and a degree set $\mathcal{A} \subseteq \{0,1,2,3\}^d$, a concept class $\mathcal{C}$ is said to be ϵ-concentrated around $\mathcal{A}$, if for each measurement $\mathcal{M} \in \mathcal{C}$ with $\mathrm{G}_M = M_1 - M_0$ as its operator representation, the following inequality holds $\sum_{\mathbf{s}\notin\mathcal{A}} g_{\mathbf{s}}^2 \leq \epsilon^2$, where $g_{\mathbf{s}}$ are the Fourier coefficients of G_M.*

3. Main Results

With the ingredients given in the previous section, we present our main results. First, in Section 3.1, we make a connection between the $0-1$ loss and the quantum Fourier expansion. Then, in Section 3.2, we study learning with respect to concentrated measurement classes. Lastly, in Section 3.3 we focus on a particular class of such measurements called k-junta POVMs.

3.1 Mislabeling probability and the quantum Fourier expansion

The first result of the paper is the fundamental connection between the Fourier expansion in Section 2.2 and the $0-1$ loss in the QPAC learning framework. We present the following theorem which is proved in Section 4.1.

Theorem 3.1 *There is a quantum operator F_Y with eigenvalues in $\{-1,1\}$ such that the $0-1$ loss of any measurement $\mathcal{M} := \{M_0, M_1\}$ equals to $L_{0-1}(\mathcal{M}) = \frac{1}{2} - \frac{1}{2}\langle \mathrm{G}_M, \mathrm{F}_Y\rangle_{\rho_{XY}}$, where $\rho_{XY} := p_0\hat{\rho}_0 \otimes |0\rangle\langle 0| + p_1\hat{\rho}_1 \otimes |1\rangle\langle 1|$ is the average feature-label state and $\mathrm{G}_M = M_1 - M_0$ is the operator representation of $\mathcal{M}$. Additionally, F_Y commutes with all such G_M's.*

We note that F_Y is viewed as the operator representation of the labeling. With this perspective, we bridge to powerful tools developed for Hilbert-Schmidt operators, namely the triangle inequality, and Cauchy-Schwarz inequality. We also note that, Theorem 3.1 resembles a similar relation in the classical settings. That is, given any two function $f, g : \mathcal{X} \mapsto \{-1,1\}$, $\mathbb{P}\{f(X) \neq g(X)\} = \frac{1}{2} - \frac{1}{2}\mathbb{E}[f(X)g(X)]$, where the expectation can be viewed as an inner-product between f and g.

From Theorem 3.1 and the quantum Fourier expansion, we immediately obtain the following corollary expressing the mislabeling probability in the Fourier domain.

Corollary 3.2 *Under maximally mixed X-system, the $0-1$ loss of of any predictor $\mathcal{M}$ decomposes as $L_{0-1}(\mathcal{M}) = \frac{1}{2} - \frac{1}{2}\sum_{\mathbf{s}} g_{\mathbf{s}} f_{\mathbf{s}}$, where $g_{\mathbf{s}}$'s are the Fourier coefficients of G_M and $f_{\mathbf{s}} = \operatorname{tr}\{\mathrm{F}_Y\sigma^{\mathbf{s}}\rho_{XY}\}$ are Fourier coefficients of F_Y.*

We use Theorem 3.1 and this corollary as the building block of our learning algorithms.

3.2 Quantum Fourier Learning Algorithm

We leverage the previous section and study learning with respect to concentrated measurement classes as in Definition 2.2. Let $\mathcal{C}$ be the concept class that is also ϵ-concentrated around a degree set $\mathcal{A}$. From Corollary 3.2, since most of the spectral power of the measurements in $\mathcal{C}$ is concentrated around $\mathcal{A}$, then the $0-1$ loss depends mostly on $f_{\mathbf{s}}, g_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$. Further, the loss is minimizes when $g_{\mathbf{s}} = f_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$. Hence, the main idea for our learning algorithm is to estimate $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$.

In the classical setting, estimating the Fourier coefficients is easily done by empirical averaging. But in quantum, estimating all such Fourier coefficients from the same set of training samples might not be feasible. The reason is that the observables corresponding to estimating each $f_{\mathbf{s}}$ may not be *compatible* and, thus, the uncertainty principle limits the accuracy of the estimations. To address this issue, we propose "compatibility covering". As a result, we derive bounds on the sample complexity that depend not only on the size but also on the "compatibility structure" of the class. That said, to present our results, we first provide the following definition from [Hol12].

Definition 3.3 *A set of observables $M_j = \{\Lambda_u^j : u \in \mathcal{U}\}, j = 1, 2, ..., k$ are called compatible if there exist a fine-grained measurement $M_{ref} = \{A_w : w \in \mathcal{W}\}$ and a function $e : \mathcal{W} \mapsto \mathcal{U}^k$ such that $\Lambda_u^j = \sum_{w \in \mathcal{W}: e(w)_j = u} A_w$ for all $u \in \mathcal{U}$ and all $j \in [k]$.*

When a group of operators are not compatible, we need a compatibility covering tackling the uncertainty principle. Since we consider the compatibility for estimating $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$, we need to present the compatibility covering as follows. The use of this covering is described in detail in Section 3.2.3.

Definition 3.4 *Any subset $\mathcal{B} \subseteq \{0, 1, 2, 3\}^d$ is said to be compatible if all $\sigma^{\mathbf{s}}, \mathbf{s} \in \mathcal{B}$ mutually commute, i.e., $\sigma^{\mathbf{s}}\sigma^{\mathbf{t}} - \sigma^{\mathbf{t}}\sigma^{\mathbf{s}} = 0, \forall \mathbf{t}, \mathbf{s} \in \mathcal{B}$. Furthermore, a compatibility covering of a degree set $\mathcal{A} \subseteq \{0, 1, 2, 3\}^d$ is a collection of distinct compatible subsets $\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_m \subseteq \mathcal{A}$ for some $m \in \mathbb{N}$ such that $\mathcal{A} = \bigcup_{j=1}^m \mathcal{B}_j$.*

Note that there always exists a covering, as the single-element subsets of $\mathcal{A}$ form a trivial one. With the above definition, we present our second main result.

Theorem 3.5 *Let $\mathcal{C}$ be a concept class that is ϵ-concentrated around a degree set $\mathcal{A}$, where $\epsilon \in [0, 1]$. Then, there exists a quantum learning algorithm that learns $\mathcal{C}$ under the maximally mixed X-system and with error up to*

$$2\mathsf{opt}_{\mathcal{C}} + 2\epsilon + \mathcal{O}\left(\min_{\mathcal{B}_j \text{ covering of } \mathcal{A}} \sum_{j=1}^{m} \sqrt{\frac{|\mathcal{B}_j|}{n} \log \frac{|\mathcal{B}_j|}{\delta}}\right). \tag{4}$$

This result suggests that the quantum sample complexity depends on the compatibility structure of the class — a profound property unique to the quantum settings. In fact, under the realizability assumption (i.e., $\mathsf{opt}_{\mathcal{C}} = 0$), if $\mathcal{A}$ is fully compatible, then the sample complexity of $\mathcal{C}$ is less than $\mathcal{O}\big(\frac{|\mathcal{A}|}{\epsilon^2} \log \frac{|\mathcal{A}|}{\delta}\big)$. If $\mathcal{A}$ is fully incompatible, then the sample complexity is less than $\mathcal{O}\big(\frac{|\mathcal{A}|^2}{\epsilon^2} \log \frac{1}{\delta}\big)$.

Road map for developing QLD algorithm: We propose our QLD algorithm to learn with respect to the concentrated measurement classes. As presented in Algorithm 1, our method consists of two main procedures: 1) estimating the quantum Fourier coefficients, and 2) constructing the predictor from the estimated Fourier coefficients. The first process is described in step-by-step starting from

one Fourier coefficient (Section 3.2.1) to a compatible set of Fourier coefficients (Section 3.2.2) and lastly with compatible covering (Section 3.2.3). The second process is presented in lines 2 to 5 of Algorithm 1.

Algorithm 1: QLD Algorithm

Input: Degree set $\mathcal{A}$ with its compatibility covering $\mathcal{B}_j, j \in [m]$, and n training samples.
Output: Predictor $\mathcal{M}$

1 Run `FourierEstimation` $(\mathcal{B}_j, j \in [m])$, as in Procedure 1, to get $\hat{f}_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$.
2 **Function** `Predictor` $(\hat{f}_{\mathbf{s}}, \mathbf{s} \in \mathcal{A})$ **:**
3 Create the quantum operator $G_M = \text{sign}[\hat{F}]$, where $\hat{F} \triangleq \sum_{\mathbf{s}\in\mathcal{A}} \hat{f}_{\mathbf{s}} \sigma^{\mathbf{s}}$.
4 Construct the projection operators $\Pi_0 = \frac{1}{2}(I_d + G_M)$ and $\Pi_1 = \frac{1}{2}(I_d - G_M)$.
5 **return** POVM $\mathcal{M} := \{\Pi_0, \Pi_1\}$ as the predictor.

Next, we bound the mislabeling probability of the algorithm's predictor using the square loss for the estimation of $f_{\mathbf{s}}$'s. For that, we establish the following result with the proof in Section 4.2.

Theorem 3.6 *For $\epsilon \in (0,1)$, let $\mathcal{C}$ be any concept class that is ϵ-concentrate around a degree set $\mathcal{A} \subseteq \{0,1,2,3\}^d$. Suppose that $\sum_{\mathbf{s}\in\mathcal{A}}(\hat{f}_{\mathbf{s}} - f_{\mathbf{s}})^2 \leq \beta^2$, where $f_{\mathbf{s}}$'s are the Fourier coefficients of F_Y and $\beta \in [0,1]$. Let $\mathcal{M}$ be the output of QLD algorithm. Then, under the maximally mixed X-system, the mislabeling error of $\mathcal{M}$ is at most $2\mathsf{opt}_{\mathcal{C}} + 2\epsilon + 5\beta$.*

With this result, it remains to estimate $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ and analyze the square loss (β^2) which is discussed in the next section.

3.2.1 Estimating one Fourier Coefficient

For the moment, suppose the objective is to estimate only one Fourier coefficient $f_{\mathbf{s}}$ for some $\mathbf{s} \in \{0,1,2,3\}^d$. For that we construct a POVM as in the following.

Definition 3.7 *For any $\mathbf{s} \in \{0,1,2,3\}^d$, the quantum observable for estimating the Fourier coefficient $f_{\mathbf{s}}$ is defined as the POVM $M_{\mathbf{s}} := \{\Lambda^{\mathbf{s}}_1, \Lambda^{\mathbf{s}}_{-1}\}$ with outcomes $\{-1, 1\}$ and operators*

$$\Lambda^{\mathbf{s}}_1 \triangleq \frac{1}{2}(I_d + \sigma^{\mathbf{s}}\mathrm{F}_Y), \qquad \Lambda^{\mathbf{s}}_{-1} \triangleq \frac{1}{2}(I_d - \sigma^{\mathbf{s}}\mathrm{F}_Y), \tag{5}$$

where $\sigma^{\mathbf{s}}$ is the Pauli operator corresponding to $\mathbf{s}$ as in (3).

We measure each sample $\rho_{y_i} \otimes |y_i\rangle\langle y_i|, i \in [n]$ using $M_{\mathbf{s}}$. Let $Z_i \in \{-1, 1\}$ be the output of the measurement on the ith sample. Note that even when the samples are fixed, Z_i's are still random variables. These random variables are connected to $f_{\mathbf{s}}$ via the following lemma which is proved in Appendix D.

Lemma 3.8 *For any $\mathbf{s} \in \{0,1,2,3\}^d$ and $i \in [n]$, the outcome of $M_{\mathbf{s}}$ on the ith sample is a random variable Z_i with $\mathbb{E}[Z_i] = f_{\mathbf{s}}$, where $f_{\mathbf{s}}$ is the Fourier coefficient of F_Y and the expectation is taken with respect to the overall randomness induced by the measurement and the sample's distribution.*

Note that since the samples are i.i.d., then so are Z_i's. Hence, to estimate $f_{\mathbf{s}}$, we need to approximate the expected value of these random variables. This is done by computing the empirical mean $\hat{f}_{\mathbf{s}} = \frac{1}{n}\sum_i z_i$ and the empirical sum converges to the expectation which is $f_{\mathbf{s}}$. To show this we have the following result which is a quantum analogous of Chernoff-Hoeffding's inequality.

Lemma 3.9 *Let $\rho_1, \rho_2, ..., \rho_n$ be i.i.d. random density operators on a finite dimensional Hilbert space H. Let $\bar{\rho} = \mathbb{E}[\rho_i]$ be the corresponding average density operator.*[1] *Let $\mathcal{M}$ be a (discrete) observable on H with outcomes belonging to $[a, b]$, where $a, b \in \mathbb{R}$. If Z_i is the outcome for measuring ρ_i, then for any $t \geq 0$*

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} Z_i - \langle\mathcal{M}\rangle_{\bar{\rho}}\Big| \geq t\Big\} \leq 2\exp\Big\{-\frac{nt^2}{2(b-a)^2}\Big\},$$

where $\langle M\rangle_{\bar{\rho}}$ is the expectation value of $\mathcal{M}$ in the average state $\bar{\rho}$.

The proof is omitted as it is a direct consequence of Theorem A.19 in [AW02]. From Lemma 3.9, with $a = -1, b = 1$, we obtain that $\mathbb{P}\Big\{|\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}| \geq t\Big\} \leq 2\exp\{-\frac{1}{8}nt^2\}$. Hence, given $\delta \in [0, 1]$, with probability $(1 - \delta)$ the following inequality holds $|\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}| \leq \sqrt{\frac{8}{n}\log(2/\delta)}$.

Naive extension to multiple coefficients: A naive method to estimate all Fourier coefficients $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ is to repeat the above process on fresh batches of samples. For that, the samples are partitioned into $|\mathcal{A}|$ batches of equal size, one batch for each $\mathbf{s} \in \mathcal{A}$. For this strategy, with n samples, it is not difficult to see that, the estimation error is bounded as $\sup_{\mathbf{s}\in\mathcal{A}} |\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}| \leq \sqrt{\frac{8|\mathcal{A}|}{n}\log\frac{2}{\delta}}$.

3.2.2 Estimating *Compatible* Fourier Coefficients

To minimize the number of samples required for each estimation, we would like to measure as many observables as possible on one sample. This is possible when $M_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ are compatible. In that case, the idea is to first perform a fine-grained measurement $M_{\mathbf{s}}$ on the sample and then extract the estimations from additional post-measurement classical processes. Figure 1 demonstrates this idea, where M_{ref} is the fine-grained measurement and e represents the post-processing function. To check whether $M_{\mathbf{s}}$'s are compatible, we need the following result.

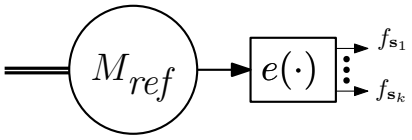

Figure 1: Measuring using compatible observables.

Proposition 3.10 *Given a set $\mathcal{A} \subseteq \{0, 1, 2, 3\}^d$, the observables $M_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ are compatible if and only if the operators $\sigma^{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ mutually commute, or equivalently $\sigma^{\mathbf{s}}\sigma^{\mathbf{t}} = \sigma^{\mathbf{t}}\sigma^{\mathbf{s}}$ for all $\mathbf{s}, \mathbf{t} \in \mathcal{A}$.*

1. Equivalently, ρ_i's are *operator-valued* random variables with values in $\mathcal{D}(H)$, the *algebra* of linear operators on H, that are self-adjoint, non-negative, and unite-trace.

Proof Since $M_{\mathbf{s}}$'s are sharp, then from Theorem 2.13 of [Hol12] they are compatible if and only if their operators commute, that is $\Lambda_z^{\mathbf{s}}$ in (5) must commute for all $\mathbf{s} \in \mathcal{A}$ and $z \in \{-1, 1\}$. Since, $\Lambda_{-1}^{\mathbf{s}} = I - \Lambda_1^{\mathbf{s}}$, then the commutativity condition is equivalent to say that $\Lambda_1^{\mathbf{s}}\Lambda_1^{\mathbf{t}} = \Lambda_1^{\mathbf{t}}\Lambda_1^{\mathbf{s}}$ for for all $\mathbf{s}, \mathbf{t} \in \mathcal{A}$. Note that $\sigma^{\mathbf{s}}$ and F_Y in (5) commute. Therefore, $\Lambda_1^{\mathbf{s}}$ and $\Lambda_1^{\mathbf{t}}$ commute if and only if $\sigma^{\mathbf{s}}$ and $\sigma^{\mathbf{t}}$ commute. With that, the proof is complete. ■

Now we are ready to explain the estimation process for a compatible degree set $\mathcal{A}$. Suppose $M_{\mathbf{s}}$, $\mathbf{s} \in \mathcal{A}$ are all compatible. From Definition 3.3, we can obtain these observables via a fine-grained measurement M_{ref} and a function e. In what follows, we construct such elements.

Without loss of generality, let $\mathcal{A} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_m\}$ for some $m \in \mathbb{N}$. Then, we construct $M_{ref} = \{\Gamma_{\mathbf{w}} : \mathbf{w} \in \{-1, 1\}^m\}$ with the operators

$$\Gamma_{\mathbf{w}} = \prod_{\ell=1}^{m} \Lambda_{w_\ell}^{\mathbf{s}_\ell}, \qquad \text{for all } \mathbf{w} \in \{-1, 1\}^m, \tag{6}$$

where $\Lambda_{w_\ell}^{\mathbf{s}_\ell}$ are the operators given in (5). By construction, M_{ref} is a valid sharp POVM. With that, the function e is simply the identity function on $\{-1, 1\}^m$, that is $e(\mathbf{w}) = \mathbf{w}, \forall \mathbf{w} \in \{-1, 1\}^m$. Therefore, the estimation process is as follows: We perform M_{ref} on each sample. Let $\mathbf{W}_i, i \in [n]$ be the outcome of this measurement on the ith sample. Then, to estimate $f_{\mathbf{s}_\ell}$, we compute the empirical mean

$$\hat{f}_{\mathbf{s}_\ell} = \frac{1}{n} \sum_{i=1}^{n} W_{i,\ell}, \qquad \text{for all } \ell \in [m], \tag{7}$$

where $W_{i,\ell}$ is the ℓth component of $\mathbf{W}_i$. We derive the following bound on the estimation error.

Lemma 3.11 *Let $\mathcal{A} \subseteq \{0, 1, 2, 3\}^d$ be a set such that $M_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ are compatible. Let $\hat{f}_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$ be the empirical estimates of $f_{\mathbf{s}}$ as in (7) and using n quantum samples. Then, for $\delta \in [0, 1]$, with probability at least $(1 - \delta)$, the following inequality holds* $\sup_{\mathbf{s}\in\mathcal{A}} |\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}| \leq \sqrt{\frac{8}{n} \log \frac{2|\mathcal{A}|}{\delta}}$.

Compared to the naive strategy, we get exponential improvements when $\mathcal{A}$ is compatible. In this case, with Lemma 3.11, the square loss of the Fourier estimations is less than $\beta^2 = \frac{8|\mathcal{A}|}{n} \log \frac{2|\mathcal{A}|}{\delta}$ as needed in Theorem 3.6. Now, if $\mathcal{A}$ is not compatible, we need compatible covering that is explained in the next section.

3.2.3 Compatible Covering

We proceed with extending the Fourier estimation process to general degree sets, consisting of incompatible measurements. The idea is to cover the degree set $\mathcal{A}$ with compatible subsets as presented in Definition 3.4. Moreover, we randomly partition the training samples into multiple batches one for each covering subset $\mathcal{B}_j, j \in [m]$. We use samples from the jth batch to estimate $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{B}_j$. Let n_j be the size of the jth batch. Among multiple coverings and choices of n_j's, we chose those that minimize the square loss. In what follows, we first explain a way to find all coverings and derive bounds on the square loss.

Finding compatible covers of $\mathcal{A}$ is a process solely depending on the structure of $\mathcal{A}$. Hence, it can be done once as a pre-processing step, independently of the samples. For that, we use an exhaustive search to find all possible compatible coverings of $\mathcal{A}$. The exhaustive search involves checking the compatibility of the observables $M_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$. From Proposition 3.10, we just need to

check the commuting relation between $\sigma^{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$. We start with an observable $M_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$, hence creating a single-element subset $\{M_{\mathbf{s}}\}$. Next, we find a compatible $M_{\mathbf{t}}, \mathbf{t} \in \mathcal{A} \backslash \{\mathbf{s}\}$ and add it to $\{M_{\mathbf{s}}\}$. We keep adding to the subset by finding observables that are compatible with all of its members. Once no more compatible observable is found, we create another subset by repeating this process for the remaining observables in $\mathcal{A}$.

Next, we proceed with analyzing the square loss for each covering and sample partitioning. Since $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{B}_j$ are estimated using the jth batch of samples, then from Lemma 3.11, we obtain a bound on the absolute difference $\sup_{\mathbf{s} \in \mathcal{B}_j} |\hat{f}_{\mathbf{s}} - f_{\mathbf{s}}|$. With these bounds, the overall square loss is bounded as

$$\sum_{\mathbf{s} \in \mathcal{A}} (\hat{f}_{\mathbf{s}} - f_{\mathbf{s}})^2 \leq \sum_{j=1}^{m} \frac{8|\mathcal{B}_j|}{n_j} \log \frac{2|\mathcal{B}_j|}{\delta}. \tag{8}$$

Hence, right-hand side in (8) gives an upper bound for β^2 as in Theorem 3.6. Therefore, we choose the coverings $\mathcal{B}_j$ and batch sizes n_j to minimize the above expression. For each covering, the optimal batch sizes n_j are given in the following lemma.

Lemma 3.12 *Having n training samples and given $\delta \in (0,1)$, the optimal batch sizes for a covering $\mathcal{B}_j, j \in [m]$ are given by $n_j^* = n \frac{\sqrt{|\mathcal{B}_j| \log(2|\mathcal{B}_j|/\delta)}}{\sum_j \sqrt{|\mathcal{B}_j| \log(2|\mathcal{B}_j|/\delta)}}$.*

Proof Let $x_j = \frac{n_j}{n}, j \in m$ and $b_j = |\mathcal{B}_j| \log(2|\mathcal{B}_j|/\delta)$. The lemma follows, by the optimization problem: $\min_{\mathbf{x} \in \mathbb{R}^m} \sum_{j=1}^m \frac{b_j}{x_j}$ subject to $x_j \geq 0$, and $\sum_{j=1}^m x_j = 1$. This can be solved via Karush–Kuhn–Tucker (KKT) approach. ■

Therefore, by replacing n_j^* into (8), the right-hand side of (8) equals to $\Big(\sum_{j=1}^m \sqrt{\frac{|\mathcal{B}_j|}{n} \log \frac{|\mathcal{B}_j|}{\delta}} \Big)^2$. As a result, the best covering is the one that minimizes this quantity. With the optimal choice of covering and sample partitioning, we complete the estimation processes which is presented Procedure 1. Moreover, we establish Theorem 3.5 by applying Theorem 3.6 with

$$\beta^2 = \min_{\mathcal{B}_j \text{covering} of \mathcal{A}} \Big(\sum_{j=1}^{m} \sqrt{\frac{|\mathcal{B}_j|}{n} \log \frac{|\mathcal{B}_j|}{\delta}} \Big)^2.$$

Procedure 1: FourierEstimation in QLD algorithm

Function `FourierEstimation` $(\mathcal{B}_j, j \in [m])$ **:**
- Partition the samples into m bathes, each of size $n_j^*, j \in [m]$, as in Lemma 3.12.
- **for** $j = 1$ **to** m **do**
  - On samples in the jth batch apply POVM M_{ref}^j with operators in (6) and $\mathbf{s}_\ell \in \mathcal{B}_j$.
  - Compute the empirical Fourier coefficient $\hat{f}_{\mathbf{s}}$ for all $\mathbf{s} \in \mathcal{B}_j$, as in (7).
- **return** $\hat{f}_{\mathbf{s}}$ for all $\mathbf{s} \in \mathcal{A}$

Note that the running time of this procedure is $\mathcal{O}\big(n 4^{|\mathcal{A}|}\big)$, as there are $4^{|\mathcal{A}|}$ that many Fourier coefficients to estimate. Moreover, the running time of Algorithm 1 is also dominated by this term.

Lastly, we point out that in QLD algorithm, the set $\mathcal{A}$ is given as an input. Whether the optimal loss $\mathsf{opt}_{\mathcal{C}}$ of a concentrated class with degree set $\mathcal{A}$ is lower than another concentrated class with a different degree set is in general open. In the next section, we answer this question for a special type of concentrated measurement classes.

3.3 Learning Junta Measurements

In this section, we focus on learning junta measurements that are examples of low-depth quantum circuits that are also ϵ-concentrated. A two-valued quantum measurement is said to be a k-junta POVM, if its outcomes depend on at most k coordinates of the underlying system H_d. To make this precise, we first define k-junta quantum operators.

Definition 3.13 *An operator A is said to be a k-junta 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 $\mathcal{J}$ sub-system $\bigotimes_{j\in\mathcal{J}} H_j$, and $I_{[d]\setminus\mathcal{J}}$ is the identity operator on the residual sub-system.*

Based on this definition, a measurement $\mathcal{M} := \{M_0, M_1\}$ is said to be a k-junta measurement if $G_M = M_1 - M_0$ is a k-junta operators.

Remark 3.14 *If A is a k-junta operator depending on a coordinate subset $\mathcal{J}$, then its quantum Fourier coefficients $a_{\mathbf{s}}$ are zero for all $\mathbf{s}$ with $\mathrm{supp}(\mathbf{s}) \nsubseteq \mathcal{J}$.*

We consider learning with respect to $\mathcal{C}_k$, the class of all k-junta measurements acting on H_d. Note that $\mathcal{C}_k$ is a ϵ-concentrated class with $\epsilon = 0$ and the degree set being all $\mathbf{s}$ with $|\mathrm{supp}(\mathbf{s})| \leq k$. In this section, we intend to find the best coordinate subset $\mathcal{J}$. In the classical setting, there are only finitely many k-juntas. Therefore, we could perform a brute-force exhaustive search over all k-juntas using ERM and find the best k-junta with the best coordinate subset. However, this is not feasible in the quantum setting as there are infinitely many k-juntas. Having this in mind, we start with characterizing the minimum loss among all k-junta measurements. We present the following lemma which is proved in Appendix E.

Lemma 3.15 *If $\mathcal{C}_k$ is the class of all k-junta measurements, then under maximally mixed X-system,*

$$\mathsf{opt}_{\mathcal{C}_k} = \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} \left\|\mathrm{F}_Y^{\subseteq\mathcal{J}}\right\|_{1,\rho}, \qquad \textit{where} \quad \mathrm{F}_Y^{\subseteq\mathcal{J}} = \sum_{\mathbf{s}:\mathrm{supp}(\mathbf{s})\subseteq\mathcal{J}} f_{\mathbf{s}}\sigma^{\mathbf{s}},$$

and $f_{\mathbf{s}} = \mathrm{tr}\{\mathrm{F}_Y\sigma^{\mathbf{s}}\rho\}$. Further, the optimal predictor is the measurement $\mathcal{M}^*$ whose operator representation is $G_{M^*} = \mathrm{sign}[\mathrm{F}_Y{}^{\subseteq\mathcal{J}^*}]$, where $\mathcal{J}^*$ is the subset maximizing the above 1-norm quantity.

Based on this result, we do not need to perform an exhaustive search over $\mathcal{C}_k$. With that, we present JuntaLearn in Algorithm 2, where the main idea is to estimate $\left\|F_Y{}^{\subseteq\mathcal{J}}\right\|_{1,\rho}$ for all k-element subsets $\mathcal{J}$. It is not difficult to see that the running time of the algorithm is $\mathcal{O}\big(n(4d)^k\big)$. In what follows, we show that this algorithm not only finds close to the best coordinate subset but also the optimal k-junta measurement. We present the following theorem which is proved in Section 4.3.

Theorem 3.16 *Under the maximally mixed X-system, Algorithm 2 learns the class of k-junta measurements with error upto $\mathsf{opt}_{\mathcal{C}_k} + \min_{\mathcal{B}_j \text{ covering of } \mathcal{A}} \sum_{j=1}^{m} \sqrt{\frac{25|\mathcal{B}_j|}{n} \log \frac{|\mathcal{B}_j|}{\delta}}$.*

Algorithm 2: JuntaLearn: Algorithm for Learning k-Junta Measurements

`JuntaLearn`:

Denote $\mathcal{B}_j, j \in [m]$ as the optimal covering of $\{\mathbf{s} \in \{0,1,2,3\}^d : |\operatorname{supp}(\mathbf{s})| \le k\}$.

Run `FourierEstimation` $(\mathcal{B}_j, j \in [m])$.

For each $\mathcal{J} \subset [d]$ with $|\mathcal{J}| = k$, compute $\left\|\hat{F}^{\subseteq \mathcal{J}}\right\|_{1,\rho}$, where $\hat{F}^{\subseteq \mathcal{J}} = \sum_{\mathbf{s}:\operatorname{supp}(\mathbf{s}) \subseteq \mathcal{J}} \hat{f}_{\mathbf{s}} \sigma^{\mathbf{s}}$ and find $\hat{\mathcal{J}}$ that maximizes it.

Run `Predictor` $(\{\hat{f}_{\mathbf{s}} : \operatorname{supp}(\mathbf{s}) \subseteq \hat{\mathcal{J}}\})$.

return POVM $\hat{\mathcal{M}} := \{\Pi_0, \Pi_1\}$ as the predictor.

4. Proofs of the Main Theorems

In this section, we present the arguments for proving the theorems in this paper.

4.1 Proof of Theorem 3.1

Note that the joint distribution of $(Y, \hat{Y})$ is given by $p_y \operatorname{tr}\{M_{\hat{y}} \rho_y\}$ for $y, \hat{y} \in \{0,1\}$, where p_y is the sample's probability. Therefore, by definition, as $M_1 = I_d - M_0$, the mislabeling probability can be written as

$$\begin{aligned} L_{0-1}(\mathcal{M}) &= \sum_y p_y \mathbb{1}\{y=1\} \operatorname{tr}\{M_0 \rho_y\} + \sum_y p_y \mathbb{1}\{y=0\} \operatorname{tr}\{(I_d - M_0)\rho_y\} \\ &= -\sum_y p_y (-1)^y \operatorname{tr}\{M_0 \rho_y\} + \mathbb{P}\Big\{Y=0\Big\}, \end{aligned} \tag{9}$$

We define a labeling operator F_Y which acts on the joint XY-system and is characterized as

$$F_Y : |x, y\rangle \mapsto -(-1)^y |x, y\rangle .$$

This mapping can be viewed as a phase detector recovering the label. Therefore, if F_Y acts on $\rho_y \otimes |y\rangle\langle y|$, it outputs $-(-1)^y \rho_y \otimes |y\rangle\langle y|$. Thus, the eigenvalues of F_Y are $\{-1, 1\}$. Now, with $G_M = M_1 - M_0 = I - 2M_0$ as the operator representation of $\mathcal{M}$ the inner product $\langle G_M, F_Y \rangle_{\rho_{XY}}$ is written as

$$\langle G_M, F_Y \rangle_{\rho_{XY}} = \operatorname{tr}\{G_M F_Y \rho_{XY}\} = \operatorname{tr}\{F_Y \rho_{XY}\} - 2 \operatorname{tr}\{M_0 F_Y \rho_{XY}\}$$

Thus, $F_Y \rho_{XY} = -\sum_y p_y (-1)^y \rho_y \otimes |y\rangle\langle y|$. Therefore, from the above equation we have that

$$\begin{aligned} \langle G_M, F_Y \rangle_{\rho_{XY}} &= -\mathbb{E}_{p_0,p_1}[(-1)^Y] + 2\sum_y p_y (-1)^y \operatorname{tr}\{M_0 \rho_y\} \\ &\overset{(a)}{=} -\mathbb{E}[(-1)^Y] + 2\Big(\mathbb{P}\Big\{Y=0\Big\} - \mathbb{P}\Big\{Y \neq \hat{Y}\Big\}\Big) \\ &= 2\mathbb{P}\Big\{Y=1\Big\} - 1 + 2\Big(\mathbb{P}\Big\{Y=0\Big\} - \mathbb{P}\Big\{Y \neq \hat{Y}\Big\}\Big) \\ &= 1 - 2\mathbb{P}\Big\{Y \neq \hat{Y}\Big\}, \end{aligned}$$

where (a) follows from (9). Hence, the mislabeling probability equals to $\frac{1}{2} - \frac{1}{2}\langle G_M, F_Y \rangle_{\rho_{XY}}$ and the proof is complete.

4.2 Proof of Theorem 3.6

We first derive a lower-bound on opt. For that we have the following lemma which is proved in Appendix C.

Lemma 4.1 *For $\epsilon \in (0,1)$, let $\mathcal{C}$ be any concept class that is ϵ-concentrate around a degree set $\mathcal{A} \subseteq \{0,1,2,3\}^d$. Then, under the maximally mixed X-system*

$$\mathsf{opt}_{\mathcal{C}} \geq \frac{1}{2} - \frac{1}{2}\big\|\mathrm{F}_Y^{\mathcal{A}}\big\|_{1,\rho} - \epsilon,$$

where $\mathrm{F}_Y^{\mathcal{A}} \triangleq \sum_{\mathbf{s}\in\mathcal{A}} f_{\mathbf{s}}\sigma^{\mathbf{s}}$ with $f_{\mathbf{s}}$ being the Fourier coefficients of F_Y.

Next, we prove the theorem by providing an upper-bound for QLD. Note that the operator representation of $\mathcal{M}$ is $G_M = \mathrm{sign}[\hat{F}]$. From Theorem 3.1, the $0-1$ loss of $\mathcal{M}$ equals to

$$\begin{aligned} L_{0-1}(\mathcal{M}) &= \frac{1}{2} - \frac{1}{2}\langle G_M, F_Y\rangle_{\rho_{XY}} = \frac{1}{4}\|F_Y - G_M\|_{2,\rho}^2 \\ &\leq \frac{1}{4}\Big(\big\|F_Y - F_Y^{\mathcal{A}}\big\|_{2,\rho} + \big\|F_Y^{\mathcal{A}} - G_M\big\|_{2,\rho}\Big)^2 \\ &\leq \frac{1}{2}\Big(\big\|F_Y - F_Y^{\mathcal{A}}\big\|_{2,\rho}^2 + \big\|F_Y^{\mathcal{A}} - G_M\big\|_{2,\rho}^2\Big), \end{aligned} \tag{10}$$

where the first equality follows from part (a) of Proposition A.1 in Appendix A and the fact that $\|F_Y\|_{2,\rho} = \|G_M\|_{2,\rho} = 1$. The second inequality follows from Minkowski's inequality for 2-norm (Proposition A.1 in Appendix A). The second inequality follows from AM-GM inequality. Observe that (10) is an upper bound on the generalization error in terms of 2-norm quantities. From the Parseval's identity (Proposition A.2), the first term in (10) equals

$$\big\|F_Y - F_Y^{\mathcal{A}}\big\|_{2,\rho}^2 = 1 - \big\|F_Y^{\mathcal{A}}\big\|_{2,\rho}^2.$$

We proceed by bounding the second term in (10). From Minkowski's inequality for 2-norm and by adding and subtracting $\hat{F}$, we have that

$$\begin{aligned} \big\|F_Y^{\mathcal{A}} - G_M\big\|_{2,\rho}^2 \leq \Big\|F_Y^{\mathcal{A}} - \hat{F}\Big\|_{2,\rho}^2 + \Big\|\hat{F} - G_M\Big\|_{2,\rho}^2 \\ + 2\Big\|F_Y^{\mathcal{A}} - \hat{F}\Big\|_{2,\rho}\Big\|\hat{F} - G_M\Big\|_{2,\rho}. \end{aligned} \tag{11}$$

We bound the first term in (11). From the Parseval's inequality (Proposition A.2 in Appendix A) we have that

$$\Big\|F_Y^{\mathcal{A}} - \hat{F}\Big\|_{2,\rho}^2 = \sum_{\mathbf{s}\in\mathcal{A}}(f_{\mathbf{s}} - \hat{f}_{\mathbf{s}})^2 \leq \epsilon'^2,$$

where the last inequality follows from the assumption.

As for the second term in (11), we expand the 2-norm and replace G_M with $\mathrm{sign}[\hat{F}]$. Therefore, we obtain that

$$\begin{aligned} \Big\|\hat{F} - G_M\Big\|_{2,\rho}^2 &= \big\|\hat{F}\big\|_{2,\rho}^2 + \Big\|\mathrm{sign}[\hat{F}]\Big\|_{2,\rho}^2 - 2\langle\hat{F}, \mathrm{sign}[\hat{F}]\rangle_\rho \\ &= \big\|\hat{F}\big\|_{2,\rho}^2 + 1 - 2\big\|\hat{F}\big\|_{1,\rho}. \end{aligned} \tag{12}$$

Next, we show that the third term in (11) is less than $4\epsilon'_n$. It suffices to show that $\left\|\hat{F} - G_M\right\|_{2,\rho} \leq 2$. For that, we use the equality in (12). By removing the last term in (12) and taking the square root we have

$$\left\|\hat{F} - G_M\right\|_{2,\rho} \leq \sqrt{1 + \left\|\hat{F}\right\|_{2,\rho}^2}.$$

From the Minkowski's inequality, we have that

$$\begin{aligned}\left\|\hat{F}\right\|_{2,\rho} &\leq \left\|F_Y^{\mathcal{A}}\right\|_{2,\rho} + \left\|F_Y^{\mathcal{A}} - \hat{F}\right\|_{2,\rho} \\ &\leq \left\|F_Y^{\mathcal{A}}\right\|_{2,\rho} + \epsilon' \leq 1 + \epsilon'.\end{aligned}$$

Hence, we get the desired bound $\left\|\hat{F} - G_M\right\|_{2,\rho} \leq \sqrt{1 + (1+\epsilon')^2} \leq 2$, assuming that $\epsilon' \leq 1/3$. Combining the bounds for each term in (11), we get

$$\begin{aligned}\left\|F_Y^{\mathcal{A}} - G_M\right\|_{2,\rho}^2 &\leq \epsilon'^2 + \left\|\hat{F}\right\|_{2,\rho}^2 + 1 - 2\left\|\hat{F}\right\|_{1,\rho} + 4\epsilon' \\ &\leq \left\|\hat{F}\right\|_{2,\rho}^2 + 1 - 2\left\|\hat{F}\right\|_{1,\rho} + 5\epsilon'.\end{aligned}$$

We plug this inequality in (10). After rearranging the terms by adding and subtracting $2\left\|F_Y^{\mathcal{A}}\right\|_{1,\rho}$, we obtain the following inequality

$$\begin{aligned}L_{0-1}(\mathcal{M}) &\leq \frac{1}{2}\Big(2 - 2\left\|F_Y^{\mathcal{A}}\right\|_{1,\rho} + 5\epsilon' + 2\big(\left\|F_Y^{\mathcal{A}}\right\|_{1,\rho} - \left\|\hat{F}\right\|_{1,\rho}\big) + \big(\left\|\hat{F}\right\|_{2,\rho}^2 - \left\|F_Y^{\mathcal{A}}\right\|_{2,\rho}^2\big)\Big) \\ &\leq 2\,\mathsf{opt}_{\mathcal{C}} + 2\epsilon + 5\epsilon',\end{aligned}$$

where the last inequality follows from Lemma 4.1 and the following argument for bounding the last two terms in the first inequality:

For the difference of square of 2-norms, we apply the Minkowski's inequality for 2-norm. Therefore, we get the following chain of inequalities:

$$\begin{aligned}\left\|\hat{F}\right\|_{2,\rho}^2 &\leq \left(\left\|F_Y^{\mathcal{A}}\right\|_{2,\rho} + \left\|F_Y^{\mathcal{A}} - \hat{F}\right\|_{2,\rho}\right)^2 \\ &\leq \left\|F_Y^{\mathcal{A}}\right\|_{2,\rho}^2 + \left\|F_Y^{\mathcal{A}} - \hat{F}\right\|_{2,\rho}^2 + 2\left\|F_Y^{\mathcal{A}}\right\|_{2,\rho}\left\|F_Y^{\mathcal{A}} - \hat{F}\right\|_{2,\rho} \\ &\leq \left\|F_Y^{\mathcal{A}}\right\|_2^2 + 3\epsilon'.\end{aligned}$$

where the last inequality holds as $\left\|F_Y^{\mathcal{A}}\right\|_{2,\rho} \leq 1$.

For the 1-norm difference, note that ρ_X is maximally mixed, that is $\rho_X = \frac{1}{2^d} I_d$. Thus, $\|A\|_{1,\rho} = \frac{1}{2^d}\|A\|_1$ for any bounded operator A, where $\|\cdot\|_1$ is the trace norm. Therefore, from the triangle-inequality for trace norm, we have that $\|A + B\|_{1,\rho} \leq \|A\|_{1,\rho} + \|B\|_{1,\rho}$. As a result, the 1-norm differences are bounded as

$$\left\|F_Y^{\mathcal{A}}\right\|_{1,\rho} - \left\|\hat{F}\right\|_{1,\rho} \leq \left\|F_Y^{\mathcal{A}} - \hat{F}\right\|_{1,\rho} \leq \left\|F_Y^{\mathcal{A}} - \hat{F}\right\|_{2,\rho} = \epsilon',$$

where the last inequality follows from the Jensen's inequality (Proposition A.1) implying that $\|\cdot\|_{1,\rho} \leq \|\cdot\|_{2,\rho}$.

4.3 Proof of Theorem 3.16

The main ingredient in the proof of this theorem is the following lemma which is proved in Appendix F.

Lemma 4.2 *Let $\mathcal{M}_{\mathcal{J}}$ be a measurement depending on only the coordinates $\mathcal{J} \subset [d]$ where its operator representation is $\mathrm{G}_M = \mathrm{sign}[B_{\mathcal{J}}]$ for some Hermitian operator $B_{\mathcal{J}}$ depending only on the coordinates of $\mathcal{J}$. Then,*

$$L_{0-1}(\mathcal{M}_{\mathcal{J}}) \leq \frac{1}{2}\Big(1 - \Big\|\mathrm{F}_{\overline{Y}}^{\subseteq \mathcal{J}}\Big\|_{1,\rho}\Big) + U\Big(\Big\|\mathrm{F}_{\overline{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_{2,\rho}\Big),$$

where U is defined as $U(x) = x^3 + \frac{3}{2}x^2 + \frac{5}{4}x$, for all $x \geq 0$.

Having this lemma, the proof of the theorem is as follows. We run `FourierEstimation`$(\mathcal{B}_j,\ j \in [m])$, where $\mathcal{B}_j$'s are the optimal covering of the degree set $\mathcal{A} = \big\{\mathbf{s} \in \{0,1,2,3\}^d : |\operatorname{supp}(\mathbf{s})| \leq k\big\}$. As a result, we obtain estimated Fourier coefficients $f_{\mathbf{s}}, \mathbf{s} \in \mathcal{A}$, with square loss satisfying the following inequality with probability $(1-\delta)$

$$\sum_{\mathbf{s}:|\operatorname{supp}(\mathbf{s})|\leq k} (f_{\mathbf{s}} - \hat{f}_{\mathbf{s}})^2 \leq \epsilon'_n := \min_{\mathcal{B}_j \text{ covering of } \mathcal{A}} \Big(\sum_{j=1}^{m} \sqrt{\frac{|\mathcal{B}_j|}{n} \log \frac{|\mathcal{B}_j|}{\delta}}\Big)^2. \tag{13}$$

For any $\mathcal{J}$, define the operator $\hat{F}^{\subseteq \mathcal{J}} = \sum_{\mathbf{s}:\operatorname{supp}(\mathbf{s})\subseteq\mathcal{J}} \hat{f}_{\mathbf{s}}\sigma^{\mathbf{s}}$ and let $\mathcal{M}$ be the measurement whose operator representation is $G_M = \mathrm{sign}[\hat{F}^{\subseteq \mathcal{J}}]$. Then, from Lemma 4.2 we have that

$$L_{0-1}(\mathcal{M}) \leq \frac{1}{2}\Big(1 - \Big\|F_{\overline{Y}}^{\subseteq \mathcal{J}}\Big\|_{1,\rho}\Big) + U\Big(\Big\|F_{\overline{Y}}^{\subseteq \mathcal{J}} - \hat{F}^{\subseteq \mathcal{J}}\Big\|_{2,\rho}\Big).$$

Next, we bound the second term in the right-hand side. From Parseval's identity (Proposition A.2 in Appendix A) and (13) we obtain that

$$\Big\|F_{\overline{Y}}^{\subseteq \mathcal{J}} - \hat{F}^{\subseteq \mathcal{J}}\Big\|_{2,\rho}^2 = \sum_{\mathbf{s}:\operatorname{supp}(\mathbf{s})\subseteq\mathcal{J}} (f_{\mathbf{s}} - \hat{f}_{\mathbf{s}})^2 \leq \epsilon'_n,$$

which implies that $L_{0-1}(\mathcal{M}) \leq \frac{1}{2}\big(1 - \big\|F_{\overline{Y}}^{\subseteq \mathcal{J}}\big\|_{1,\rho}\big) + U(\sqrt{\epsilon'_n})$. Next, we search over all $\mathcal{J}$ with k elements and find the one that maximizes $\big\|\hat{F}^{\subseteq \mathcal{J}}\big\|_{1,\rho}$. Let $\hat{\mathcal{J}}$ be such a subset and $\hat{\mathcal{M}}$ be the measurement with $G_{\hat{M}} = \mathrm{sign}[\hat{F}^{\subseteq \hat{\mathcal{J}}}]$. Then

$$L_{0-1}(\hat{\mathcal{M}}) \leq \frac{1}{2}\big(1 - \big\|F_Y{}^{\subseteq \hat{\mathcal{J}}}\big\|_{1,\rho}\big) + U(\sqrt{\epsilon'_n}). \tag{14}$$

In what follows, we show that the right-hand side is less than $\mathsf{opt}(k) + \sqrt{\epsilon'_n} + U(\sqrt{\epsilon'_n})$. Note that form the triangle inequality and Jensen's inequality (Proposition A.1) for any $\mathcal{J}$, we have that

$$\Big|\ \big\|\hat{F}^{\subseteq \mathcal{J}}\big\|_{1,\rho} - \big\|F_{\overline{Y}}^{\subseteq \mathcal{J}}\big\|_{1,\rho}\Big| \leq \big\|F_{\overline{Y}}^{\subseteq \mathcal{J}} - \hat{F}^{\subseteq \mathcal{J}}\big\|_{1,\rho} \leq \big\|F_{\overline{Y}}^{\subseteq \mathcal{J}} - \hat{F}^{\subseteq \mathcal{J}}\big\|_{2,\rho} \leq \sqrt{\epsilon'_n}$$

Let $\mathcal{J}^* = \arg\max_{\mathcal{J}:|\mathcal{J}|=k} \left\|F_Y^{\subseteq\mathcal{J}}\right\|_{1,\rho}$. Then

$$\left\|F_Y{}^{\subseteq\hat{\mathcal{J}}}\right\|_{1,\rho} \geq \left\|\hat{F}^{\subseteq\hat{\mathcal{J}}}\right\|_{1,\rho} - \sqrt{\epsilon'_n} \geq \left\|\hat{F}^{\subseteq\mathcal{J}^*}\right\|_{1,\rho} - \sqrt{\epsilon'_n} \geq \left\|F_Y{}^{\subseteq\mathcal{J}^*}\right\|_{1,\rho} - 2\sqrt{\epsilon'_n}.$$

As a result, using (14), we have that

$$\begin{aligned} L_{0-1}(\hat{\mathcal{M}}) &\leq \frac{1}{2}\big(1 - \left\|F_Y{}^{\subseteq\mathcal{J}^*}\right\|_{1,\rho} + 2\sqrt{\epsilon'_n}\big) + U(\sqrt{\epsilon'_n}) \\ &= \mathsf{opt}(k) + \sqrt{\epsilon'_n} + U(\sqrt{\epsilon'_n}), \end{aligned}$$

where the last equality follows from Lemma 3.15. The proof is complete by noting that $U(x) \leq 4x$, for any $x \in [0,1]$. This implies that $L_{0-1}(\hat{\mathcal{M}}) \leq \mathsf{opt}(k) + 5\sqrt{\epsilon'_n}$.

References

[Aar07] Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, sep 2007.

[AdW17] Srinivasan Arunachalam and Ronald de Wolf. A survey of quantum learning theory. *arXiv:1701.06806*, 2017.

[ADW18] Srinivasan Arunachalam and Ronald De Wolf. Optimal quantum sample complexity of learning algorithms. *J. Mach. Learn. Res.*, 19(1):2879–2878, January 2018.

[ANSV08] 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.

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, mar 2002.

[BC09] Stephen M. Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238, feb 2009.

[BCL20] Sebastien Bubeck, Sitan Chen, and Jerry Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, nov 2020.

[BJ98] Nader H Bshouty and Jeffrey C Jackson. Learning dnf over the uniform distribution using a quantum example oracle. *SIAM Journal on Computing*, 28(3):1136–1153, 1998.

[Bla09] Eric Blais. Testing juntas nearly optimally. In *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing - STOC '09*. ACM Press, 2009.

[BOW10] Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. *Machine learning*, 80(2-3):273–294, 2010.

[BOW19] Costin Badescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, jun 2019.

[BV97] Ethan Bernstein and Umesh Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, oct 1997.

[CHY15] Hao-Chung Cheng, Min-Hsiu Hsieh, and Ping-Cheng Yeh. The learnability of unknown quantum measurements. *QIC, Vol. 16, No. 7-8, 0615-0656 (2016)*, January 2015.

[CvSW+20] Ophelia Crawford, Barnaby van Straaten, Daochen Wang, Thomas Parks, Earl Campbell, and Stephen Brierley. Efficient quantum measurement of pauli operators in the presence of finite sampling error. *arXiv:1908.06942*, 2020.

[DW08] Ronald De Wolf. *A Brief Introduction to Fourier Analysis on the Boolean Cube*. Number 1 in Graduate Surveys. Theory of Computing Library, 2008.

[FKR+04] Eldar Fischer, Guy Kindler, Dana Ron, Shmuel Safra, and Alex Samorodnitsky. Testing juntas. *Journal of Computer and System Sciences*, 68(4):753–787, jun 2004.

[Gam08] Sébastien Gambs. Quantum classification. *0809.0444 [quant-ph]*, September 2008.

[GK10] Mădălin Guţă and Wojciech Kotłowski. Quantum learning: asymptotically optimal classification of qubit states. *New Journal of Physics*, 12(12):123032, dec 2010.

[GQJ+18] Jun Gao, Lu-Feng Qiao, Zhi-Qiang Jiao, Yue-Chi Ma, Cheng-Qiu Hu, Ruo-Jing Ren, Ai-Lin Yang, Hao Tang, Man-Hong Yung, and Xian-Min Jin. Experimental machine learning of quantum states. *Physical Review Letters*, 120(24):240501, jun 2018.

[GR06] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 2006.

[Hås01] Johan Håstad. A slight sharpening of LMN. *Journal of Computer and System Sciences*, 63(3):498–508, nov 2001.

[HHJ+16] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016.

[Hol12] Alexander S. Holevo. *Quantum Systems, Channels, Information*. DE GRUYTER, jan 2012.

[HPS21] Mohsen Heidari, Arun Padakandla, and Wojciech Szpankowski. A theoretical framework for learning from quantum data. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[KBI+20] Zhaxylyk A. Kudyshev, Simeon I. Bogdanov, Theodor Isacsson, Alexander V. Kildishev, Alexandra Boltasseva, and Vladimir M. Shalaev. Rapid classification of quantum sources enabled by machine learning. *Advanced Quantum Technologies*, 3(10):2000067, sep 2020.

[KKYS05] A. T. Kalai, A. R. Klivans, Yishay Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proc. 46th Annual IEEE Symp. Foundations of Computer Science (FOCS'05)*, pages 11–20, October 2005.

[KRS19] Varun Kanade, Andrea Rocchetto, and Simone Severini. Learning dnfs under product distributions via μ-biased quantum fourier sampling. *arXiv:1802.05690v3*, 2019.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[MdW16] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *Theory of Computing*, 1(1):1–81, 2016.

[MO10] Ashley Montanaro and Tobias J. Osborne. Quantum boolean functions. *arXiv:0810.2435*, 2010.

[MOS03] Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proc. ACM Symp. on Theory of Computing*, pages 206–212, 2003.

[MY18] Yue-Chi Ma and Man-Hong Yung. Transforming bell's inequalities into state classifiers with machine learning. *npj Quantum Information*, 4(1), jul 2018.

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016.

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography II. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, jun 2017.

[PMS+14] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1), jul 2014.

[SG04] Rocco A. Servedio and Steven J. Gortler. Equivalences and separations between quantum and classical learnability. *SIAM J. Comput.*, 33(5):1067–1092, May 2004.

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

[Val12] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, oct 2012.

[Wil13] Mark Wilde. *Quantum information theory*. Cambridge University Press, Cambridge, UK, 2013.

Appendices

A. Useful Norm and Trace Inequalities

Let ρ be a density operator. Then, given $q > 0$, the q-norm of an operator A with respect to ρ is defined as

$$\|A\|_{q,\rho} \triangleq \mathrm{tr}\Big\{(A^\dagger A)^{q/2}\rho\Big\}^{1/q}.$$

This norm is similar to the Schatten norm; however there is an additional component which is the density operator ρ. One can see that for $q = 2$, the above $\|A\|_{2,\rho}$ reduces to the Schatten 2-norm of $A\sqrt{\rho}$. In what follows, we develop a number of useful properties for the q-norm above.

Proposition A.1 *Given any density operator ρ, and a pair of bounded Hermitian operators A and B, the following properties hold:*

(a) $\|A+B\|_{2,\rho}^2 = \|A\|_{2,\rho}^2 + \|B\|_{2,\rho}^2 + 2\langle A, B\rangle_\rho$

(b) Cauchy-Schwarz inequality: $\langle A, B\rangle_\rho \leq \|A\|_{2,\rho}\|B\|_{2,\rho}$

(c) Minkowski inequality for 2-norm : $\|A+B\|_{2,\rho} \leq \|A\|_{2,\rho} + \|B\|_{2,\rho}$.

(d) Jensen's inequality: $\|A\|_{1,\rho} \leq \|A\|_{2,\rho}$.

Proof (a): Follows from the definition. **(b):** Follows from the noncommutative Cauchy-Schwarz inequality.
(c): Using (a) and (b) we can write:

$$\begin{aligned}\|A+B\|_{2,\rho}^2 &= \|A\|_{2,\rho}^2 + \|B\|_{2,\rho}^2 + 2\langle A, B\rangle_\rho \\ &\leq \|A\|_{2,\rho}^2 + \|B\|_{2,\rho}^2 + 2\|A\|_{2,\rho}\|B\|_{2,\rho} \\ &= \Big(\|A\|_{2,\rho} + \|B\|_{2,\rho}\Big)^2\end{aligned}$$

(d): We have that $\|A\|_{1,\rho} = \mathrm{tr}\{|A|\rho\}$. Since A is Hermitian, it admits a *spectral decomposition* as $A = \sum_i \mu_i |\phi_i\rangle\langle\phi_i|$, where $\mu_i \in \mathbb{R}$ and $|\phi_i\rangle$'s are an orthonormal basis. Hence, $|A|$ admits the same decomposition but with $|\mu_i|$. Thus, we obtain that

$$\|A\|_{1,\rho} = \sum_j |\mu_j| \underbrace{\langle\phi_i|\rho|\phi_i\rangle}_{\alpha_i}, \qquad \|A\|_{2,\rho} = \sqrt{\sum_j |\mu_j|^2 \langle\phi_i|\rho|\phi_i\rangle}$$

As ρ is a density operator, then $\alpha_i \in [0,1]$ and $\sum_i \alpha_i = 1$. Hence, from the Jensen's inequality for square root function we have that

$$\|A\|_{2,\rho} = \sqrt{\sum_j |\mu_j|^2 \alpha_i} \geq \sum_j |\mu_j| \alpha_i = \|A\|_{1,\rho}.$$

■

From the Fourier expansion, we immediately get the following results. We omit the proof as they are straightforward.

Proposition A.2 *Let $A, B \in \mathcal{B}(H_d)$ with Fourier coefficients $a_{\mathbf{s}}$ and $b_{\mathbf{s}}$, respectively. Then the following statements hold*

(a) $\langle A, B \rangle = \sum_{\mathbf{s}} a_{\mathbf{s}}^* b_{\mathbf{s}}$. *Hence, Parseval's identity holds:* $\|A\|_2^2 = \sum_{\mathbf{s}} |a_{\mathbf{s}}|^2$.

(b) If A is Hermitian, then $a_{\mathbf{s}}$ are real.

B. Proof of Proposition 2.1

We start with $d = 1$. Note that $\{|\ell\rangle\langle k| : \ell, k \in \{0,1\}\}$ is the standard basis for all operators on H_1. Hence, it suffices to write $|\ell\rangle\langle k|$ in terms of the Pauli operators. For any $k \neq \ell \in \{0,1\}$ we have that

$$|\ell\rangle\langle\ell| = \frac{1}{2}\sigma^0 + \frac{1}{2}(-1)^\ell \sigma^3, \qquad |\ell\rangle\langle k| = \frac{1}{2}\sigma^1 + \frac{1}{2}(-1)^\ell i\sigma^3,$$

where $i \in \mathbb{C}$ with $i^2 = -1$. This completes the proof for $d = 1$. For $d > 1$, the set of all d-fold tensor product of $|\ell\rangle\langle k|$'s is the standard basis for $\mathcal{B}(H_d)$. Hence, from the above relations, such basis elements can be written as a linear combinations of $\sigma^{\mathbf{s}}, \mathbf{s} \in \{0,1\}^d$. That completes the proof for $d > 1$. The uniqueness follows as $\sigma^{\mathbf{s}}$ are orthogonal.

C. Proof of Lemma 4.1

From Theorem 3.1, the generalization error of any measurement $\mathcal{M} \in \mathcal{C}$ can be written as

$$L_{0-1}(\mathcal{M}) = \frac{1}{2} - \frac{1}{2}\langle F_Y, G_M \rangle_\rho,$$

where $G_M = M_1 - M_0$ is the operator representation of $\mathcal{M} := \{M_0, M_1\}$. Let $G_M^{\mathcal{A}} = \sum_{\mathbf{s}\in\mathcal{A}} g_{\mathbf{s}} \sigma^{\mathbf{s}}$ be the partial Fourier decomposition of G_M restricted to degree set $\mathcal{A}$. Since, $\mathcal{C}$ is ϵ-concentrated around $\mathcal{A}$, then from Definition 2.2 we have that $\sum_{\mathbf{s}\in\mathcal{A}} g_{\mathbf{s}}^2 \leq \epsilon^2$. From from the Parseval's identity (Proposition A.2 in Appendix A), this condition is equivalent to $\left\|G_M - G_M^{\mathcal{A}}\right\|_{2,\rho} \leq \epsilon$. Then, by adding and subtracting $G_M^{\mathcal{A}}$, we obtain that

$$\begin{aligned}\langle F_Y, G_M \rangle &= \langle F_Y, G_M^{\mathcal{A}} \rangle_\rho + \langle F_Y, (G_M - G_M^{\mathcal{A}}) \rangle_\rho \\ &\leq \langle F_Y, G_M^{\mathcal{A}} \rangle_\rho + \|F_Y\|_{2,\rho} \left\|G_M - G_M^{\mathcal{A}}\right\|_{2,\rho} \leq \langle F_Y, G_M^{\mathcal{A}} \rangle_\rho + \epsilon, \end{aligned} \tag{15}$$

where the first inequality follows from the Cauchy–Schwarz inequality (Proposition A.1 in Appendix A) and the second inequality follows because $\|F_Y\|_{2,\rho} = 1$. From properties of the quantum Fourier (Proposition 2.1), the above inner-product decomposes as

$$\langle F_Y, G_M^{\mathcal{A}} \rangle_\rho = \sum_{\mathbf{s} \in \mathcal{A}} f_{\mathbf{s}} g_{\mathbf{s}} = \langle F_Y^{\mathcal{A}}, G_M^{\mathcal{A}} \rangle_\rho.$$

Consequently, from the above equality and (15), we obtain that

$$\begin{aligned}
\langle F_Y, G_M \rangle_\rho &\leq \langle F_Y^{\mathcal{A}}, G_M^{\mathcal{A}} \rangle_\rho + \epsilon = \langle F_Y^{\mathcal{A}}, G_M \rangle_\rho + \langle F_Y^{\mathcal{A}}, (G_M^{\mathcal{A}} - G_M) \rangle_\rho + \epsilon \\
&\overset{(a)}{\leq} \langle F_Y^{\mathcal{A}}, G_M \rangle_\rho + \big\|F_Y^{\mathcal{A}}\big\|_{2,\rho} \big\|G_M^{\mathcal{A}} - G_M\big\|_{2,\rho} + \epsilon \\
&\overset{(b)}{\leq} \langle F_Y^{\mathcal{A}}, G_M \rangle_\rho + \epsilon \big\|F_Y^{\mathcal{A}}\big\|_{2,\rho} + \epsilon \\
&\overset{(c)}{\leq} \langle F_Y^{\mathcal{A}}, G_M \rangle_\rho + 2\epsilon \\
&\overset{(d)}{\leq} \big\|F_Y^{\mathcal{A}}\big\|_{1,\rho} + 2\epsilon, \qquad (16)
\end{aligned}$$

where (a) follows from the Cauchy–Schwarz inequality, (b) follows from the assumption of the lemma, stating that $\big\|G_M^{\mathcal{A}} - G_M\big\|_{2,\rho} \leq \epsilon$, (c) holds because of the Parseval's identity implying that $\big\|F_Y^{\mathcal{A}}\big\|_{2,\rho} \leq \|F_Y\|_{2,\rho} = 1$ and (d) holds because of the following argument:

As $\rho_X = \frac{1}{2^d} I_d$ and $F_Y^{\mathcal{A}}$ and G_M act only on the X-system, then $\langle F_Y^{\mathcal{A}}, G_M \rangle_\rho = \frac{1}{2^d} \operatorname{tr}\{F_Y^{\mathcal{A}} G_M\}$. From properties of the trace-norm, we have that $|\operatorname{tr}\{F_Y^{\mathcal{A}} G_M\}| \leq \big\|F_Y^{\mathcal{A}} G_M\big\|_1 \leq \big\|F_Y^{\mathcal{A}}\big\|_1 \|G_M\|_\infty$. By definition $G_M = M_1 - M_0$ where $M_1, M_0 \geq 0$ and $M_1 + M_0 = I$. Therefore, $|G_M| \leq I_d$, implying that $\|G_M\|_\infty \leq 1$. Therefore, $\langle F_Y^{\mathcal{A}}, G_M \rangle_\rho \leq \frac{1}{2^d} \big\|F_Y^{\mathcal{A}}\big\|_1 = \big\|F_Y^{\mathcal{A}}\big\|_{1,\rho}$.

Lastly, from (16), we get the desired bound on the generalization error; hence, the proof is complete.

D. Proof of Lemma 3.8

Note that the conditional expectation of Z_i given the ith sample $\hat{\rho}_{y_i} \otimes |y_i\rangle\langle y_i|$ can be written as

$$\mathbb{E}[Z_i | \hat{\rho}_{y_i} \otimes |y_i\rangle\langle y_i|] = \sum_{z \in \{-1,1\}} z \ \operatorname{tr}\big\{\Lambda_z^{\mathbf{s}} \big(\hat{\rho}_{y_i} \otimes |y_i\rangle\langle y_i|\big)\big\}.$$

Note that the above value in fact is the *expectation value* of the observable $M_{\mathbf{s}}$ in state $\hat{\rho}_{y_i} \otimes |y_i\rangle\langle y_i|$. With that the overall expectation of Z_i equals

$$\begin{aligned}
\mathbb{E}[Z_i] &= \sum_{y \in \mathcal{Y}} \sum_{z \in \{-1,1\}} z \ p_y \operatorname{tr}\big\{\Lambda_z^{\mathbf{s}} \big(\hat{\rho}_y \otimes |y\rangle\langle y|\big)\big\} \\
&= \sum_{z \in \{-1,1\}} z \ \operatorname{tr}\{\Lambda_z^{\mathbf{s}} \rho_{XY}\} \\
&= \operatorname{tr}\Big\{\big(\Lambda_1^{\mathbf{s}} - \Lambda_{-1}^{\mathbf{s}}\big)\rho_{XY}\Big\} \\
&= \operatorname{tr}\Big\{\sigma^{\mathbf{s}} F_Y \rho_{XY}\Big\}
\end{aligned}$$

where we used the linearity of the trace and (5). Hence, the proof is complete as the right-hand side is $f_{\mathbf{s}}$.

E. Proof of Lemma 3.15

We start with proving a lower bound on $\mathsf{opt}(k)$. Fix a k element coordinate subset $\mathcal{I} \subset [d]$ and consider a k-junta measurement that depends only on coordinates $\mathcal{I}$. Such measurement belongs to the ε-concentrated measurement class with $\varepsilon = 0$ and degree set

$$\mathcal{A} = \left\{ \mathbf{s} \in \{0,1,2,3\}^d : \mathrm{supp}(\mathbf{s}) \subseteq \mathcal{I} \right\}.$$

Therefore, we can apply Lemma 4.1 to get the following lower bound

$$L_{0-1}(\mathcal{M}) \geq \frac{1}{2} - \frac{1}{2}\left\| F_{\overline{Y}}^{\subseteq \mathcal{I}} \right\|_{1,\rho} \geq \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J} \subset [d] : |\mathcal{J}| = k} \left\| F_{\overline{Y}}^{\subseteq \mathcal{J}} \right\|_{1,\rho},$$

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 $\mathsf{opt}(k)$:

$$\mathsf{opt}(k) \geq \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J} \subset [d] : |\mathcal{J}| = k} \left\| F_{\overline{Y}}^{\subseteq \mathcal{J}} \right\|_{1,\rho}. \tag{17}$$

Next, we establish the achievability of the lower bound. Again fix a k-element subset $\mathcal{J} \subset [d]$ and let $G_{M_{\mathcal{J}}} = \mathrm{sign}[F_{\overline{Y}}^{\subseteq \mathcal{J}}]$. 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 Theorem 3.1,

$$\begin{aligned} L_{0-1}(\mathcal{M}_{\mathcal{J}}) &= \frac{1}{2} - \frac{1}{2}\langle G_{M_{\mathcal{J}}}, F_Y \rangle_\rho \\ &= \frac{1}{2} - \frac{1}{2}\langle \mathrm{sign}[F_{\overline{Y}}^{\subseteq \mathcal{J}}], F_{\overline{Y}}^{\subseteq \mathcal{J}} \rangle_\rho \\ &= \frac{1}{2} - \frac{1}{2}\left\| F_{\overline{Y}}^{\subseteq \mathcal{J}} \right\|_{1,\rho}, \end{aligned}$$

where the last equality follows from the identity $\|A\|_{1,\rho} = \mathrm{tr}\{A\,\mathrm{sign}[A]\rho\}$ that holds for any Hermitian and bounded operator A. With the above inequality, optimizing over $\mathcal{J}$ gives

$$\mathsf{opt}(k) \leq \min_{\mathcal{J} \subset [d] : |\mathcal{J}| = k} L_{0-1}(\mathcal{M}_{\mathcal{J}}) = \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J} \subset [d] : |\mathcal{J}| = k} \left\| F_{\overline{Y}}^{\subseteq \mathcal{J}} \right\|_{1,\rho}. \tag{18}$$

As the upper bound in (18) matches with the lower bound in (17), then we obtain the equality in (18). This proves the expression for $\mathsf{opt}(k)$ and that $\mathcal{M}_{\mathcal{J}^*}$ is the best k-junta measurement.

F. Proof of Lemma 4.2

From Theorem 3.1 in the main text, the generalization error of $\mathcal{M}_{\mathcal{J}}$ can be written as $\frac{1}{2} - \frac{1}{2}\langle F_Y, \mathrm{sign}[B_{\mathcal{J}}] \rangle_\rho$. Since $\mathrm{sign}[B_{\mathcal{J}}]$ is a function depending only on the coordinates of $\mathcal{J}$, then in the inner product, we can replace F_Y with $F_{\overline{Y}}^{\subseteq \mathcal{J}}$. As short-hand we remove the subscribe ρ in the norm quantities. Hence, we obtain that

$$\begin{aligned} L_{0-1}(\mathcal{M}_{\mathcal{J}}) &= \frac{1}{2} - \frac{1}{2}\langle F_{\overline{Y}}^{\subseteq \mathcal{J}}, \mathrm{sign}[B_{\mathcal{J}}] \rangle \\ &= \frac{1}{4}\Big(1 - \left\| F_{\overline{Y}}^{\subseteq \mathcal{J}} \right\|_2^2 + \left\| F_{\overline{Y}}^{\subseteq \mathcal{J}} - \mathrm{sign}[B_{\mathcal{J}}] \right\|_2^2\Big), \end{aligned} \tag{19}$$

where the last equality holds because of part (a) of Proposition A.1 in Appendix A and the fact that $\|\text{sign}[B_{\mathcal{J}}]\|_2 = 1$. The 2-norm quantity above is upper-bounded as follows

$$\begin{aligned}\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}} - \text{sign}[B_{\mathcal{J}}]\right\|_2^2 &\overset{(a)}{\leq} \left(\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}} - B_{\mathcal{J}}\right\|_2 + \|B_{\mathcal{J}} - \text{sign}[B_{\mathcal{J}}]\|_2\right)^2, \\ &= \left(\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}} - B_{\mathcal{J}}\right\|_2^2 + \|B_{\mathcal{J}} - \text{sign}[B_{\mathcal{J}}]\|_2^2 + 2\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}} - B_{\mathcal{J}}\right\|_2 \|B_{\mathcal{J}} - \text{sign}[B_{\mathcal{J}}]\|_2\right), \quad (20)\end{aligned}$$

where (a) follows from the Minkowski's Inequality inequality for 2-norm (Proposition A.1). Therefore, from part (a) of Proposition A.1, we have that

$$\begin{aligned}\|B_{\mathcal{J}} - \text{sign}[B_{\mathcal{J}}]\|_2^2 &= \|B_{\mathcal{J}}\|_2^2 + 1 - 2\langle B_{\mathcal{J}}, \text{sign}[B_{\mathcal{J}}]\rangle \\ &= 1 + \|B_{\mathcal{J}}\|_2^2 - 2\|B_{\mathcal{J}}\|_1, \quad (21)\end{aligned}$$

where the last equality follows, because for any Hermitian operator A the identity $|A| = A\,\text{sign}[A]$ holds. Consequently from (19), (21) and (20), we obtain the following upper bound

$$\begin{aligned}4L_{0-1}(\mathcal{M}_{\mathcal{J}}) \leq 2 - 2\|B_{\mathcal{J}}\|_1 &+ \underbrace{\|B_{\mathcal{J}}\|_2^2 - \left\|F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2^2}_{\text{(I)}} \\ &+ \left\|F_{\overline{Y}}^{\subseteq\mathcal{J}} - B_{\mathcal{J}}\right\|_2^2 + 2\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}} - B_{\mathcal{J}}\right\|_2 \underbrace{\|B_{\mathcal{J}} - \text{sign}[B_{\mathcal{J}}]\|_2}_{\text{(II)}}. \quad (22)\end{aligned}$$

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

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

$$\begin{aligned}\|B_{\mathcal{J}}\|_2^2 &\leq \left(\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2 + \left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2\right)^2 \\ &= \left\|F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2^2 + \left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2^2 + 2\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2 \left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2 \\ &\leq \left\|F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2^2 + \left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2^2 + 2\left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2\end{aligned}$$

where the second inequality is due Bessel's inequality implying that $\left\|F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2 \leq 1$. Hence, the term (I) in (22) is upper bounded as

$$\text{(I)} \leq \lambda_1 \triangleq \left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2^2 + 2\left\|B_{\mathcal{J}} - F_{\overline{Y}}^{\subseteq\mathcal{J}}\right\|_2. \quad (23)$$

Bounding (II): From (21), we have

$$\begin{aligned}
\|B_{\mathcal{J}} - \operatorname{sign}[B_{\mathcal{J}}]\|_2^2 &= 1 + \|B_{\mathcal{J}}\|_2^2 - 2\|B_{\mathcal{J}}\|_1 \\
&\overset{(a)}{\leq} 1 + 2\Big(\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_2^2 + \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2\Big) - 2\|B_{\mathcal{J}}\|_1 \\
&\overset{(b)}{=} 1 + 2\Big(\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_2^2 + \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2\Big) - 2\Big(\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1 + \big(\|B_{\mathcal{J}}\|_1 - \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1\big)\Big) \\
&= 1 + 2\Big(\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_2^2 - \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1\Big) + 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 - 2\Big(\|B_{\mathcal{J}}\|_1 - \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1\Big) \\
&\overset{(c)}{\leq} 1 + 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 - 2\Big(\|B_{\mathcal{J}}\|_1 - \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1\Big) \\
&\overset{(d)}{\leq} 1 + 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 + 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2, \qquad (24)
\end{aligned}$$

where (a) follows from the Minkowski's inequality for 2-norm and the AM-GM inequality. Equality (b) follows by adding and subtracting $\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1$. Inequality (c) holds because of the inequality $\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_2^2 \leq \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1$. Lastly, inequality (d) holds because of the following chain of inequalities

$$\Big|\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1 - \|B_{\mathcal{J}}\|_1\Big| \leq \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_1 \leq \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2, \qquad (25)$$

where the first inequality is due to the triangle inequality for trace norm and the second inequality is due to Jensen's inequality (Proposition A.1).

Next, we show that the quantity $\big\|B_{\mathcal{J}} - \operatorname{sign}[h_{\mathcal{J}}]\big\|_2$ without the square is upper bounded by the same term as in the right-hand side of (24). That is

$$\text{(II)} = \big\|B_{\mathcal{J}} - \operatorname{sign}[h_{\mathcal{J}}]\big\|_2 \leq \lambda_2 \triangleq 1 + 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 + 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2. \qquad (26)$$

The argument is as follows: if $\big\|B_{\mathcal{J}} - \operatorname{sign}[h_{\mathcal{J}}]\big\|_2$ is less than one, then the upper bound holds trivially as $\lambda_2 \geq 1$; otherwise, this quantity is less than its squared and, hence, the upper-bound holds.

As a result of the bounds in (22), (23), and (26) we obtain that

$$\begin{aligned}
4L_{0-1}(\mathcal{M}_{\mathcal{J}}) &\leq 2 - 2\|B_{\mathcal{J}}\|_1 + \lambda_1 + \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 + 2\lambda_2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2 \\
&= 2 - 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1 + \Big(\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1 - \|B_{\mathcal{J}}\|_1\Big) + \lambda_1 + \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 + 2\lambda_2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2 \\
&\leq 2 - 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1 + \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2 + \lambda_1 + \Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2^2 + 2\lambda_2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2,
\end{aligned}$$

where the last inequality is due to (25). Therefore, from the definition of λ_1 and λ_2, and the function U in the statement of the lemma, we obtain

$$4L_{0-1}(\mathcal{M}_{\mathcal{J}}) \leq 2 - 2\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}}\Big\|_1 + 4U\Big(\Big\|F_{\bar{Y}}^{\subseteq \mathcal{J}} - B_{\mathcal{J}}\Big\|_2\Big).$$

This completes the proof.