Quantum Information and Computation, Vol. 16, No. 7&8 (2016) 0615–0656

THE LEARNABILITY OF UNKNOWN QUANTUM MEASUREMENTS

HAO-CHUNG CHENG
Graduate Institute Communication Engineering, National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.) &
Centre for Quantum Computation & Intelligent Systems (QCIS),
Faculty of Engineering and Information Technology, University of Technology Sydney
85 Broadway, Ultimo NSW 2007, Australia

MIN-HSIU HSIEH
Centre for Quantum Computation & Intelligent Systems (QCIS),
Faculty of Engineering and Information Technology, University of Technology Sydney
85 Broadway, Ultimo NSW 2007, Australia

PING-CHENG YEH
Graduate Institute Communication Engineering, National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.)

Received (October 2, 2015)
Revised (February 19, 2016)

In this work, we provide an elegant framework to analyze learning matrices in the Schatten class by taking advantage of a recently developed methodology—matrix concentration inequalities. We establish the fat-shattering dimension, Rademacher/Gaussian complexity, and the entropy number of learning bounded operators and trace class operators. By characterising the tasks of learning quantum states and two-outcome quantum measurements into learning matrices in the Schatten-1 and ∞ classes, our proposed approach directly solves the sample complexity problems of learning quantum states and quantum measurements.

Our main result in the paper is that, for learning an unknown quantum measurement, the upper bound, given by the fat-shattering dimension, is linearly proportional to the dimension of the underlying Hilbert space. Learning an unknown quantum state becomes a dual problem to ours, and as a byproduct, we can recover Aaronson's famous result [*Proc. R. Soc. A* ***463****, 3089–3144 (2007)*] solely using a classical machine learning technique. In addition, other famous complexity measures like covering numbers and Rademacher/Gaussian complexities are derived explicitly under the same framework. We are able to connect measures of sample complexity with various areas in quantum information science, e.g. quantum state/measurement tomography, quantum state discrimination and quantum random access codes, which may be of independent interest. Lastly, with the assistance of general Bloch-sphere representation, we show that learning quantum measurements/states can be mathematically formulated as a neural network. Consequently, classical ML algorithms can be applied to efficiently accomplish the two quantum learning tasks.

Keywords: Quantum Machine Learning, Sample Complexity, Quantum Tomography, Matrix Concentration Inequalities

Communicated by: R Cleve & A Harrow

1. Introduction

Statistical learning theory [1, 2] or *Machine Learning* (ML) [3] is a branch of artificial intelligence which aims to devise algorithms for machines to systematically learn from historic data. Typically, ML has been separated into *unsupervised learning* and *supervised learning*. In unsupervised learning, the machine is most useful for finding the hidden structure, e.g. clustering or density estimation, within unlabeled data. In supervised learning, the machine is equipped with more power to predict the class or to infer the characteristics from the structured data. The figures of merit for a learning machine include: (i) *computational complexity* which measures the run-time efficiency of a learning algorithm; (ii) *sample complexity* which determines the number of queries to a membership made by the learning algorithm such that the hypothesis function is Probably Approximately Correct (PAC) [4]; and (iii) *model complexity* (otherwise called the generalisation error [5]) which is defined as the discrepancy between the out-of-sample error and the in-sample error. Note that the model complexity is closely related to the sample complexity in the sense that a learning machine with large model complexity requires more samples to accurately approximate the target function, which results in high sample complexity. Current research trends include the reduction of computational complexity due a large volume data set (big data) as well as the high dimensional features of each data point, and how to balance the model complexity with the in-sample error such that the training data set can be trained well without the occurrence of overfitting.

To appropriately estimate the *sample complexity* of the hypothesis space, the most plausible quantity to measure the sample complexity of learning Boolean functions is the *Vapnik-*

Chervonenkis (VC) dimension [6]. Later on, complexity measures such as *fat-shattering dimensions* [7], *covering numbers* [8], and *Rademacher complexities* [9] are introduced to generalise the VC dimension for real-valued functions. To appropriately upper bound the complexity measures, Gurvits *et al.* [10, 11, 12] proposed a probabilistic approach for the class of linear functionals defined on Euclidean space:

$$\mathcal{F} = \left\{ x \mapsto w \cdot x : \|w\| \leq 1,\, x, w \in \mathbb{R}^d \right\}.$$

This method has been successfully applied to analyze sample complexities of the celebrated *support vector machines* (SVM) and large-margin classifiers in ML.

In this work, we extend Gurvits' work to consider learning linear functionals defined on matrix spaces:

$$\mathcal{F} = \left\{ X \mapsto \langle W, X \rangle : \|W\| \leq 1,\, X, W \in \mathbb{C}^{d \times d} \right\},$$

where $\langle \cdot, \cdot \rangle$ denotes the Hilbert-Schmidt inner product. Our novelty is that we adopt a powerful methodology—*matrix concentration inequalities* (MCIs) [13]—to derive sample complexity measures for learning matrices with norm constraints. The major advantage of using MCIs is that the method generalises standard statistical tools of learning real-valued functions defined on Euclidean space $\mathbb{R}^d$ so that we can directly generalise Gurvits' probabilistic approach to matrix spaces. Hence, complexity measures, such as the fat-shattering dimensions, Rademacher/Gaussian complexities, and entropy numbers, can be derived in a simple and elegant way.

The matrix learning problem has a strong connection with quantum information science. Quantum information science (QIS) is an active field that studies the computational capability in quantum systems. In recent years, QIS has achieved significant breakthroughs: factorizing large prime integers with an exponential speed-up and searching an unstructured database with a quadratic speed-up are the two most famous examples. Owing to the successful achievements of QIS, researchers have begun to explore whether QIS can advance other subjects of classical computer science. Consequently, the interdisciplinary area of quantum machine learning has attracted substantial interest lately. For example, *quantum tomography* is an essential task in physics for inferring the state of a quantum system or the measurement apparatus. When the target is to identify the preparation of a quantum system (resp. measurement instrument), it is called quantum state (resp. measurement) tomography. In large quantum systems, tomography is fundamentally difficult and practically infeasible because the number of parameters for describing the quantum system grows exponentially with the size of the system. Aaronson first pointed out that performing quantum state tomography in the ML setting can be exponentially efficient in the number of measurements [14] (we compare these two schemes in Section 1.2). Aaronson's method mainly relies on the entropic inequalities in quantum random access coding [15]. In this work, the proposed matrix learning framework covers Aaronson's result and shows that the problem of learning quantum states can be embedded into learning matrices with Schatten 1-norms (i.e. trace-class operators). Moreover, we push further to investigate learning quantum measurements and the connections with other areas of QIS.

The key to a successful development of efficient learning algorithms for a certain hypothesis class relies on finding an efficient representation for it. In this paper, we consider learning hypothesis classes that consist of normalized or subnormalized positive semi-definite matrices

of finite dimensions. Using the Bloch sphere representations [16, 17], we can transform the Hilbert-Schmidt inner product of two matrices into a standard inner product of two vectors in Euclidean space, and show that its resulting form has the structure of a neural network. This not only allows us to apply existing neural network algorithms to efficiently learn the hypothesis class, but also provides an elegant paradigm for the problems of matrix recovering.

1.1. *Contributions of this work*

Let $S_p^d = \{M \in \mathbb{C}^{d\times d} : \|M\|_p \leq 1\}$, where $\|\cdot\|_p$ is the Schatten p-norm, be a unit ball. We aim to learn an unknown matrix element $W \in S_q^d$ given the training data set $\{X_i, \langle W, X_i\rangle\}_{X_i\in S_p^d}$, where $1/p + 1/q = 1$.

Our results are

- We obtain major complexity measures for learning Schatten 1-norm and Schatten ∞-norm matrices in Table 1 (see Section 3). We show that the sample complexity of learning matrices in S_∞^d is proportional to the dimension d, while the sample complexity of learning elements in S_1^d is logarithmically proportional to d.

Table 1. Complexity Measures of Matrix Learning with Norm Constraints.

	Learning $\mathcal{X} = S_1^d$	Learning $\mathcal{X} = S_\infty^d$
Pseudo-Dimension	d^2	$d^2 - 1$
Fat-Shattering Dimension	d/ϵ^2	$\log d/\epsilon^2$
Uniform Entropy Number	d/ϵ^2	$\log d/\epsilon^2$
Rademacher/Gaussian Complexity	$\sqrt{d}$	$\sqrt{\log d}$
Sample Complexity $m_{\mathcal{F}}(\epsilon, \delta)$	$\max\{d, \log(1/\delta)\}/\epsilon^2$	$\max\{\log d, \log(1/\delta)\}/\epsilon^2$

- We show that the theoretical outcomes of matrix learning problems can answer important questions in quantum information science in Section 4 and 5. Firstly, learning Schatten 1-norm and Schatten ∞-norm matrices correspond to quantum state and measurement tomography. Thus the sample complexities derived in Table 1 provide theoretical upper bounds for these quantum tomographic tasks. Moreover, some of the complexity measures are directly related to problems in quantum set discrimination and quantum random access coding.

- We propose an efficient neural network formulation for learning matrices with norm constraints based on the Bloch sphere representations, and present numerical studies for several cases in Section 6.

There are several fields that may relate to or benefit from our work.

Quantum State/Measurement Tomography. Quantum state tomography is a difficult task in physics because the number of unknown parameters in a multi-partite quantum system grows exponentially. Aaronson pointed out that quantum ML can serve as an alternative approach to quantum state tomography [14]. Surprisingly, learning an unknown target state within a given accuracy requires only the number of measurements that grows logarithmically with the dimension d. In this work, we push Aaronson's result one step further and consider

the application of machine learning framework to study quantum measurement tomography. To the best of our knowledge, there are very few results in this direction. We hope that our result in learning quantum measurements will stimulate further investigation into this problem.

Quantum State Discrimination. The goal of quantum state discrimination is to identify a state in an ensemble. Whenever states are not mutually orthogonal, they cannot be perfectly discriminated. Therefore, a possible way is ambiguous state discrimination with the goal of minimizing the error probability. Given an $\epsilon > 0$, we show that the fat-shattering dimension guarantees the maximum number of quantum states that can be discriminated into two subsets with the worst error probability no greater than $1/2 - \epsilon$. Following the same reasoning, the quantum states in the hypothesis set can be used to distinguish between two-outcome measurements.

Quantum Random Access Coding. The (n, m, p)-quantum random access (QRA) coding stands for encoding an n-bit sequence into m-qubit so that the receiver can recover any one of the bits with successful probability at least p. The information-theoretic inequalities of n and m provide an upper bound for the fat-shattering dimension of learning quantum states. Alternatively, we can use the complexity measure—pseudo dimension—to show that there exists no (n, m, p)-QRA coding scheme, with $n \geq 2^{2m}$. The result coincides with the work of Hayashi *et al.* [18]. See Section 5.4 for further discussions.

1.2. *Comparisons between the Learning Setup and Quantum Tomography*

In the paradigm of learning an unknown quantum state ρ, the set of two-outcome measurements $(E_1, \ldots, E_n)$ are generated from an unknown distribution μ with the corresponding outcome statistics $(\mathrm{Tr}(E_1\rho), \ldots, \mathrm{Tr}(E_n\rho))$. The learning algorithm will produce a hypothesis state σ such that the *in-sample error*

$$\widehat{L}_n(\sigma) = \frac{1}{n}\sum_{i=1}^{n} \left(|\mathrm{Tr}(E_i\sigma) - \mathrm{Tr}(E_i\rho)|\right)$$

is minimized. The sample complexity of learning quantum states is the smallest number satisfying:

$$\Pr\left\{\sup_{\sigma} \left|L(\sigma) - \widehat{L}_n(\sigma)\right| \geq \epsilon\right\} \leq \delta,$$

where $L(\sigma) = \mathbb{E}|\,\mathrm{Tr}(E\sigma) - \mathrm{Tr}(E\rho)|$ is the *out-sample error*, and ϵ, δ are the accuracy and confidence respectively.

On the other hand, in the scheme of quantum state tomography, a series of quantum measurements (e.g. Pauli matrices) are designed and then performed on the unknown state ρ. Hence the hypothesis state σ is determined according to the measurement outcomes such that the distance measure, e.g. the trace distance $\frac{1}{2}\|\sigma - \rho\|_1 \leq \epsilon$, is within a certain level. We list the differences between the quantum learning setup and quantum state tomography in Table 2.

1.3. *Related Works*

Table 2. The comparison between learning quantum states and quantum state tomography.

	Learning Quantum States	**Quantum State Tomography**
Measurements	randomly generated	designed and deterministic
Distortion Measure	$\lvert L(\sigma) - \widehat{L}_n(\sigma)\rvert$	e.g. $\frac{1}{2}\lVert\sigma - \rho\rVert_1$
Other Assumptions	outcome statistics	finite copies of the unknown state

Our work is closely related with the problems of matrix recovering and learning matrices with norm constraints. This research topic has recently attracted substantial attention in ML with an increasing number of statistical tasks that organize data into matrices. The sample complexity of the matrix learning problem was addressed by [19], where the authors derive the Rademacher complexities of learning Schatten p-norm matrices using the techniques of *strong convexity duality*. In this paper, the proposed method with MCIs not only recovers the Rademacher complexities, but also solves fat-shattering dimensions and entropy numbers in the same framework. Furthermore, our approach is more general and can attack problems of the matrix space with certain structures. For example, our upper bound will be improved if the considering matrix is low rank or has small *intrinsic dimensions* [20]. The subset of the matrices with norm constraints such as positive partial transpose (PPT) states and separable states might also be treated in the same way. We leave this problem as future work.

The interdisciplinary area of quantum machine learning [21, 22] has attracted substantial interest lately. The central problems are two-fold. The first kind of problem investigates how quantum information processing can improve or accelerate classical ML tasks by converting classical algorithms into quantum regime [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]. On the other hand, certain fundamental quantum problems, such as inferencing an unknown quantum states or operations, or discovering the hidden structure of the underlying quantum system, can be assisted with ML techniques [42, 14, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 37, 40, 54, 55, 56, 57, 58, 59, 60]. In this work, we start from a machine learning point of view to formalize the problems of learning quantum measurements and quantum states as learning real-valued functions on matrix spaces. Hence, the sample complexities of these two learning problems are solved.

1.4. *Notation*

In this paper, we denote a Hilbert space by $\mathcal{H}$. The trace of an operator M on $\mathcal{H}$ is calculated as

$$\mathrm{Tr}(M) := \sum_k e_k M e_k,$$

where $\{e_k\}$ is any orthonormal basis on $\mathcal{H}$. Let $\mathbb{M}_d$ denote the set of all self-adjoint operators on $\mathbb{C}^d$. The Hilbert-Schmidt inner product on $\mathbb{M}_d$ can be defined as $\langle A, B\rangle_{\mathrm{HS}} := \mathrm{Tr}(AB)$, where the subscript 'HS' will be omitted when the context is clear. For $p \in [1, \infty)$, we denote the Schatten p-norm of an self-adjoint operator M as

$$\lVert M\rVert_p := \left(\sum_{i\geq 1} |\lambda_i(M)|^p\right)^{1/p},$$

where $\lambda_i(M)$ is the eigenvalue of M. We denote by $\|M\|_\infty := \sup_i |\lambda_i(M)|$ the operator norm. Clearly, $\|\cdot\|_1$ and $\|\cdot\|_2$ correspond to the trace norm and the Hilbert-Schmidt norm $\|\cdot\|_{\mathrm{HS}}$ respectively. Slightly abusing the notation, we also denote the conventional ℓ_p norm on $\mathbb{R}^d$ by $\|\cdot\|_p$ for $p \in [1,\infty]$. We define the unit ball associated with the Schatten norms as $S_p^d = \{M \in \mathbb{M}_d : \|M\|_p \leq 1\}$. The set of bounded operators on $\mathcal{H}$ is denoted as $\mathcal{B}(\mathcal{H})$, which is the set operators with finite Schatten ∞-norm. Likewise, the set of operators with finite Schatten 1-norm is called the set of trace class operators, $\mathcal{T}(\mathcal{H})$.

A *quantum state* (also called *density operators*) on the Hilbert space $\mathcal{H}$ is a positive semi-definite operator with unit trace. We identify the *state space* as the set of all quantum states on $\mathcal{H}$, i.e.

$$\mathcal{Q}(\mathcal{H}) := \{\rho \in \mathcal{T}(\mathcal{H}) : \rho \succeq 0,\ \mathrm{Tr}(\rho) = 1\}.$$

A positive operator-valued measure (POVM) on $\mathcal{H}$ is a finite set of positive semi-definite operators $\{\Pi_i\}_{i\in I}$ such that

$$\sum_{i\in I} \Pi_i = \mathcal{I},$$

where $\mathcal{I}$ denotes the identity operator on $\mathcal{H}$. Each POVM element Π_i is called a *quantum effect*, which serves as an instrument to perform a yes-no measurement. We denote the set of all effects as an *effect space*:

$$\mathcal{E}(\mathcal{H}) := \{E \in \mathcal{B}(\mathcal{H}) : \mathcal{O} \preceq E \preceq \mathcal{I}\}.$$

All constants are denoted as C or c and are independent from other parameters. Their values may change from line to line. The notation $A \lesssim B$ means there is a constant c such that $A \leq cB$ and $A \simeq B$ means both $A \lesssim B$ and $A \gtrsim B$. We summarise all the notation in table A.1 in Appendix 1.

The paper is organized as follows. In Section 2 we introduce the background of statistical learning theory (especially on supervised learning) and describe important complexity measures. In Section 3, we formalize a unified framework to relate the problems of learning quantum measurements and learning quantum states with the learning real-valued functions. Based on the proposed approach, we prove the main results of learning quantum measurements in Section 4. In addition, we discuss the interpretations of the learning tasks to ambiguous set discrimination and also derive the covering numbers and the Rademacher complexity. In Section 5, we consider the problem of learning quantum states and describe its relationship with QRA codes. In Section 6, we formulate the learning problem into Bloch-sphere representation and propose possible algorithms (e.g. neural networks) to implement the quantum learning tasks. We conclude this paper in Section 7.

2. Background of Statistical Learning Theory

The starting point of this section is the mathematical formalism of the *supervised machine learning*. We describe the efficiency of a learning machine and examine the number of samples required to produce an almost optimal function with an error rate below the desired accuracy. As will be shown later, the bound of the sample complexity is closely related to the complexity measures which characterise the "effective size" of a function class.

2.1. *Supervised Machine Learning*

Generally speaking, supervised learning is a ML task that infers a function (or a learning model) by observing the data and the response to the data. In this work, we focus on the definitions of agnostic PAC learnability and the sample complexity for supervised machine learning. For more comprehensive introduction to ML, we refer the readers to literature such as Refs. [61, 2, 62, 63, 64, 65, 66].

Consider a probability space $(\mathcal{Z}, \mu)$, where $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$ (called the *input space*) a measurable space and $\mathcal{Y}$ (called the *output space*) a closed subset of real line $\mathbb{R}$. The probability distribution μ over $\mathcal{Z}$ is assumed to be fixed but known only through the *training data set*, i.e. $\mathsf{Z}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \in \mathcal{Z}^n$ sampled independently and identically according to the measure μ. Supervised learning aims to construct a function $f : \mathcal{X} \to \mathcal{Y}$ which approximates the functional relationship between the input variable $X \in \mathcal{X}$ and the output variable $Y \in \mathcal{Y}$ from the observed training data set. To evaluate the performance of the approximation, we define the loss function as a measurable map $\ell_f : \mathcal{Z} \to [0, +\infty)$ and the *expected risk* (also called the *out-of-sample error*):

$$L(f) = \mathbb{E}_\mu \ell_f(X, Y).$$

The loss function is usually taken as the absolute error or square error, i.e.

$$\ell_f(X, Y) = |f(X) - Y| \quad \text{or} \quad \ell_f(X, Y) = (f(X) - Y)^2.$$

For convenience, we only consider the square error in this work. Other loss functions that satisfy the Lipschitz condition can be easily generalised.*

Since we are interested in minimizing the expected risk, hence the *target function* (or *Bayes function*) as $t(x) = \mathbb{E}[Y|X = x]$ can be defined to achieve the minimum expected risk (called the *Bayes risk*), i.e.

$$L_{\text{Bayes}} := L(t) = \inf_f L(f), \tag{1}$$

where the infimum is taken over all possible measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. When y is a deterministic function of X, then $Y = t(X)$ almost surely and $L(t) = 0$.

The goal of the learner is to identify the target function t from a collection of functions $\mathcal{F}$, called the *hypothesis set*,† which is a set of real-valued functions defined on the input space

* A loss function $\ell_f : \mathcal{Z} \to (0, \infty)$ is a Lipschitz function if it satisfies the Lipschitz condition

$$|\ell_f(X, Y) - \ell_g(X, Y)| \leq \mathsf{L}|f(X) - g(X)|$$

for all possible $(X, Y) \in \mathcal{Z}$ and the quantity $\mathsf{L} \in \mathbb{R}$ is called the Lipschitz constant. Denote by $\ell_{\mathcal{F}}$ the set $\{\ell_f : f \in \mathcal{F}\}$. Then the complexity measures (e.g. the covering number and Rademacher complexity; see Definition 6, 7, and 8) of the class $\ell_{\mathcal{F}}$ are different from that of the hypothesis set $\mathcal{F}$ by the Lipschitz constant L [67, 9], i.e.

$$\mathcal{N}_p(\epsilon, \ell_{\mathcal{F}}, m) \leq \mathcal{N}_p(\epsilon/\mathsf{L}, \mathcal{F}, m) \quad \text{for } p \geq 1,\ m \in \mathbb{N}$$

and

$$\mathcal{R}_n(\ell_{\mathcal{F}}) \leq \mathsf{L}\mathcal{R}_n(\mathcal{F}).$$

Therefore, by homogeneity we may assume the loss function is the absolute error with $\mathsf{L} = 1$ or the square error $\mathsf{L} = 2$ for deriving the sample complexity problems.

† Note that we use the term 'hypothesis set' and 'function class' interchangeably throughout the paper.

$\mathcal{X}$. A *learning algorithm* A for hypothesis set $\mathcal{F}$ is a mapping that assigns to every training data Z_n some candidate function $\mathsf{A}(\mathsf{Z}_n) \in \mathcal{F}$, i.e.

$$\mathsf{A} : \cup_{n=1}^{\infty} \mathsf{Z}^n \to \mathcal{F}.$$

The effectiveness of the learning algorithm is measured by the number of data required to produce an optimal function with the minimum expected risk, see Eq. (1). Therefore, we introduce one of the most fundamental concepts in supervised machine learning—*Agnostic Probably Approximately Correct (PAC)* learning model [4, 68]:

Definition 1 (Agnostic PAC Learnability [66], Def. 3.3) *A hypothesis set $\mathcal{F}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{F}} : \mathbb{R} \times \mathbb{R} \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution μ over $\mathcal{Z}$, when running the learning algorithm on $n \geq m_{\mathcal{F}}(\epsilon, \delta)$ samples generated by μ, the algorithm returns a hypothesis $\hat{f}$ such that, with probability of at least $1-\delta$ (over the choice of the n training examples),*

$$L(\hat{f}) \leq \inf_{f \in \mathcal{F}} L(f) + \epsilon.$$

However, the expected risk $L(f) = \mathbb{E}_{\mu}[\ell_f(X, Y)]$ cannot be calculated since the measure μ is unknown. We can only evaluate the agreement of a candidate function over the training data set, which is called the *empirical risk* (also called the *in-sample error*):

$$\widehat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_f(X_i, Y_i).$$

For example, one of the most well-known learning algorithms is the Empirical Risk Minimization (ERM) principle [2] that assigns a function $f_n \in \mathcal{F}$ to each training data set which is "almost optimal" on the data, i.e.

$$f_n = \arg\min_{f \in \mathcal{F}} \widehat{L}_n(f). \tag{2}$$

One way to evaluate the performance of the learning algorithm is to relate the risk $L(f_n)$ to the empirical risk $\widehat{L}_n(f_n)$. Following the reasoning of agnostic PAC model, our goal is hence to estimate the *generalisation error* ϵ:

$$L(f_n) \leq \widehat{L}_n(f_n) + \epsilon(n, \mathcal{F}).$$

For any algorithm that outputs a $f_n \in \mathcal{F}$, we have

$$L(f_n) - \widehat{L}_n(f_n) \leq \sup_{f \in \mathcal{F}} \{L(f) - \widehat{L}_n(f)\},$$

which leads to the definition of *uniform Glivenko-Cantelli class* (uGC class).

Definition 2 *We say that the hypothesis set $\mathcal{F}$ is a uniform Glivenko-Cantelli class if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} \sup_{\mu} \Pr \left\{ \sup_{f \in \mathcal{F}} \left| L(f) - \widehat{L}_n(f) \right| \geq \epsilon \right\} = 0.$$

The uniformity is with respect to all members of $\mathcal{F}$ and over all possible probability measures μ on the domain $\mathcal{Z}$. In addition to the conditions of the learnability, we also consider the bound on the rate of uniform convergence. For every $0 < \epsilon, \delta < 1$, let $m_{\mathcal{F}}(\epsilon, \delta)$ be the first integer such that for every $n \geq m_{\mathcal{F}}(\epsilon, \delta)$ and any probability measure μ,

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|L(f) - \widehat{L}_n(f)\right| \geq \epsilon\right\} \leq \delta. \tag{3}$$

The quantity $m_{\mathcal{F}}(\epsilon, \delta)$ satisfied Eq. (3) is called the (Glivenko-Cantelli) *sample complexity* of the hypothesis set $\mathcal{F}$ with accuracy ϵ and confidence δ. The sample complexity encapsulates the number of samples required to learn a set of functions.

Vapnik studied the relation between the uGC class and learnability [69, 1, 2] and showed that if a hypothesis set $\mathcal{F}$ is a uGC class, then it is sufficient for the agnostic PAC learnability[‡].

Theorem 1 (Uniform Convergence [66, Corollary 4.4]) *A training data set* Z_n *is called* ϵ*-representative (with respect to domain* $\mathcal{Z}$*, hypothesis set* $\mathcal{F}$*, loss function* ℓ*, and distribution* μ*) if*

$$\forall f \in \mathcal{F}, \quad \left|\widehat{L}_n(f) - L(f)\right| \leq \epsilon.$$

Then, for every $\epsilon, \delta \in (0, 1)$ *and every probability distribution* μ *over* $\mathcal{Z}$*, a uGC class* $\mathcal{F}$ *that guarantees an* $\epsilon/2$*-representative set with probability of at least* $1-\delta$ *is agnostic PAC learnable. Furthermore, the ERM algorithm is an agnostic PAC learner for* $\mathcal{F}$*.*

As a result, we consider the generalisation error $\epsilon(n, \mathcal{F})$ and the sample complexity $m_{\mathcal{F}}(\epsilon, \delta)$ of the hypothesis set $\mathcal{F}$ as the performance criterion to investigate whether the underlying learning problem is agnostic PAC learnability.

In summary, the fundamental problems in ML are two-fold. The first is under what conditions the machine is agnostic PAC learnable. Secondly, the sample complexity determines the rate of the uniform convergence and the information-theoretic efficiency of the hypothesis set $\mathcal{F}$. In the next subsection, several complexity measures are introduced to characterise the "richness" or "effective size" of the hypothesis set. In Section 2.3, we show that the sample complexity can be further expressed in terms of the complexity measures.

2.2. *Measures of Sample Complexity*

As discussed before, we are interested in the parameters which effectively measure the size of a given hypothesis set. There are some well-known measures of the (information) complexity[§] of function classes: *combinatorial parameters*, *covering numbers*, and *Rademacher complexity*.

The first combinatorial parameter—*Vapnik-Chervonenkis (VC) dimension*—was introduced by Vapnik and Chervonenkis [6] for learning Boolean functions.

[‡] Agnostic PAC learnable is also called *learnable with ERM*, or we can say that the ERM algorithm is *consistent*. Recent works consider the *stability* issues of the learning algorithm as one of the criterion of learnability. However, in this paper we do not deal with issues of stability and hence refer interested readers to Refs. [70, 71] and the references therein.

[§] The complexity measures introduced in this section and the generalisation bounds derived in Section 2.3 are information-theoretic in the sense that the learning algorithms are based on the agnostic PAC model regardless of the computational resources.

Definition 3 (VC Dimension) *Let $\mathcal{F}$ be a set of $\{0,1\}$-valued functions on a domain $\mathcal{X}$. We say that $\mathcal{F}$ shatters a set $\{x_1,\ldots,x_n\}\subseteq\mathcal{X}$ if for every subset $B\subseteq\{1,\ldots,n\}$ there exists a function $f_B\in\mathcal{F}$ for which $f_B(x_i)=1$ if $i\in B$, and $f_B(x_i)=0$ if $i\notin B$. Let*

$$\mathrm{VCdim}(\mathcal{F})=\sup\{|\mathcal{S}|:\mathcal{S}\subseteq\mathcal{X},\,\mathcal{S}\text{ is shattered by }\mathcal{F}\}.$$

The VC dimension of $\mathcal{F}$ (on the domain $\mathcal{X}$) is denoted as VCdim($\mathcal{F}$).

Pollard [72] generalised the concept of VC dimension and introduced the *pseudo dimension* to quantify the sample complexity of a real-valued function class. The parameterised version of Pollard's pseudo-dimension is the *scale-sensitive dimension* (also called the *fat-shattering dimension*) introduced by Kearns and Schapire [73].

Definition 4 (Pseudo Dimension) *Let $\mathcal{F}$ be a set of real-valued functions on a domain $\mathcal{X}$. We say a set $\mathcal{S}=\{x_1,\ldots,x_n\}\subseteq\mathcal{X}$ is pseudo-shattered by $\mathcal{F}$ if there exists a set $\{\alpha_i\}_{i=1}^n$ such that for every $B\subseteq\{1,\ldots,n\}$ there is some function $f_B\in\mathcal{F}$ for which $f_B(x_i)\geq\alpha_i$ if $i\in B$, and $f_B(x_i)<\alpha_i$ if $i\notin B$. Define the pseudo dimension of $\mathcal{F}$ as*

$$\mathrm{Pdim}(\mathcal{F})=\sup\{|\mathcal{S}|:\mathcal{S}\subseteq\mathcal{X},\,\mathcal{S}\text{ is pseudo-shattered by }\mathcal{F}\}.$$

f_B is called the shattering function of the set $\mathcal{S}$.

There is a desirable property of the pseudo dimension that will be useful in our main theorems.

Theorem 2 (Pollard [72])

(i) If $\mathcal{F}$ is a vector space of real-valued functions then $\mathrm{Pdim}(\mathcal{F})=\dim(\mathcal{F})$.

(ii) If $\mathcal{F}$ is a subset of a vector space $\mathcal{F}'$ of real-valued functions then $\mathrm{Pdim}(\mathcal{F})\leq\dim(\mathcal{F}')$.

Definition 5 (Fat-Shattering Dimension) *Let $\mathcal{F}$ be a set of real-valued functions on a domain $\mathcal{X}$. For every $\epsilon>0$, a set $\mathcal{S}=\{x_1,\ldots,x_n\}\subseteq\mathcal{X}$ is said to be ϵ-shattered by the $\mathcal{F}$ if there exists a set $\{\alpha_i\}_{i=1}^n\subset\mathbb{R}$ such that for every $B\subseteq\{1,\ldots,n\}$ there is some function $f_B\in\mathcal{F}$ for which $f_B(x_i)\geq\alpha_i+\epsilon$ if $i\in B$, and $f_B(x_i)<\alpha_i-\epsilon$ if $i\notin B$. Define the fat-shattering dimension of $\mathcal{F}$ on the domain $\mathcal{X}$ as*

$$\mathrm{fat}_{\mathcal{F}}(\epsilon,\mathcal{X})=\sup\{|\mathcal{S}|:\mathcal{S}\subseteq\mathcal{X},\,\mathcal{S}\text{ is }\epsilon\text{-shattered by }\mathcal{F}\}.$$

f_B is called the shattering function of the set B and the set $\{\alpha_i\}_{i=1}^n$ is called a witness to the ϵ-shattering. When the underlying space is clear, we denote it by $\mathrm{fat}_{\mathcal{F}}(\epsilon)$. If the witness set $\{\alpha_i\}$ are all equal to a constant, we call it as the level fat-shattering dimension, $\underline{\mathrm{fat}}_{\mathcal{F}}(\epsilon)$.

In Ref. [62], a relationship between the fat-shattering dimension and the pseudo-dimension can be given.

Theorem 3 (Anthony and Bartlett [62, Theorem 11.13]) *Let $\mathcal{F}$ be a set of real-valued functions. Then:*

(i) For all $\epsilon>0$, $\mathrm{fat}_{\mathcal{F}}(\epsilon)\leq\mathrm{Pdim}(\mathcal{F})$.

(ii) If a finite set $\mathcal{S}$ is pseudo-shattered then there is ϵ_0 such that for all $\epsilon<\epsilon_0$, $\mathcal{S}$ is ϵ-shattered.

(iii) *The function* $\mathrm{fat}_{\mathcal{F}}(\epsilon)$ *is non-increasing with* ϵ.

(iv) $\mathrm{Pdim}(\mathcal{F}) = \lim_{\epsilon \downarrow 0} \mathrm{fat}_{\mathcal{F}}(\epsilon)$ *(where both sides may be infinite).*

Note that it is possible for the pseudo-dimension to be infinite, even when the fat-shattering dimension is finite for all positive ϵ.

In addition to the combinatorial parameters bounding the sample complexity, there are other quantities called *covering number* which measure the size of the function class by the finite approximating set. The concept of covering number dates back to Kolmogorov *et al.* [8] and has been used in many areas of mathematics.

Definition 6 (Covering Number) *Let* $(\mathfrak{M}, d)$ *be a metric space and let* $\mathcal{F} \subset \mathfrak{M}$. *For every* $\epsilon > 0$, *the set* $\{y_1, \ldots, y_n\}$ *is called an* ϵ*-cover of* $\mathcal{F}$ *if every* $f \in \mathcal{F}$ *has some* y_i *such that* $d(f, y_i) < \epsilon$. *The covering number* $\mathcal{N}(\epsilon, \mathcal{F}, \tau)$ *is the minimum cardinality of a* ϵ*-covering set for* $\mathcal{F}$ *with respect to the metric* τ.

To characterise the size of the function class $\mathcal{F}$ in machine learning, we are interested in the metrics endowed by the samples; for every sample $\{x_1, \ldots, x_n\} \in \mathcal{X}$, let $\mu_n = n^{-1} \sum_{i=1}^{n} \delta_{x_i}$ be the empirical measure supported on that sample. For $1 \leq p < \infty$ and a function f, denote $\|f\|_{L_p(\mu_n)} = \left(n^{-1} \sum_{i=1}^{n} |f(x_i)|^p\right)^{1/p}$ and $\|f\|_{\infty} = \max_{1 \leq i \leq n} |f(x_i)|$. Then, $\mathcal{N}\left(\epsilon, \mathcal{F}, L_p(\mu_n)\right)$ is the covering number of $\mathcal{F}$ at scale ϵ with respect to the $L_p(\mu_n)$ norm.

Definition 7 (Entropy Number) *For every class* $\mathcal{F}$, $1 \leq p \leq \infty$ *and* $\epsilon > 0$, *let*

$$\mathcal{N}_p(\epsilon, \mathcal{F}, n) = \sup_{\mu_n} \mathcal{N}\left(\epsilon, \mathcal{F}, L_p(\mu_n)\right),$$

and

$$\mathcal{N}_p(\epsilon, \mathcal{F}) = \sup_{n} \sup_{\mu_n} \mathcal{N}\left(\epsilon, \mathcal{F}, L_p(\mu_n)\right).$$

We call $\log \mathcal{N}_p(\epsilon, \mathcal{F}, n)$ *the entropy number of* $\mathcal{F}$ *with respect to* $L_p(\mu_n)$ *and* $\log \mathcal{N}_p(\epsilon, \mathcal{F})$ *the uniform entropy number.*

Bartlett and Mendelson [9] considered the techniques of concentration of measures for empirical processes and proposed a random average quantity—*Rademacher complexity*, which capture the size of the uGC class more directly and leads to sharp complexity bounds.

Definition 8 (Rademacher Complexity[¶] [73, 9, 74]) *Let* μ *be a probability measure on* $\mathcal{X}$ *and* $\mathcal{F}$ *be a set of uniformly bounded functions on* $\mathcal{X}$. *For every positive integer* n, *define*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \gamma_i f(x_i) \right|,$$

where $\{x_i\}_{i=1}^{n}$ *are independent random variables distributed according to* μ *and* $\{\gamma_i\}_{i=1}^{n}$ *independently takes values in* $\{-1, +1\}$ *with equal probability (which are also independent of* $\{x_i\}_{i=1}^{n}$*). The quantity* $\mathcal{R}_n(\mathcal{F})$ *is called the Rademacher complexity associated with the class* $\mathcal{F}$.

[¶] Some authors define the Rademacher complexity with the normalization term as n rather than $\sqrt{n}$. Here we follow the notation used in Ref. [74], which is more convenient to bound the sample complexity (e.g. Eq. (7)).

We remark that the complexity measures can be related among each other [75, 76, 77]:

$$\mathrm{fat}_{\mathcal{F}}(\epsilon) \lesssim \log \mathcal{N}_2(\epsilon, \mathcal{F}, n) \lesssim \frac{\mathcal{R}_n^2(\mathcal{F})}{\epsilon^2} \lesssim \mathrm{fat}_{\mathcal{F}}(\epsilon) \cdot \log\left(\frac{1}{\epsilon}\right).$$

To sum up the results we have presented so far, the complexity measures, such as the combinatorial parameters (e.g. VC dimension and fat-shattering dimension), covering numbers and the Rademacher complexity of the hypothesis set control the rate of uniform convergence. By computing those quantities of the given hypothesis set and according to Eqs. (4), (5), (6) and (7) in Section 2.3, we can estimate the bounds on the sample complexity of the learning problems.

2.3. *Sample Complexity in Terms of Complexity Measure*

Previously, we introduce several complexity measures. In this section, we list some well-known deviation formula to express the generalisation error and sample complexity in terms of those complexity measures.

It has been established that any set of Boolean functions is a uGC class (i.e. PAC learnable) if and only if it has a finite VC dimension [78, 79]. Additionally, the finite VC dimension provides an upper bound for the sample complexity of the Boolean function class.

Theorem 4 (Vapnik *et al.* [78, 79, 9]) *Let C be an absolute constant and $\mathcal{F}$ be a class of Boolean functions which has a finite VC dimension d. Then, for every $0 < \epsilon, \delta < 1$,*

$$\sup_{\mu} \Pr\left\{\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_n(f)\right| \geq \epsilon\right\} \leq \delta,$$

provided that $n \geq \frac{C}{\epsilon^2}\left(d\log(2/\epsilon) + \log(2/\delta)\right)$.

Therefore, the sample complexity is bounded by

$$m_{\mathcal{F}}(\epsilon, \delta) \leq \frac{C}{\epsilon^2} \max\left\{d\log\frac{1}{\epsilon}, \log\frac{1}{\delta}\right\}. \tag{4}$$

Following the same reasoning as in Theorem 4, the analogous results can be drawn: the hypothesis set $\mathcal{F}$ is a uGC class if and only if it has a finite fat-shattering dimension for every $\epsilon > 0$ [80, 7, 77]. We have the following theorem:

Theorem 5 (Bartlett *et al.* [80, 7, 77]) *There is an absolute constant C such that for every $\mathcal{F}$ consisting of bounded functions and every $0 < \epsilon, \delta < 1$,*

$$\sup_{\mu} \Pr\left\{\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_n(f)\right| \geq \epsilon\right\} \leq \delta,$$

provided that $n \geq \frac{C}{\epsilon^2}\left(fat_{\mathcal{F}}(\epsilon/8)\cdot\log(2/\epsilon) + \log(8/\delta)\right)$.

Therefore, the sample complexity is bounded by

$$m_{\mathcal{F}}(\epsilon, \delta) \leq \frac{C}{\epsilon^2} \max\left\{\mathrm{fat}_{\mathcal{F}}(\epsilon)\cdot\log\frac{1}{\epsilon}, \log\frac{1}{\delta}\right\}. \tag{5}$$

The entropy number is distribution-independent and is closely related to the learnability of the function class. Dudley et al. [81] showed that a class $\mathcal{F}$ consisting of bounded functions is a uGC class if and only if that there is some $1 \leq p \leq \infty$ such that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{\log \mathcal{N}_p(\epsilon, \mathcal{F}, n)}{n} = 0.$$

In addition, we have the following theorem:

Theorem 6 (Polland [72]) *Let $\mathcal{F}$ be a set of bounded functions.*

(i) For every $0 < \epsilon < 1$, any $n \geq 8/\epsilon^2$, and any probability measure μ,

$$\Pr\left\{\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_n(f)\right| \geq \epsilon\right\} \leq 8\mathcal{N}_1(\epsilon/8, \mathcal{F}, n) \exp\left(-\frac{n\epsilon^2}{128}\right).$$

(ii) For every $0 < \epsilon, \delta < 1$,

$$\sup_{\mu} \Pr\left\{\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_n(f)\right| \geq \epsilon\right\} \leq \delta,$$

provided that $n \geq \frac{C}{\epsilon^2}\left(\log \mathcal{N}_1(\epsilon, \mathcal{F}) + \log(2/\delta)\right)$.

Therefore, the sample complexity is bounded by

$$m_{\mathcal{F}}(\epsilon, \delta) \leq \frac{C}{\epsilon^2} \max\left\{\mathcal{N}_1(\epsilon, \mathcal{F}), \log\frac{1}{\delta}\right\}. \tag{6}$$

Theorem 7 (Bartlett and Mendelson [9]) *For any $0 < \delta < 1$, with probability at least $1 - \delta$ and for all $f \in \mathcal{F}$ we have,*

$$\Pr\left\{\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_n(f)\right| \geq \epsilon\right\} \leq \delta,$$

provided that $n \geq \frac{C}{\epsilon^2} \max\{\mathcal{R}_n(\mathcal{F}), \log(1/\delta)\}$.

Therefore, the sample complexity is bounded by

$$m_{\mathcal{F}}(\epsilon, \delta) \leq \frac{C}{\epsilon^2} \max\left\{\mathcal{R}_n(\mathcal{F}), \log\frac{1}{\delta}\right\} \tag{7}$$

3. The Framework for Learning Matrices in Schatten Class and the Quantum Learning Model

This section proposes a framework of learning unknown matrix elements in the Schatten classes. Specifically, for every $1/p + 1/q = 1$, we aim to learn a target matrix $W \in S_q^d$ with the input $X \in S_p^d$ and the corresponding label

$$f_W : X \mapsto \langle W, X\rangle.$$

We connect the problem of learning matrices in the Schatten class with learning (real-valued) linear functionals on the input space in Section 3.1. Afterwards, we unify the two quantum learning problems at hand into learning linear functionals in Section 3.2. We also provide a justification of the the proposed quantum learning model in practical situations. The interested readers can refer to Appendix B.

3.1. *Learning Linear Functionals on Banach Space*

According to the duality theorem between bounded operators and trace class operators (see Theorem 8 below), we can identify the element in the Banach space as the membership in the dual space of the input space, i.e. linear functionals on the input space. For example, assume the input space is the unit ball of the Schatten p-class, i.e. $\mathcal{X} = S_p^d$. Then the hypothesis set can be represented as the linear functionals that are polar[‖] to S_p^d, i.e. for all $x \in S_p^d$ and $1/p + 1/q = 1$,

$$\mathcal{F} = \left\{ x \mapsto \langle E, x \rangle : E \in S_q^d \right\} = \left(S_p^d \right)^\circ .$$

Under this duality formalism, the problems of estimating the complexity measures of the subset in a Banach space can be transformed into the following question: Whether a set of linear functionals is agnostic PAC learnable?

Theorem 8 (Duality of Bounded Operator and Trace class [82, Thm. 19.1 & 19.2]) *Fix a Hilbert space $\mathcal{H}$. The map $E \mapsto f_E$ is an isometric isomorphism from the space of bounded operators, $\mathcal{B}(\mathcal{H})$, to the dual space of the set of trace classes operators, $\mathcal{T}(\mathcal{H})^*$. Conversely, the map $\rho \mapsto f_\rho$ is an isometric isomorphism from $\mathcal{T}(\mathcal{H})$ to $\mathcal{B}(\mathcal{H})^*$.*

Mendelson and Schechtman [83] first investigated the fat-shattering dimension of sets of linear functionals on Banach space and proposed the following useful result.

Lemma 1 (Mendelson and Schechtman [83, Coro. 3.2]) *The set $\mathcal{S} = \{x_1, \ldots, x_n\} \subset B_X$ is ϵ-shattered by B_{X^*} if and only if $\{x_i\}_{i=1}^n$ are linearly independent and for every $a_1, \ldots, a_n \in \mathbb{R}$,*

$$\epsilon \sum_{i=1}^n |a_i| \leq \left\| \sum_{i=1}^n a_i x_i \right\|_X ,$$

where B_X is the unit ball of some Banach space X and B_{X^*} is its dual unit ball.

By restricting the values of the set $\{a_i\}_{i=1}^n$ to $\{+1, -1\}$, the core idea of Lemma 1 is to calculate the *Rademacher series* on the Banach space, where the n points Rademacher series

[‖] In convex analysis, a *convex body* $K \subset \mathbb{R}^n$ is a convex compact set with nonempty interior. The *gauge* of a convex body K, also known as the *Minkowski functional*, is defined by $\|x\|_K := \inf\{t \geq 0 : x \in tK\}$. If K is symmetric with respect to the origin ($-K = K$), then K is a unit ball associated with the norm $\|\cdot\|_K$ and the inner product $\langle \cdot, \cdot \rangle$. We define the *polar* of K as

$$K^\circ = \left\{ x \in \mathbb{R}^n : \sup_{k \in K} \langle k, x \rangle \leq 1 \right\}.$$

In the symmetric case, K° is the unit ball of the dual space of $(\mathbb{R}^n, \|\cdot\|_K)$. Here, S_1^d is a unit ball of Schatten 1-class and S_∞^d is a unit ball of Schatten ∞-class. Considering the Hilbert-Schmidt inner product, S_1^d and S_∞^d are polar to each other.

on $\mathcal{X}$ is defined as $\sum_{i=1}^{n} \gamma_i x_i$, where $\{\gamma_i\}_{i=1}^{n}$ are the symmetric $\{+1, -1\}$-valued random variables. Additionally, with the following duality formula for the Schatten p-norm, we can estimate the range of the linear functional, which is helpful to further derive the complexity measures.

Theorem 9 (Duality Formula for $\|A\|_p$ [84, Theorem 7.1]) *For all $p \geq 1$, define q by $1/q + 1/p = 1$. Then for all $A \in \mathbb{M}_d$,*

$$\|A\|_p = \sup_{B \in \mathbb{M}_d} \{\mathrm{Tr}(BA) : \|B\|_q = 1\}.$$

The techniques from Mendelson and Schechtman (Lemma 1) and the duality formula (Theorem 9) can be used to upper bound the fat-shattering dimension and the Rademacher complexity via the Rademacher series. What remains is to compute the Rademacher series on the Banach space for both complexity measures, and we leave the details to Sections 4 and 5.

3.2. *The Quantum Learning Problem as learning Linear Functional on Matrices*

Recall that a physical theory aims to predict events observed in the experiments by describing three types of apparatus: preparation, transformation, and measurement. The preparation process of a system can be embodied by a state, while an *effect* is a measurement that produces either 'yes' or 'no' outcomes in order to observe the physical experiment. However, according to the statistical nature of Quantum Theory, only probabilities of the occurrence can be predicted (counting multiple measurements). More precisely, assume that a system is prepared in the state $\rho \in \mathcal{Q}(\mathcal{H})$. Then the outcome of every two-outcome measurement $E \in \mathcal{E}(\mathcal{H})$ takes the form of the probability distribution:

$$f_E(\rho) = \mathrm{Tr}(E\rho) = \langle E, \rho \rangle \in [0, 1].$$

Note that it is indeed a linear functional on the state space, i.e. $f_E : \mathcal{Q}(\mathcal{H}) \to \mathbb{R}$. In ML, such $[0, 1]$-valued functions are called *probabilistic concepts* [73].

The following proposition establishes the one-to-one correspondence between $f_E \leftrightarrow E$.

Proposition 1 ([85, Prop. 2.30]) *Given a Hilbert space $\mathcal{H}$, let f_E be an effect, i.e. a linear map from $\mathcal{Q}(\mathcal{H})$ to the interval $[0, 1]$. Then there exists a bounded operator $E \in \mathcal{E}(\mathcal{H})$ such that*

$$f_E(\rho) = \mathrm{Tr}(E\rho) = \langle E, \rho \rangle \quad \forall \rho \in \mathcal{Q}(\mathcal{H}).$$

Furthermore, the operator E is unique in the following sense. Let $E_1, E_2 \in \mathcal{E}(\mathcal{H})$. If $\langle \varphi, E_1 \varphi \rangle = \langle \varphi, E_2 \varphi \rangle$ for every $|\varphi\rangle \in \mathcal{H}$, then $E_1 = E_2$.

The proposition states that every two-outcome measurement can be identified as a linear functional on the state space. Consequently, the problem of learning an unknown (two-outcome) quantum measurement is equivalent to learning a real-valued linear functional on quantum states. Here and subsequently, we call an effect to represent either the linear functionals on $\mathcal{Q}(\mathcal{H})$ or the two-outcome measurement $E \in \mathcal{E}(\mathcal{H})$.

Conversely, if the measurement apparatus is chosen as some $E \in \mathcal{E}(\mathcal{H})$, then the measurement outcome of every state ρ is distributed as

$$f_\rho(E) = \mathrm{Tr}(\rho E) = \langle \rho, E \rangle \in [0, 1].$$

Therefore, we take the state space as the set of linear functionals on the effect space by the following proposition:

Proposition 2 ([86]) *Given a Hilbert space $\mathcal{H}$, let f_ρ be probability measure on $\mathcal{E}(\mathcal{H})$. Then there exists a quantum state $\rho \in \mathcal{Q}(\mathcal{H})$ such that*

$$f_\rho(E) = \mathrm{Tr}(\rho E) = \langle \rho, E \rangle \quad \forall E \in \mathcal{E}(\mathcal{H}).$$

Furthermore, different $\rho_1, \rho_2 \in \mathcal{Q}(\mathcal{H})$ determines different probability measures, i.e. there exists an operator $E \in \mathcal{E}(\mathcal{H})$ such that $\mathrm{Tr}(E\rho_1) \neq \mathrm{Tr}(E\rho_2)$.

Similarly, according to the one-to-one correspondence between $\rho \leftrightarrow f_\rho$, learning an unknown quantum state coincides with learning a real-valued linear functional on the effect space.

4. Learning Quantum Measurements

In this section, we follow the quantum learning framework presented in Section 3 and explicitly show how to derive the upper bound for the fat-shattering dimension, Rademacher complexity and the covering/entropy number. We then discuss how the relation of the complexity measures and quantum state discrimination.

Recall that, in the problem of learning an unknown quantum measurement, the goal is to learn a fixed but unknown effect $\Pi \in \mathcal{E}(\mathbb{C}^d)$ through the training data set is $\mathsf{Z}_n = \{(\rho_i, \mathrm{Tr}(\Pi\rho_i))\}_{i=1}^n$, where $\{\rho_i\}_{i=1}^n \in \mathcal{Q}(\mathbb{C}^d) \equiv \mathcal{X}$ distribute independently according to the unknown measure μ. Note that learning Π is equivalent to learning a two-outcome POVM $\{\Pi, \mathcal{I} - \Pi\}$. Due to the correspondence between a quantum effect $E \in \mathcal{E}(\mathbb{C}^d)$ and the linear functional $f_E : \rho \mapsto \langle E, \rho \rangle$ on the input space $\mathcal{X}$ (Proposition 1), we consider the hypothesis set that consists of all quantum effects;[**] that is,

$$\mathcal{F} = \{f_E : E \in \mathcal{E}(\mathbb{C}^d)\}.$$

In the following, we present our main result to the question: "*how many quantum states are needed to learn a quantum measurement?*" This is exactly the sample complexity problem introduced in Section 2.1. To tackle this problem, we have to estimate the complexity measures that characterise the size of the hypothesis set.

4.1. *The Fat-Shattering Dimension for Learning Quantum Measurements*

Our first step is to use a common trick in convex analysis; namely, "symmetrisation" of the state space and the effect space, to embed them into a subset of the Banach space. In other words, the symmetric convex hull of the state space is contained in a unit ball of Schatten 1-class:

$$S_1^d \subset \mathrm{conv}(-\mathcal{Q}(\mathbb{C}^d) \cup \mathcal{Q}(\mathbb{C}^d)),$$

where $\mathrm{conv}(\cdot)$ denotes the convex hull operation. Similarly, we have

$$S_\infty^d \subset \mathrm{conv}(-\mathcal{E}(\mathbb{C}^d) \cup \mathcal{E}(\mathbb{C}^d)).$$

[**]The hypothesis set can be chosen as a subset of the effects space, to which the target effect Π may not belong. Then the goal is to choose an effect in the hypothesis set that approximates the target well. We discuss this issue in Section 6. Also note that we sometimes denote $\mathcal{F}$ as the subset of $\mathcal{E}(\mathbb{C}^d)$ and sometimes denote it as the linear functionals formed by that subset.

Now the input space $\mathcal{X} \subset S_1^d$ and the hypothesis set $\mathcal{F}$ consists of linear functionals which can be paremeterised by the elements in S_∞^d. That is,

$$\mathcal{F} = \{f_E : E \in S_\infty^d\}.$$

The main reason for introducing S_1^d and S_∞^d is that they are unit balls which are *polar* to each other (through the Hilbert-Schmidt inner product). Thus, we can apply Mendelson and Schechtman's result (Lemma 1) to estimate the fat-shattering dimension.

The following is our main result in this section.

Theorem 10 (Fat-Shattering Dimension for Learning Quantum Measurements) *For all $0 < \epsilon < 1/2$, and integer $d \geq 2$, we have*

$$\mathrm{Pdim}(\mathcal{E}(\mathbb{C}^d)) \leq d^2,$$

and

$$\mathrm{fat}_{\mathcal{E}(\mathbb{C}^d)}(\epsilon, \mathcal{Q}(\mathbb{C}^d)) = \min\{O(d/\epsilon^2), d^2\}.$$

The operational meaning of the two quantities in Theorem 10 will become clear in Sections 4.4 and 5.4, where the pseudo-dimension and fat-shattering dimension are related to tasks of the quantum set discrimination and quantum random access codes, respectively.

Proof: We first present the outline of the proof. According to the definition of the fat-shattering dimension, it follows that the function $\mathrm{fat}_{\mathcal{F}}(\epsilon)$ is non-increasing in ϵ. Hence, our first objective is to check whether the fat-shattering dimension is unbounded for arbitrarily small ϵ. Equivalently, it suffices to find the pseudo dimension which bounds the fat-shattering dimension (Theorem 3). Second, assume there is a set of n points that can be ϵ-shattered; we will find an inequality to relate n with ϵ, which proves our claim.

(i) Pseudo Dimension: Since $\mathbb{M}_d$ is a vector space with dimension d^2 and S_∞^d is a subset of $\mathbb{M}_d$, we can embed S_∞^d into a real vector space of dimension d^2. Hence, by Theorem 2 we obtain $\mathrm{Pdim}(\mathcal{F}) \leq d^2$.

(ii) Fat-Shattering Dimension: Consider any set $\mathcal{S} = \{x_1, \ldots, x_n\} \subset S_1^d$ is ϵ-shattered by S_∞^d, where $n \leq d^2$. Denote a *Rademacher series* as $\sum_{i=1}^n \gamma_i x_i$, where $\{\gamma_i\}_{i=1}^n$ are independent and uniform $\{+1, -1\}$ random variables (also called Rademacher random variables). By selecting $a_i = \gamma_i$ in Lemma 1, we have

$$\epsilon n \leq \left\| \sum_{i=1}^n \gamma_i x_i \right\|_1 . \tag{8}$$

We adopt a probabilistic method to upper bound the right-hand side of Eq. (8). If we can find a quantity $C(n,d)$ that upper bounds $\mathbb{E}\left\|\sum_{i=1}^n \gamma_i x_i\right\|_1$, then there is a realization of $\{\gamma_i\}_{i=1}^n$ such that $\left\|\sum_{i=1}^n \gamma_i x_i\right\|_1 \leq C(n,d)$. As a result, it remains to find an upper bound for the expected norm of the Rademacher series $\mathbb{E}\left\|\sum_{i=1}^n \gamma_i x_i\right\|_1$.

In order to upper bound the Rademacher series, we need the powerful *Noncommutative Khintchine Inequalities* [87]:

Proposition 3 (Noncommutative Khintchine Inequalities [87, 88]) *Let $\{x_i\}_{i=1}^n$ be deterministic $d \times d$ matrices, $\{\gamma_i\}_{i=1}^n$ be independent Rademacher random variables. Then*

$$\mathbb{E}\left\| \sum_{i=1}^n \gamma_i x_i \right\|_p \approx_p \begin{cases} \left(\left\| \left(\sum_{i=1}^n x_i x_i^\dagger\right)^{1/2} \right\|_p^p + \left\| \left(\sum_{i=1}^n x_i^\dagger x_i\right)^{1/2} \right\|_p^p \right)^{1/p}, & \text{if} \quad 2 \leq p < \infty \\ \inf_{x_i = a_i + b_i} \left(\left\| \left(\sum_{i=1}^n a_i a_i^\dagger\right)^{1/2} \right\|_p^p + \left\| \left(\sum_{i=1}^n b_i^\dagger b_i\right)^{1/2} \right\|_p^p \right)^{1/p}, & \text{if} \quad 1 \leq p \leq 2. \end{cases}$$

where $\approx_p$ *means that the equality holds up to an absolute constant depending on* p*, and* † *denotes the complex conjugate operation.*

Note that Haagerup and Musat [88] proved that the result also holds as $\{\gamma_i\}_{i=1}^n$ *are independent standard complex Gaussian random variables*

By Proposition 3, it is not hard to obtain the following inequality:

$$\mathbb{E}\left\|\sum_{i=1}^n \gamma_i x_i\right\|_1 \lesssim \left\|\left(\sum_{i=1}^n x_i^2\right)^{1/2}\right\|_1.$$

Since the square operation preserves S_1^d, i.e. $x_i^2 \in S_1^d$, for all $x_i \in S_1^d$, by the convexity of S_1^d, we have $\frac{1}{n}\sum_{i=1}^n x_i^2 \in S_1^d$. Then the problem is reduced to finding

$$\max_{\{x_i\}\in S_1^d} \sqrt{n}\left\|\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{1/2}\right\|_1 = \max_{x\in S_1^d} \sqrt{n}\|\sqrt{x}\|_1,$$

which is essentially a convex optimization problem

$$\max_{x\in S_1^d} \sqrt{n}\sum_{j=1}^d \sqrt{|\lambda_i|}, \text{ subject to } \sum_{j=1}^d |\lambda_j| = 1.$$

Since the square root is concave, we attain the maximum when $|\lambda_j| = 1/d$, for $j = 1, \ldots, d$. That is,

$$\mathbb{E}\left\|\sum_{i=1}^n \gamma_i x_i\right\|_1 \lesssim \max_{x\in S_1^d} \sqrt{n}\sum_{i=1}^d \sqrt{\lambda_i} = \sqrt{nd}. \tag{9}$$

Consequently, there is a realization of $\{\gamma_i\}_{i=1}^n$ such that $\|\sum_{i=1}^n \gamma_i x_i\|_1 \leq \sqrt{nd}$, $\forall x_i \in S_1^d$. Combined with Eq. (8), we have $n \leq d/\epsilon^2$ which proves our claim.

□

In the following proposition, we will demonstrate that the upper bound of the fat-shattering dimension in Theorem 10 is tight.

Proposition 4 *Considering a Hilbert space* $\mathbb{C}^d$*, there exist infinitely many sets of* d *quantum states that can be* $1/2$*-shattered by the effect space.*

Proof: Consider arbitrary d mutually orthogonal rank-1 projection operators (i.e. pure states) $\{\rho_i\}_{i=1}^d$ on $\mathbb{C}^d$ as the input states. Now for every $B \subseteq \{1, \ldots, d\}$, denote $f_B : \rho \to \langle \sum_{i\in B} \rho_i, \rho\rangle$, for some $\rho \in \mathcal{Q}(\mathbb{C}^d)$. Note that one can easily check $\sum_{i\in B} \rho_i \in \mathcal{E}(\mathbb{C}^d)$. Then for $i \in B$, we have

$$\begin{aligned} f_B(\rho_i) &= \left\langle \sum_{i\in B} \rho_i, \rho \right\rangle \\ &= \langle \rho_i, \rho_i \rangle \\ &= 1. \end{aligned}$$

Similarly, $f_B(\rho_i) = 0$ if $i \notin B$. As a result, $\{\rho_i\}_{i=1}^d$ is $1/2$-shattered by $\{f_B\}$. □

4.2. *The Rademacher Complexity*

Following the paradigm in Section 4.1, we calculate the Rademacher complexity of the effect space $\mathcal{E}(\mathbb{C}^d)$ via the duality formula, Theorem 9, and the noncommutative Khintchine inequality, Proposition 3.

Theorem 11 (Rademacher Complexity for Learning Quantum Measurements) *Assume the input space is the state space $\mathcal{X} = \mathcal{Q}(\mathbb{C}^d)$ and the hypothesis set $\mathcal{F} = \{f_E : \forall E \in \mathcal{E}(\mathbb{C}^d)\}$. Then the Rademacher complexity is*

$$\mathcal{R}_n(\mathcal{E}(\mathbb{C}^d)) = O\left(\sqrt{d}\right).$$

Proof:

Recall the definition of the Rademacher complexity (Definition 8). We have

$$\begin{aligned}
\sqrt{n}\mathcal{R}_n(S_\infty^d) &= \mathbb{E} \sup_{E\in S_\infty^d} \left| \sum_{i=1}^n \gamma_i f_E(x_i) \right| \\
&= \mathbb{E} \sup_{E\in S_\infty^d} \left| \sum_{i=1}^n \gamma_i \langle E, x_i \rangle \right| \\
&= \mathbb{E} \sup_{E\in S_\infty^d} \left| \left\langle E, \sum_{i=1}^n \gamma_i x_i \right\rangle \right| \\
&\leq \mathbb{E} \left\| \sum_{i=1}^n \gamma_i x_i \right\|_1 \\
&\lesssim \sqrt{nd}.
\end{aligned}$$

The third line is due to the duality formula (Theorem 9), and the last relation follows from Eq. (9). This completes the proof. □

4.3. *The Entropy Number*

The covering number (and the related entropy number) follows directly from the Rademacher complexity by the *Sudakov's minoration theorm* (see e.g. [19]).

Corollary 1 (Entropy Number for Learning Quantum Measurements) *Assume the input space is the state space $\mathcal{X} = \mathcal{Q}(\mathbb{C}^d)$ and the hypothesis set $\mathcal{F} = \{f_E : \forall E \in \mathcal{E}(\mathbb{C}^d)\}$. Then for each $\epsilon > 0$, the covering number of the function class is*

$$\log \mathcal{N}_2(\epsilon, \mathcal{E}(\mathbb{C}^d), n) = O(d/\epsilon^2)$$

for all positive integers n.

Proof: The upper bound of the empirical L_2 entropy number by the Rademacher complexity follows directly from the Sudakov's minoration theorem:

Theorem 12 (Sudakov's Minoration Theorem [76, 89, 90]) *Let $\mathcal{T}$ be an index set. Let $X = (X_t)_{t\in\mathcal{T}}$ be a sub-Gaussian process[††] with L_2-metric d_X (i.e. $d_X(s,t) = \|X_s - X_t\|_2$) for*

[††] A stochastic process is called *sub-Guassian* if there exists $\sigma > 0$ such that $\mathbb{E}\exp(\theta X_t) \leq \exp(\sigma^2\theta^2/2)$ for all $\theta \in \mathbb{R}$ and $t \in \mathcal{T}$. Note that both Gaussian process and Rademacher process belong to sub-Gaussian process.

$s, t \in \mathcal{T}$). Then for each $\epsilon > 0$,

$$\epsilon(\log \mathcal{N}(\epsilon, \mathcal{T}, d_X))^{1/2} \leq C\mathbb{E} \sup_{t\in\mathcal{T}} \|X_t\|_1,$$

for some constant C.

Denote the (vector-valued) stochastic process by

$$X_f := \frac{1}{\sqrt{n}}(\gamma_1 f(x_1), \ldots, \gamma_n f(x_n)),$$

where $x_1, \ldots, x_n$ are independently drawn from $\mathcal{X}$ according to some distribution μ. Then the distance measure can be calculated as

$$d_X(f, g) = \|X_f - X_g\|_2 = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} |f(x_i) - g(x_i)|^2 \right)^{1/2} = \|f - g\|_{L_2(\mu_n)}.$$

Invoke Theorem 12 and 11 to obtain

$$\begin{aligned} \log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n)) &= \log \mathcal{N}(\epsilon, \mathcal{F}, d_X) \\ &\leq C^2 \frac{\left(\mathbb{E} \sup_{f\in\mathcal{F}} \|X_f\|_1\right)^2}{\epsilon^2} \\ &= C^2 \frac{\mathcal{R}_n(\mathcal{F})^2}{\epsilon^2} \\ &\leq C^2 \frac{d}{\epsilon^2}. \end{aligned}$$

Note that the right-hand side in the last line does not depend on the distribution μ. Hence the entropy number $\log \mathcal{N}_2(\epsilon, \mathcal{F}, n) = \sup_{\mu_n} \log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n)) = O(d/\epsilon^2)$ follows. □

Remark. The pseudo dimension of the effect space $\mathrm{Pdim}(\mathbb{C}^d) = d^2$ means that we need d^2 parameters to exactly determine a POVM element. Note that it coincides with the number of measurements in the quantum measurement tomography (since $\mathcal{E}(\mathbb{C}^d)$ lies in a d^2-dimensional real vector space).

On the other hand, the covering number provides a geometric perspective in the learning problem. That is, if we relax the criterion by tolerating an ϵ accuracy, then the effect space can be covered by $\mathcal{N}_2(\epsilon, \mathcal{E}(\mathbb{C}^d)) = \exp(d/\epsilon^2)$ balls each with radius ϵ. In other words, we need $\log \mathcal{N}_2(\epsilon, \mathcal{E}(\mathbb{C}^d)) \leq d/\epsilon^2$ samples to identify which ball the target POVM element lies in. Consequently, the entropy number guarantees that we can specify a POVM element, satisfying the "PAC" criterion with accuracy ϵ and confidence δ, with only d/ϵ^2 samples. This provides a quadratical speed-up over conventional quantum tomography.

4.4. *The Relationship to Quantum State Discrimination*

Quantum State Discrimination studies how to optimally distinguish a set of quantum states according to a figure of merit [91, 92].

There are nevertheless some limitations in quantum state discrimination because the states cannot always be perfectly discriminated. Moreover, it may not be necessary to find the exact

state in some scenario. Therefore, Zhang and Ying [93] considered *quantum set discrimination*, where the goal is to identify which set the given state belongs to. Now we relate the concepts of the fat-shattering dimension to *quantum set discrimination*.

Definition 9 (ϵ-separable Set) *A set $\mathcal{S} = \{x_1, \ldots, x_n\} \subset \mathbb{M}_d$ is ϵ-*(linearly) separable *with respect to the set $\mathcal{W} \subseteq \mathbb{M}_d$ if and only if for any subset $B \subseteq \mathcal{S}$ there exists an ϵ-strip which separates B from its complement $\mathcal{S} \setminus B$. In other words, there exist $w \in \mathcal{W}$ and $a \in \mathbb{R}$ such that $\langle w, x\rangle \geq a + \epsilon/2$ when $x \in B$ and $\langle w, x\rangle \leq a - \epsilon/2$ when $x \in \mathcal{S} \setminus B$.*

It is not difficult to see that an 2ϵ-separable set correspond to the task of quantum set discrimination with ensemble $\mathcal{S} = \{x_1, \ldots, x_n\}$, where the error probability that a given state can be classified to a set is no greater than $(1-\epsilon)/2$. One interesting question to ask is what the maximum cardinality of the 2ϵ-separable set is. The following proposition shows that the fat-shattering dimension equals this quantity.

Proposition 5 *Denote the function class $\mathcal{F} = \{\rho \to \langle E, \rho\rangle : E \in \mathcal{E}(\mathbb{C}^d)\}$. Assume there exists a set $\mathcal{S} = \{x_1, \ldots, x_n\} \subset \mathcal{Q}(\mathbb{C}^d)$ that is 2ϵ-separable with respect to $\mathcal{E}(\mathbb{C}^d)$. Then the maximum cardinality of the set $\mathcal{S}$ is $fat_{\mathcal{F}}(\epsilon)$.*

Proof: Recall from Definition 5 that the set $\mathcal{S} = \{x_1, \ldots, x_n\}$ is 2ϵ-separable with respect to $\mathcal{E}(\mathbb{C}^d)$ if and only if $\underline{\text{fat}}_\epsilon(\mathcal{F}) \geq n$. Then the proposition is equivalent to show that $\text{fat}_\epsilon(\mathcal{F}) = \underline{\text{fat}}_\epsilon(\mathcal{F})$.

Because $\underline{\text{fat}}_\epsilon(\mathcal{F}) \leq \text{fat}_\epsilon(\mathcal{F})$ by definition, it suffices to show $\underline{\text{fat}}_\epsilon(\mathcal{F}) \geq \text{fat}_\epsilon(\mathcal{F})$. Given $\epsilon > 0$, choose a set $\mathcal{S} = \{x_1, \ldots, x_n\}$ with the largest integer n such that $\mathcal{S}$ is ϵ-shattered by $\mathcal{F}$ (with $\{s_i\}_{i=1}^n$ witnessing the shattering). Without loss of generality, we assume some $s_i \neq 1/2$. We then choose an arbitrary subset $B \subseteq \{1, \ldots, n\}$ that contains i. By the definition of fat-shattering dimension, there exists $s_i := s(x_i)$ such that there is some function $E_B \in \mathcal{F}$ for each set $B \subset \mathcal{S}$ so that $\langle E_B, x_i\rangle \geq s_i + \epsilon$, if $i \in B$. Also, we have $\langle E_{\bar{B}}, x_i\rangle \leq s_i - \epsilon$, where $\bar{B} = \mathcal{S} \setminus B$. Now denote $\overline{E_{\bar{B}}} := \mathcal{I} - E_{\bar{B}}$ such that

$$\langle \overline{E_{\bar{B}}}, x_i\rangle = 1 - \langle E_{\bar{B}}, x_i\rangle \geq 1 - s_i + \epsilon.$$

Since $\mathcal{F}$ is convex, set $E'_B := \frac{1}{2}(E_B + \overline{E_{\bar{B}}}) \in \mathcal{F}$ which satisfies

$$\langle E'_B, x_i\rangle \geq 1/2 + \epsilon.$$

Similarly, let $E'_{\bar{B}} := \mathcal{I} - E'_B$, we have

$$\langle E'_{\bar{B}}, x_i\rangle \leq 1/2 - \epsilon.$$

The same argument holds for other $s_i \neq 1/2$. It follows that the level fat-shattering dimension (witnessed by $1/2$) also achieves the cardinality n of the ϵ-shattered set, which completes the proof. □

5. Learning Quantum States

In this section, we consider the problem of learning an unknown quantum state $\rho' \in \mathcal{Q}(\mathbb{C}^d)$ through the training data set $\mathsf{Z}_n = \{(E_i, \text{Tr}(\rho' E_i))\}_{i=1}^n$, where $\{E_i\}_{i=1}^n \in \mathcal{X} = \mathcal{E}(\mathbb{C}^d)$ are independently sampled according to an unknown distribution μ'. By Proposition 2, the hypothesis set consists of the linear functional $f_\rho : E \mapsto \langle E, \rho\rangle$ on $\mathcal{E}(\mathbb{C}^d)$:

$$\mathcal{F}' = \{f_\rho : \forall \rho \in \mathcal{Q}(\mathbb{C}^d)\}.$$

Similarly, we embed the input space into the unit ball of Schatten ∞-class, i.e. $\mathcal{X} = S_\infty^d$. Then the hypothesis set is the collection of linear functionals on the input space, i.e. S_1^d. In the following, we aim to calculate the complexity measures of S_1^d, which characterise the sample complexity of learning quantum states. It is interesting to see that the proofs derived in this section (i.e. the complexity measures of learning quantum states) parallel with that in the previous section (i.e. the complexity measures of learning quantum measurements) due to the duality relation in Theorem 8. Finally, we discuss the relationship of the fat-shattering dimension with quantum random access codes.

5.1. *The Fat-Shattering Dimension for Learning Quantum States*

Under the framework presented in Section 3, we characterising the input space $\mathcal{X} \subset S_\infty^d$ and the hypothesis set $\mathcal{F}'$ consisting of the linear functionals with elements in S_1^d. That is,

$$\mathcal{F}' = \{f_\rho : \rho \in S_1^d\}.$$

Therefore, we have the main result of deriving the fat-shattering dimension of the state space.

Theorem 13 (Fat-Shattering Dimension for Learning Quantum States) *For all* $0 < \epsilon < 1/2$ *and integer* $d \geq 2$*, we have*

$$\mathrm{Pdim}(\mathcal{Q}(\mathbb{C}^d)) \leq d^2 - 1,$$

and

$$\mathrm{fat}_{\mathcal{Q}(\mathbb{C}^d)}(\epsilon, \mathcal{E}(\mathbb{C}^d)) = \min\{O(\log d/\epsilon^2), d^2 - 1\}.$$

Proof: Following the same fashion as in the proof of Theorem 10, we first estimate the pseudo dimension and then the fat-shattering dimension.

(i) Pseudo Dimension: The state space lies in the set $\{x \in \mathcal{M}^d : \|x\|_1 = 1\}$, which is the sphere of S_1^d, i.e. $\mathcal{Q}(\mathbb{C}^d) \subset \partial S_1^d$. Since ∂S_1^d can be embedded into a real vector space of dimension $d^2 - 1$, we have $\mathrm{Pdim}(\mathcal{Q}(\mathbb{C}^d)) \leq d^2 - 1$.

(ii) Fat-Shattering Dimension: For every $\{x_i\}_{i=1}^n \in S_\infty^d$, we have to calculate the Rademacher series $\mathbb{E}\left\|\sum_{i=1}^n \gamma_i x_i\right\|_\infty$. However, in the scenario of learning quantum states the input space lies in the Schatten ∞-class. We have to estimate the spectral norm of the Rademacher series. Benefiting from the recent development of matrix concentration inequalities, Tropp [20] proved the following results:

Proposition 6 (Upper Bound for Rademacher Series [20]) *Consider a finite sequence* $\{x_i\}$ *of deterministic Hermitian matrices with dimension* d*, and let* $\{\gamma_i\}$ *be independent Rademacher variables. Form the matrix Rademacher series*

$$Y = \sum_i \gamma_i x_i.$$

Compute the variance parameter

$$\sigma^2 = \sigma^2(Y) = \|\mathbb{E}\left(Y^2\right)\|_\infty.$$

Then

$$\mathbb{E}\|Y\|_\infty \leq \sqrt{2\sigma^2 \log d}.$$

Note that the result also holds for the case $\{\gamma_i\}$ *being standard complex Gaussian variables.*

Invoking Tropp's development of matrix concentration inequalities (see Proposition 6), we have

$$\mathbb{E}\left\|\sum_{i=1}^{n}\gamma_i x_i\right\|_\infty \leq \sqrt{2\sigma^2\log d}, \tag{10}$$

where $\sigma^2 := \left\|\mathbb{E}\left(\sum_{i=1}^{n}\gamma_i x_i\right)^2\right\|_\infty$. Straightforward computation shows that

$$\sigma^2 = \left\|\mathbb{E}\left(\sum_{i=1}^{n}\gamma_i x_i\right)^2\right\|_\infty = \left\|\mathbb{E}\left(\sum_{i,j}\gamma_i\gamma_j x_i x_j\right)\right\|_\infty = \left\|\sum_{i=1}^{n}x_i^2\right\|_\infty \leq n.$$

We get

$$\mathbb{E}\left\|\sum_{i=1}^{n}\gamma_i x_i\right\|_\infty \leq \sqrt{2n\log d}.$$

Then there is a realization of $\{\gamma_i\}_{i=1}^n$ such that $\left\|\sum_{i=1}^{n}\gamma_i x_i\right\|_\infty \leq \sqrt{2n\log d}$, $\forall x_i \in S_\infty^d$.

From Lemma 1, by selecting $a_i = \gamma_i$, $\epsilon n \leq \left\|\sum_{i=1}^{n}\gamma_i x_i\right\|_\infty$. Combining the inequalities, we have $n \leq O(\log d/\epsilon^2)$ completing the proof. □

5.2. *The Rademacher Complexity*

By repeating the procedure introduced in Section 4.2, we can compute the Rademacher complexity of the state space.

Theorem 14 (Rademacher Complexity for Learning Quantum States) *Assume the input space is the effect space $\mathcal{X} = \mathcal{E}(\mathbb{C}^d)$. The hypothesis set $\mathcal{F}$ defined on $\mathcal{X}$ is the state space $\mathcal{Q}(\mathbb{C}^d)$. Then the Rademacher complexity of hypothesis set is*

$$\mathcal{R}_n(\mathcal{Q}(\mathbb{C}^d)) = O\left(\sqrt{\log d}\right).$$

Proof: Recall from the definition of the Rademacher complexity. We have

$$\begin{aligned}
\sqrt{n}\mathcal{R}_n(S_1^d) &= \mathbb{E}\sup_{\rho\in S_1^d}\left|\sum_{i=1}^{n}\gamma_i f_\rho(E_i)\right| \\
&= \mathbb{E}\sup_{\rho\in S_1^d}\left|\sum_{i=1}^{n}\gamma_i\langle E_i,\rho\rangle\right| \\
&= \mathbb{E}\sup_{\rho\in S_1^d}\left|\left\langle\sum_{i=1}^{n}\gamma_i E_i,\rho\right\rangle\right| \\
&\leq \mathbb{E}\left\|\sum_{i=1}^{n}\gamma_i E_i\right\|_\infty \\
&\lesssim \sqrt{n\log d}.
\end{aligned}$$

The forth line is due to the duality formula, Theorem 9. The last relation follows from Eq. (10), which completes the proof. □

5.3. *The Entropy Number*

Corollary 2 (Entropy Number for Learning Quantum States) *Assume the input space is $\mathcal{X} = \mathcal{E}(\mathbb{C}^d)$. The function class $\mathcal{F}$ defined on $\mathcal{X}$ is the state space $\mathcal{Q}(\mathbb{C}^d)$. Then for each $\epsilon > 0$, the covering number of the function class is*

$$\log \mathcal{N}_2(\epsilon, \mathcal{Q}(\mathbb{C}^d), n) = O(\log d/\epsilon^2).$$

for all positive integers n.

Compared with the entropy number of the effect space, the result of the state space is proportional to the logarithmic dimension. The intuition behind this is that the unit ball of Schatten ∞-class is much larger than the unit ball of Schatten 1-class. Thus, it requires more ϵ-radius ball to cover the whole effect space than the state space. From the volumetric perspective, the fact will be more evident. Denote $|\cdot|$ as the Lebesgue measure on the Banach space of the Schatten class. The volume of the Schatten balls are estimated to be $|S_p^d| \simeq d^{-1/2-1/p}$ for $0 < p \leq \infty$ [94]. Hence, it can be calculated that (see also [95]):

$$\frac{|\mathcal{E}(\mathbb{C}^d)|^{1/d^2}}{|\mathcal{Q}(\mathbb{C}^d)|^{1/(d^2-1)}} \simeq \left(\frac{|S_\infty^d|}{|S_1^d|}\right)^{1/d^2} \simeq d,$$

which shows that the volume of the effect space is essentially exponential (in the dimension d) to the state space. Recall that the complexity measures are the quantity to estimate the effective size of the hypothesis set. Accordingly, it is reasonable that the complexity measures of the effect space are exponentially compared with that of the state space. In other words, the results of Theorem 10 demonstrate the richness of the effect space.

5.4. *The Relationship to Quantum Random Access Coding*

The learnability of quantum states was first addressed by Aaronson [14]. Ingeniously, he applied the results of Quantum Random Access coding [15] to provide an information-theoretic upper bound on the fat-shattering dimension for learning m-qubit quantum states. We first give the definitions of QRA coding then discuss Aaronson's result.

Definition 10 (Quantum Random Access Coding) *An (n, m, p)-QRA coding is a function that maps n-bit strings $x \in \{0,1\}^n$ to m-qubit states ρ_x satisfying the following: For every $i \in \{1, \ldots, n\}$ there exists a POVM $E^i = \{E_0^i, E_1^i\}$ such that $\mathrm{Tr}(E_{x_i}^i \rho_x) \geq p$ for all $x \in \{0,1\}^n$, where x_i is the i-th bit of x.*

If there exists an (n, m, p)-QRA coding, we have the fact that the sets $\{E_i\}_{i=1}^n$ are $(p-1/2)$-shattered by $\{\rho_y\}$ and the constant value $1/2$ witnesses the shattering. That is,

$$m \geq (1 - H(\epsilon + 1/2))n \geq c \cdot \epsilon^2 n. \tag{11}$$

Therefore, the inequality gives an upper bound on the level fat-shattering dimension, i.e. $\underline{\mathrm{fat}}_{\mathcal{Q}(\mathbb{C}^d)}(p - 1/2) = O(m/\epsilon^2)$. Conversely, the fat-shattering dimension with scale $(p - 1/2)$ does not guarantee the existence of an (n, m, p)-QRA coding (since there may be some $\alpha_i < 1/2$), while provide an upper bound on the success probability p if it exists.

However, in the case that functions in $\mathcal{F}$ have a bounded range of $[0,1]$, Gurvits [10] utilized the Pigeonhole principle to relate the level fat-shattering dimension with the fat-shattering dimension.

Theorem 15 (Gurvits [10]) *For any hypothesis set $\mathcal{F}$ consisting of $[0,1]$-valued functions, we have*

$$(2(1-2\epsilon)/\epsilon)^{-1}\mathrm{fat}_{\mathcal{F}}(2\epsilon) \leq \underline{\mathrm{fat}}_{\mathcal{F}}(\epsilon/2) \leq \mathrm{fat}_{\mathcal{F}}(\epsilon/2). \tag{12}$$

By definition, $\underline{\mathrm{fat}}_{\mathcal{F}}(\epsilon) \leq \mathrm{fat}_{\mathcal{F}}(\epsilon)$. However, from the above theorem, the dependencies on the dimension d are of the same order for both the level fat-shattering dimension and the fat-shattering dimension. Consequently, from Eq. (11) we have $\underline{\mathrm{fat}}_{\mathcal{F}}(\epsilon) = O(m/\epsilon^2)$, which leads to $\mathrm{fat}_{\mathcal{F}}(\epsilon) = O(m/\epsilon^2)$ according to the inequalities in Eq. (12). Thus we recover Aaronson's result.

Theorem 16 (Aaronson [14]) *The fat-shattering dimension for learning the class of all m-qubits, $\mathcal{F}$, is $fat_{\mathcal{F}}(\epsilon) = O(m/\epsilon^2)$.*

We remark that it is unknown whether $\underline{\mathrm{fat}}_{\mathcal{F}}(\epsilon) = \mathrm{fat}_{\mathcal{F}}(\epsilon)$ for $\mathcal{F} = \mathcal{Q}(\mathbb{C}^d)$.

Proposition 7 *There is no $(2^{2m}, m, p)$-QRA coding for $1/2 < p \leq 1$ and positive integer m.*

Hayashi *et al.* [18] showed that there is no $(2^{2m}, m, p)$-QRA coding for $1/2 < p \leq 1$. This result can be directly derived from Theorem 13, which shows that $\mathrm{Pdim}(\mathcal{Q}(\mathbb{C}^d)) \leq d^2 - 1$. The dimension d of m-qubit is 2^m. Then the upper bound of the pseudo dimension shows that there is no $d^2 = 2^{2m}$ two-outcome POVMs that can be shattered (by the function class of the state space), which coincides with Hayashi *et al's* result.

6. The Algorithms for Quantum Machine Learning

In the previous sections, we demonstrate the information-theoretical analysis of the quantum learning problems. In this section, provide a constructive way to implement quantum ML tasks by representing the learning framework in Bloch space.

We gather all the materials and derivations concerning the Bloch-sphere representation into Appendix C. Recall from Eq. (C.5) that the function class of rank-k effects and their mixture can be represented as the following affine functional:

$$\mathcal{F}_k = \mathrm{conv}\left(\{\mathbf{r} \mapsto \frac{k}{d}\left(1 + (d-1)\mathbf{r}\cdot\mathbf{n}_{(k)}\right)\}\right),$$

where $\mathbf{r}$ is the Bloch vector of the quantum state; $\mathbf{n}_{(k)}$ (see Eq. (C.2)) parameterises the function in the hypothesis set $\mathcal{F}_k$. Moreover, it can in turn be written as

$$\mathcal{F}_k = \sigma(\mathbf{v}\cdot\mathbf{r} + v_0),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is called the *activation function*. The Bloch vector $\mathbf{r} \in \mathbb{R}^{d^2-1}$ is the input vector; $[v_0, \mathbf{v}] \in \mathbb{R}^{d^2}$ is the input weights. Each map $\mathbf{r} \mapsto \sigma(\mathbf{v}\cdot\mathbf{r} + v_0)$ can be thought of as a function computed by the linear perceptron. Using the terminology from the theory of neural network [62], each $\mathcal{F}_k$ is called the *single-layer neural network* (see Appendix D for more details).

Considering the function class of the whole effect space, we exploit the convexity of the effect space, and obtain the following result:

$$\mathcal{F} = \sum_{k=0}^{d} \frac{w_k \cdot k}{d} \left(1 + (d-1)\mathbf{r} \cdot \mathbf{n}_{(k)}\right) =: \frac{1}{d}\left(n_0 + (d-1)\mathbf{r} \cdot \mathbf{n}\right), \tag{13}$$

where $\sum_{k=0}^{d} w_k = 1$. This is called the *two-layer neural network* (also called the *single-hidden layer net*). Based on this formulation, the task of learning Schatten ∞-norm matrices is equivalent to learning the weighted coefficients $[n_0, \mathbf{n}]$ of a neural network, and a corresponding neural network algorithm can be given (see Algorithm 1). Therefore, the matrix learning problem can be implemented by existing neural network algorithms or other multivariate regression techniques. We note that the neural network formulation for learning quantum states follows in the same way by virtue of the duality.

Additionally, the fat-shattering dimension for $\mathcal{F}_k$ can easily be bounded from the classical results in neural networks. We have the following corollary.

Corollary 3 *Suppose the hypothesis set $\mathcal{F}_k$ consists of rank-k projection operators and their mixture. We have*

$$fat_{\mathcal{F}_k}(\epsilon) \leq \frac{k(d-1)(d-k)}{(d\epsilon)^2},\ k = \{0, 1, \ldots, d\}.$$

Proof: Since $\mathcal{F}_k$ is a linear function class on $\mathbb{R}^{d-1}$, invoking the classical results from Anthony and Bartlett [62]:

$$\mathrm{fat}_{\mathcal{F}}(\epsilon) \leq \frac{a^2 b^2}{\epsilon^2},$$

where $\mathcal{F} = \{\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{x}\|_2 \leq b, \|\mathbf{w}\|_2 \leq a,\ \mathbf{x}, \mathbf{w} \in \mathbb{R}^{d^2-1}\}$.

Therefore, it remains to calculate the coefficients in Eq. (C.3). Since $\|\mathbf{r}\|_2 \leq 1$, and

$$\left\| \frac{k(d-1)}{d} \sqrt{\frac{d-k}{k(d-1)}} \right\|_2 = \sqrt{\frac{k(d-1)(d-k)}{d^2}},$$

the result follows. □

We can see from the corollary that the fat-shattering dimension increases when the the rank k approaches a half of the Hilbert space dimension d, which means that the classes $\{\mathcal{F}_k\}$ form a hierarchical structure. Operationally, the hypothesis set $\mathcal{F}_1$ can be chosen at first. It can then be enlarged into $\mathrm{conv}(\mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{F}_2)$ and so forth until the whole effect space is considered. This is called the structural risk minimization (SRM [2]), and is usually adopted in classical ML to avoid overfitting. Here we give two examples to illustrate the concepts in Corollary 3.

Example 1 (Learning rank-1 Projection Valued Measures (PVMs): Qubit system attains the upper bound): The fat-shattering dimension of rank-1 projection operators and their mixture in a qubit system can be bounded by

$$\mathrm{fat}_{\mathcal{F}_1}(\epsilon) \leq \frac{(N-1)^2}{(N\epsilon)^2} = \frac{1}{4\epsilon^2}.$$

Consider two quantum states $\rho_{\mathbf{r}_1} = |1\rangle\langle 1|$, $\rho_{\mathbf{r}_2} = |-\rangle\langle -|$ with corresponding Bloch vectors $\mathbf{r}_1 = (0,0,-1)$, $\mathbf{r}_2 = (-1,0,0)$. To shatter these two quantum states, we construct four quantum effects with the Bloch vectors:

$$\mathbf{n}_{00} = \frac{1}{\sqrt{2}}(1,0,1),\ \mathbf{n}_{10} = \frac{1}{\sqrt{2}}(1,0,-1),$$
$$\mathbf{n}_{11} = \frac{1}{\sqrt{2}}(-1,0,-1),\ \mathbf{n}_{01} = \frac{1}{\sqrt{2}}(-1,0,1).$$

Since the angles between the states and effects are either $\pi/4$ or $3\pi/4$, we have

$$(\mathrm{Tr}(E_{\mathbf{n}_{00}}\rho_{\mathbf{r}_1}), \mathrm{Tr}(E_{\mathbf{n}_{00}}\rho_{\mathbf{r}_2})) = (\frac{1}{2}(1-\frac{1}{\sqrt{2}}), \frac{1}{2}(1-\frac{1}{\sqrt{2}})),$$
$$(\mathrm{Tr}(E_{\mathbf{n}_{10}}\rho_{\mathbf{r}_1}), \mathrm{Tr}(E_{\mathbf{n}_{10}}\rho_{\mathbf{r}_2})) = (\frac{1}{2}(1+\frac{1}{\sqrt{2}}), \frac{1}{2}(1-\frac{1}{\sqrt{2}})),$$
$$(\mathrm{Tr}(E_{\mathbf{n}_{11}}\rho_{\mathbf{r}_1}), \mathrm{Tr}(E_{\mathbf{n}_{11}}\rho_{\mathbf{r}_2})) = (\frac{1}{2}(1+\frac{1}{\sqrt{2}}), \frac{1}{2}(1+\frac{1}{\sqrt{2}})),$$
$$(\mathrm{Tr}(E_{\mathbf{n}_{01}}\rho_{\mathbf{r}_1}), \mathrm{Tr}(E_{\mathbf{n}_{01}}\rho_{\mathbf{r}_2})) = (\frac{1}{2}(1-\frac{1}{\sqrt{2}}), \frac{1}{2}(1+\frac{1}{\sqrt{2}})).$$

Clearly these four quantum effects $\frac{1}{2\sqrt{2}}$-shatter $(\mathbf{r}_1, \mathbf{r}_2)$ and achieve the fat-shattering dimension $\mathrm{fat}_{\mathcal{F}_1}(\frac{1}{2\sqrt{2}}) = 2$.

The case of three quantum states follows similarly. Consider $\mathbf{r}_1 = (1,0,0)$, $\mathbf{r}_2 = (0,1,0)$, $\mathbf{r}_3 = (0,0,1)$, and $\mathbf{n}_{ijk} = (i,j,k)$ for $i,j,k \in \{0,1\}$. With some calculations, the eight quantum effects $\frac{1}{2\sqrt{3}}$-shatter $(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ and achieve the fat-shattering dimension $\mathrm{fat}_{\mathcal{F}_1}(\frac{1}{2\sqrt{3}}) = 3$.

It is worth emphasising that the dual problem of learning quantum states is equivalent to learning quantum measurements when the hypothesis set consists of rank-1 projections and their mixture. The reason is that the two mathematical objects are exactly the same, i.e. $\mathrm{conv}(\mathcal{F}_1) = \mathcal{Q}(\mathbb{C}^d)$. In this scenario, the dual problem has the same results, which is optimal in the sense of Quantum Random Access coding (i.e. (2,1,0.85)-QRA coding [96]). Furthermore, we note that the measurements in the (2,1,0.85)-QRA coding and the input states $(\rho_{\mathbf{r}_1}, \rho_{\mathbf{r}_1}^{\perp})$, $(\rho_{\mathbf{r}_2}, \rho_{\mathbf{r}_2}^{\perp})$ in this example are mutually unbiased bases (MUB) which attain the upper bound of the qubit system.

Example 2 (Rank equals a half the Hilbert space dimension): Consider a quaternary Hilbert space, i.e. $\mathbb{C}^4$. First, we show that there exist no two quantum states that can be 1/2-shattered by the convex hull of rank-1 projection operators. Consider two arbitrary different quantum states $\mathcal{S} = \{\rho_i\}_{i=1}^2$. If the function class $\mathcal{F}_1$ can 1/2-shatter the set $\mathcal{S}$, then there must be an effect $E \in \mathcal{F}_1$ such that $\mathrm{Tr}(E\rho_1) = \mathrm{Tr}(E\rho_2) = 1$. Clearly, it can be achieved only when E is a rank-1 projection and the two quantum states are both equal to E, which contradicts the assumption.

Second, we show there exist two quantum states that can be 1/2-shattered by the rank-2 projection operators. Assume $\rho_i = |i-1\rangle\langle i-1|$, $i = 1,2$. We construct four quantum effects as follows:

$$E_{11} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 0 & \\ & & & 0 \end{pmatrix},\ E_{01} = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & 0 \end{pmatrix},\ E_{10} = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 1 & \\ & & & 0 \end{pmatrix},\ E_{00} = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

in the computational basis. The two quantum states can then be 1/2-shattered by these four quantum effects. This example demonstrates that the set of rank-2 projections is richer than the set of rank-1 projections in terms of the complexity measures.

Remark. The readers may contemplate the pros and cons of Bloch-sphere representation when analysing the fat-shattering dimension. Indeed, Bloch-sphere representation provides a geometric picture so that we have more concrete ideas of the linear relation between quantum measurements and states. Furthermore, in Example 1 we see how the extreme points (projection operators) and MUB play the role in the fat-shattering dimension. However, it is difficult to fully characterise the region of the Bloch space. To the best of our knowledge, the most convenient metric used in Bloch-sphere representation is the Euclidean norm, which corresponds to the Hilbert-Schmidt norm (Schatten 2-norm) in the state space, i.e.

$$\|\rho_{\mathbf{r}_1} - \rho_{\mathbf{r}_2}\|_{\mathrm{HS}} = \sqrt{\frac{d-1}{2d}}\|\mathbf{r}_1 - \mathbf{r}_2\|_2.$$

Recalling that $\mathrm{conv}\big(-\mathcal{Q}(\mathbb{C}^d) \cup \mathcal{Q}(\mathbb{C}^d)\big) = S_1^d \subset S_2^d \subset S_\infty^d = \mathrm{conv}\left(-\mathcal{E}(\mathbb{C}^d) \cup \mathcal{E}(\mathbb{C}^d)\right)$, the Hilbert-Schmidt norm is not efficient in characterising the state space (that is why some regions in the Bloch sphere are not representative as valid states). On the other hand, the unit ball of Schatten 2-class is not sufficient to contain S_∞^d, so we have to scale up the Hilbert-Schmidt norm by a factor $\sqrt{d}$ (since $\|\cdot\|_2 \leq \sqrt{d}\|\cdot\|_\infty$). Then we may overestimate the effective size of the effect space. As a result, directly analyzing the linear functionals between S_1^d and S_∞^d is the most efficient way of calculating the fat-shattering dimension. We emphasise that with Bloch-sphere representation, all the quantum measurements/states are transformed into Euclidean space, where existing ML algorithms (e.g. perceptron learning algorithm, neural network, SVM, etc.) can be applied to conduct the learning tasks. It is also worth considering other metrics (e.g. Bures metric, or other ℓ_p norms in Bloch-sphere representation) and parameterization methods (e.g. Weyl operator basis, polarization operator basis, Majorana representation, etc.) in our quantum ML framework. We leave it as future work.

When learning an $(M+1)$-outcome POVM measurement $\{\Pi_j\}_{j=0}^M$, with $\sum_{j=0}^M \Pi_j = \mathcal{I}$, we can simply follow the procedure discussed so far. Now the training data set consists of $\{(\rho_i, \mathrm{Tr}(\mathbf{\Pi}\rho_i)\}_{i=1}^n$, where

$$\mathrm{Tr}(\mathbf{\Pi}\rho_i) := \left(\mathrm{Tr}(\Pi_1\rho_i), \ldots, \mathrm{Tr}(\Pi_n\rho_i)\right).$$

This is called *multi-target prediction* or *multi-label classification*. Each target Π_j can be independently learned by the individual function class $\mathcal{F}$.

It is worth mentioning that Gross and Flammia *et al.* [47, 48] proposed a quantum state tomography method via *compressed sensing*, which is similar to our setting of learning quantum states. The main goal of the work is to concentrate on states ρ that can be well approximated by density matrices of rank $r \ll d$ and to reconstruct a density matrix $\widehat{\rho}$ based on m randomly sampled Pauli operators. With certain constraint coefficients λ and $m \geq Crd\log^6 d$, they show

$$\|\widehat{\rho} - \rho\|_1 \leq C_0 r\lambda + C_1\|\rho_c\|_1,$$

where $\rho_c = \rho - \rho_r$ is the residual part and ρ_r is the best rank-r approximation to ρ.

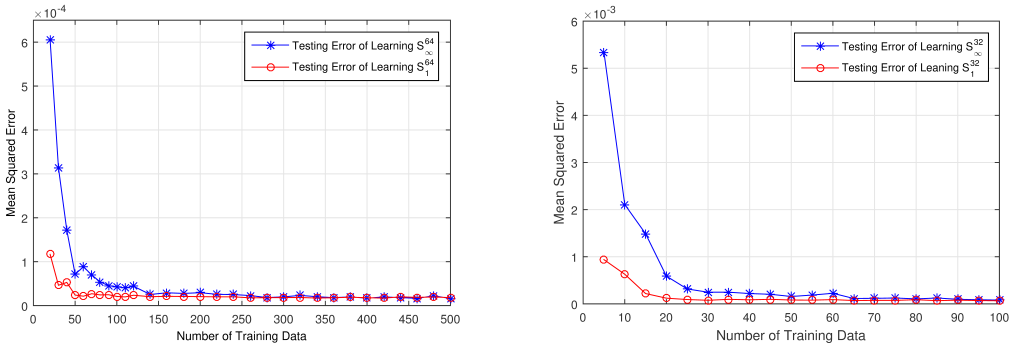
6.1. *Numerical Results*

In this section, we present the numerical results of the proposed neural network learning algorithms for learning Schatten ∞-norm matrices in S_∞^d and Schatten 1-norm matrices in S_1^d for $d = 32$ and 64, which correspond to learning 6-qubit quantum measurements and 6-qubit quantum states respectively. The training data in both cases are sampled according to the Haar measure (i.e. invariant from any unitary transformation) on the Hilbert space $\mathbb{C}^d$. The loss function is the squared error function, and gradient descent algorithms are used to find the optimal empirical minimizer.

The simulation results for $d = 64$ and $d = 32$ are shown in Figures (1a) and (1b). We can observe that the testing error for learning elements in S_1^d decays faster than those in S_∞^d. That is because the sample complexity for learning Schatten 1-norm matrices is logarithmically proportional to the sample complexity for learning Schatten ∞-norm matrices. We remark that $d = 64$ and $d = 32$ correspond to learning a six and five qubits quantum system, respectively. It can be observed from Figures (1a) and (1b) that the estimation error converges very quickly. Hence it is beneficial to consider quantum tomography using the ML approach because it significantly reduces the number of experiments needed.

Algorithm 1 Algorithms for Learning Matrices in Schatten ∞-norm Class

Input: Training data $(X_{\mathbf{r}_i}, \langle W_{\mathbf{n}}, X_{\mathbf{r}_i}\rangle)$, size n
for $i = 1$ **to** n **to**
  Transform $X_{\mathbf{r}_i}$ to Bloch vector $\mathbf{r}_i$
end for
Set the input vectors $\{\mathbf{r}_i\}$ and output variables $\{\langle W_{\mathbf{n}}, X_{\mathbf{r}_i}\rangle\}$
do gradient descent algorithms with boundary constraints to obtain the target coefficients $\mathbf{n}$
Transform the Bloch vector $\mathbf{n}$ to $W_{\mathbf{n}}$
Output: $W_{\mathbf{n}}$



(a) Learning matrices in S_∞^{64} and S_1^{64}. (b) Learning matrices in S_∞^{32} and S_1^{32}.

Fig. 2. The task of learning 64×64 and 32×32 matrices in Schatten ∞-norm and 1-norm classes.

7. Conclusions

Table 3. The Complexity Measures of The Quantum Learning Problems.

	Learning Quantum Measurements	Learning Quantum States
Pseudo Dimension	d^2	d^2-1
Fat-Shattering Dimension $\mathrm{fat}_{\mathcal{F}}(\epsilon)$	d/ϵ^2	$\log d/\epsilon^2$
Uniform Entropy Number $\log \mathcal{N}_2(\epsilon,\mathcal{F})$	d/ϵ^2	$\log d/\epsilon^2$
Rademacher Complexity $\mathcal{R}_n(\mathcal{F})$	$\sqrt{d}$	$\sqrt{\log d}$
Sample Complexity $m_{\mathcal{F}}(\epsilon,\delta)$	$\max\{d,\log(1/\delta)\}/\epsilon^2$	$\max\{\log d,\log(1/\delta)\}/\epsilon^2$

In this work, we developed a series of technical proofs to establish the fat-shattering dimension, Rademacher complexity, Gaussian complexity and entropy numbers for learning Schatten 1 and ∞ matrices. Moreover, we showed that the tasks of learning quantum measurements and states can be appropriately described into the framework of learning matrices with norm constraints, and hence answered their learnabilities. Our results show that the fat-shattering dimension of of learning (two-outcome) quantum measurements is $\min\left\{O\left(d/\epsilon^2\right), d^2\right\}$. On the other hand, the fat-shattering dimension for its dual problem—learning quantum states—is $\min\left\{O\left(\log d/\epsilon^2\right), d^2-1\right\}$. Our proof is entirely based on tools from classical learning theory, and provides an alternative proof for Aaronson's result [14]. Other important complexity measures for these two tasks are summarised in Table 3. Our results demonstrated that learning an unknown measurement is a more daunting task than learning an unknown quantum state. The intuition is that, since the effect space is much larger than the state space, it is reasonable that the fat-shattering dimension of the effect space is larger, too.

Finally, by exploiting general Bloch-sphere representation, we show that our learning problems are equivalent to a *neural network* so that classical ML algorithms can be applied to learn the unknown quantum measurement or state. Our work could provide a new viewpoint to the study of quantum state and measurement tomography. We also discuss connections between the quantum learning problems and other fields in QIS such as existence of QRA coding and quantum state discrimination. We hope that the development of our results would stimulate more theoretical studies in quantum statistical learning, and more applications in quantum information processing and related areas can be discovered.

Acknowledgements

MH is supported by an ARC Future Fellowship under Grant FT140100574.

References

1. V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995.
2. ——, *Statistical Learning Theory.* Wiley-Interscience, 1998.
3. T. M. Mitchell, *Machine Learning.* McGraw-Hill, 1997.
4. L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, nov 1984. [Online]. Available: http://dx.doi.org/10.1145/1968.1972
5. Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from Data: A Short Course.* AMLBook.com, 2012.

6. V. N. Vapnik and A. Y. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Probability and Its Applications*, vol. 26, no. 3, pp. 532–553, jan 1981. [Online]. Available: http://dx.doi.org/10.1137/1126059
7. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *Journal of ACM*, vol. 44, no. 4, pp. 615–631, jul 1997. [Online]. Available: http://dx.doi.org/10.1145/263867.263927
8. A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in functional spaces," *American Mathematical Society Translations*, vol. 17, no. 2, pp. 277–364, 1961.
9. P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 463–482, nov 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944944
10. L. Gurvits, "A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces," *Theoretical Computer Science*, vol. 261, no. 1, pp. 81–90, jun 2001.
11. C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "Rate of approximation results motivated by robust neural network learning," in *Proceedings of the sixth annual conference on Computational learning theory - COLT'93*. Association for Computing Machinery (ACM), 1993, pp. 303–309. [Online]. Available: http://dx.doi.org/10.1145/168304.168357
12. ——, "Rates of convex approximation in non-Hilbert spaces," *Constructive Approximation*, vol. 13, no. 2, pp. 187–220, jun 1997. [Online]. Available: http://dx.doi.org/10.1007/bf02678464
13. J. A. Tropp, "An introduction to matrix concentration inequalities," *Foundations and Trends in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015. [Online]. Available: http://dx.doi.org/10.1561/2200000048
14. S. Aaronson, "The learnability of quantum states," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 463, no. 2088, pp. 3089–3114, sep 2007. [Online]. Available: http://dx.doi.org/10.1098/rspa.2007.0113
15. A. Ambainis, A. Nayak, A. Ta-Shma, and U. Vazirani, "Dense quantum coding and quantum finite automata," *Journal of ACM*, vol. 49, no. 4, pp. 496–511, jul 2002. [Online]. Available: http://dx.doi.org/10.1145/581771.581773
16. G. Kimura, "The Bloch vector for n-level systems," *Physcis Letters A*, vol. 314, no. 56, pp. 339–349, 2003. [Online]. Available: http://dx.doi.org/10.1016/s0375-9601(03)00941-1
17. G. Kimura and A. Kossakowski, "The Bloch-vector space for n-level systems: The spherical-coordinate point of view," *Open Systems & Information Dynamics*, vol. 12, no. 3, pp. 207–229, jun 2005. [Online]. Available: http://dx.doi.org/10.1007/s11080-005-0919-y
18. M. Hayashi, K. Iwama, H. Nishimura, R. Raymond, and S. Yamashita, "(4,1)-quantum random access coding does not exist—one qubit is not enough to recover one of four bits," *New Journal Physics*, vol. 8, no. 8, pp. 129–129, aug 2006. [Online]. Available: http://dx.doi.org/10.1088/1367-2630/8/8/129
19. S. M. Kakade, S. Shalev-Shwartz, and A. Tewarj, "Regularization techniques for learning with matrices," *Journal of Machine Learning Research*, vol. 13, no. 1, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2343703
20. J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, aug 2011. [Online]. Available: http://dx.doi.org/10.1007/s10208-011-9099-z
21. P. Wittek, *Quantum Machine Learning: What Quantum Computing Means to Data Mining*. Academic Press, 2014.
22. M. Schuld, I. Sinayskiy, and F. Petruccione, "An introduction to quantum machine learning," *Contemporary Physics*, vol. 56, no. 2, pp. 172–185, oct 2014. [Online]. Available: http://dx.doi.org/10.1080/00107514.2014.964942
23. R. A. Servedio and S. J. Gortler, "Quantum versus classical learnability," in *Proceedings 16th Annual IEEE Conference on Computational Complexity*. IEEE Computer Society, 2000, pp. 138–148. [Online]. Available: http://dx.doi.org/10.1109/ccc.2001.933881
24. R. A. Servedio, "Separating quantum and classical learning," in *Automata, Languages*

and Programming. Springer Berlin Heidelberg, 2001, pp. 1065–1080. [Online]. Available: http://dx.doi.org/10.1007/3-540-48224-5_86

25. R. A. Servedio and S. J. Gortler, "Equivalences and separations between quantum and classical learnability," *SIAM Journal on Computing*, vol. 33, no. 5, pp. 1067–1092, jan 2004. [Online]. Available: http://dx.doi.org/10.1137/s0097539704412910
26. A. Atici and R. A. Servedio, "Improved bounds on quantum learning algorithms," *Quantum Information Process*, vol. 4, no. 5, pp. 355–386, oct 2005. [Online]. Available: http://dx.doi.org/10.1007/s11128-005-0001-2
27. D. Anguita, S. Ridella, F. Rivieccio, and R. Zunino, "Quantum optimization for training support vector machines," *Neural Networks*, vol. 16, no. 5-6, pp. 763–770, jun 2003. [Online]. Available: http://dx.doi.org/10.1016/s0893-6080(03)00087-x
28. E. Aïmeur, G. Brassard, and S. Gambs, "Quantum clustering algorithms," in *Proceedings of the 24th international conference on Machine learning - ICML '07*. ACM Press, 2007, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1145/1273496.1273497
29. K. L. Pudenz and D. A. Lidar, "Quantum adiabatic machine learning," *Quantum Information Process*, vol. 12, no. 5, pp. 2027–2070, nov 2012. [Online]. Available: http://dx.doi.org/10.1007/s11128-012-0506-4
30. E. Aïmeur, G. Brassard, and S. Gambs, "Quantum speed-up for unsupervised learning," *Machine Learning*, vol. 90, no. 2, pp. 261–287, aug 2012. [Online]. Available: http://dx.doi.org/10.1007/s10994-012-5316-5
31. N. Wiebe, D. Braun, and S. Lloyd, "Quantum algorithm for data fitting," *Physical Review Letters*, vol. 109, no. 5, p. 050505, aug 2012.
32. S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum algorithms for supervised and unsupervised machine learning," 2013. [Online]. Available: http://arxiv.org/abs/1307.0411
33. P. Rebentrost, M. Mohseni, and S. Lloyd, "Quantum support vector machine for big data classification," *Physical Review Letters*, vol. 113, no. 13, p. 130503, sep 2014. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.113.130503
34. S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum principal component analysis," *Nature Physics*, vol. 10, no. 9, pp. 631–633, jul 2014. [Online]. Available: http://dx.doi.org/10.1038/nphys3029
35. N. Wiebe, A. Kapoor, and K. M. Svore, "Quantum nearest-neighbor algorithms for machine learning," *Quantum Information & Computation*, vol. 15, no. 3-4, pp. 316–356, mar 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2871400
36. G. Wang, "Quantum algorithms for curve fitting," 2014. [Online]. Available: http://arxiv.org/abs/1402.0660
37. M. Schuld, I. Sinayskiy, and F. Petruccione, "The quest for a quantum neural network," *Quantum Information Process*, vol. 13, no. 11, pp. 2567–2586, aug 2014. [Online]. Available: http://dx.doi.org/10.1007/s11128-014-0809-8
38. S. Lloyd, S. Garnerone, and P. Zanardi, "Quantum algorithms for topological and geometric analysis of data," *Nature Communications*, vol. 7, p. 10138, jan 2016. [Online]. Available: http://dx.doi.org/10.1038/ncomms10138
39. A. W. Cross, G. Smith, and J. A. Smolin, "Quantum learning robust against noise," *Physical Review A*, vol. 92, no. 1, jul 2015. [Online]. Available: http://dx.doi.org/10.1103/physreva.92.012327
40. M. Schuld, I. Sinayskiy, and F. Petruccione, "Quantum computing for pattern classification," in *PRICAI 2014: Trends in Artificial Intelligence*. Springer International Publishing, 2014, pp. 208–220. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13560-1_17
41. N. Wiebe, A. Kapoor, and K. M. Svore, "Quantum deep learning," 2014. [Online]. Available: http://arxiv.org/abs/1412.3489
42. E. Aïmeur, G. Brassard, and S. Gambs, "Machine learning in a quantum world," in *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2006, pp. 431–442. [Online]. Available: http://dx.doi.org/10.1007/11766247_37

43. S. Gambs, "Quantum classification," 2008. [Online]. Available: http://arxiv.org/abs/0809.0444
44. M. Guţă and W. Kotłowski, "Quantum learning: Asymptotically optimal classification of qubit states," *New Journal Physics*, vol. 12, no. 12, p. 123032, dec 2010. [Online]. Available: http://dx.doi.org/10.1088/1367-2630/12/12/123032
45. A. Bisio, G. Chiribella, G. M. D'Ariano, S. Facchini, and P. Perinotti, "Optimal quantum learning of a unitary transformation," *Physical Review A*, vol. 81, p. 032324, mar 2010. [Online]. Available: http://dx.doi.org/10.1103/PhysRevA.81.032324
46. A. Bisio, G. M. D'ariano, P. Perinotti, and M. Sedlák, "Quantum learning algorithms for quantum measurements," *Physics Letters A*, vol. 375, no. 39, pp. 3425–3434, sep 2011. [Online]. Available: http://dx.doi.org/10.1016/j.physleta.2011.08.002
47. D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Physical Review Letter*, vol. 105, no. 15, oct 2010. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.105.150401
48. S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, "Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators," *New Journal Physics*, vol. 14, no. 9, p. 095022, sep 2012. [Online]. Available: http:/dx.doi.org/10.1088/1367-2630/14/9/095022
49. G. Sentís, J. Calsamiglia, R. Muñoz-Tapia, and E. Bagan, "Quantum learning without quantum memory," *Scientific Reports*, vol. 2, no. 708, oct 2012. [Online]. Available: http://dx.doi.org/10.1038/srep00708
50. G. Sentís, M. Guţă, and G. Adesso, "Quantum learning of coherent states," *EPJ Quantum Technology*, vol. 2, no. 1, jul 2015. [Online]. Available: http://dx.doi.org/10.1140/epjqt/s40507-015-0030-4
51. S. Lu and S. L. Braunstein, "Quantum decision tree classifier," *Quantum Information Process*, vol. 13, no. 3, pp. 757–770, nov 2013. [Online]. Available: http://dx.doi.org/10.1007/s11128-013-0687-5
52. E. C. Behrman and J. E. Steck, "A quantum neural network computes its own relative phase," in *2013 IEEE Symposium on Swarm Intelligence (SIS)*. IEEE, apr 2013, pp. 119–124. [Online]. Available: http://dx.doi.org/10.1109/sis.2013.6615168
53. M. V. Altaisky, N. E. Kaputkina, and V. A. Krylov, "Quantum neural networks: Current status and prospects for development," *Physics of Particles and Nuclei*, vol. 45, no. 6, pp. 1013–1032, nov 2014. [Online]. Available: http://dx.doi.org/10.1134/s1063779614060033
54. A. Monras, A. Beige, and K. Wiesner, "Hidden quantum markov models and non-adaptive read-out of many-body states," *Applied Mathematical and Computational Sciences*, vol. 3, no. 93, 2010.
55. J. Barry, D. T. Barry, and S. Aaronson, "Quantum partially observable Markov decision processes," *Physical Review A*, vol. 90, no. 3, p. 032311, sep 2014. [Online]. Available: http://dx.doi.org/10.1103/physreva.90.032311
56. L. A. Clark, W. Huang, T. M. Barlow, and A. Beige, "Hidden quantum markov models and open quantum systems with instantaneous feedback," in *Emergence, Complexity and Computation*. Springer, 2015, pp. 143–151. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10759-2_16
57. A. Monràs and A. Winter, "Quantum learning of classical stochastic processes: The completely positive realization problem," *Journal of Mathematical Physics*, vol. 57, no. 1, p. 015219, jan 2016. [Online]. Available: http://dx.doi.org/10.1063/1.4936935
58. A. Hentschel and B. C. Sanders, "Machine learning for precise quantum measurement," *Physical Review Letters*, vol. 104, no. 6, p. 063603, feb 2010. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.104.063603
59. E. Magesan, J. M. Gambetta, A. D. Córcoles, and J. M. Chow, "Machine learning for discriminating quantum measurement trajectories and improving readout," *Physical Review Letters*, vol. 114, no. 20, p. 200501, may 2015. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.114.200501
60. X.-D. Cai, D. Wu, Z.-E. Su, M.-C. Chen, X.-L. Wang, L. Li, N.-L. Liu, C.-Y. Lu, and J.-W. Pan, "Entanglement-based machine learning on a quantum computer,"

Physical Review Letters, vol. 114, no. 11, p. 110504, mar 2015. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.114.110504
61. L. Devroye, L. Gy orfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1997.
62. M. Anthony and P. L. Barlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
63. O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced Lectures on Machine Learning*, ser. Lecture Notes in Computer Science, vol. 3176. Springer, 2003, pp. 169–207.
64. S. Mendelson, "Geometric parameters in learning theory," in *Geometric Aspects of Functional Analysis*. Springer, 2004, pp. 193–235. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-44489-3_17
65. T. Hasite, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2011.
66. S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
67. B. K. Natarajan, "Occam's razor for functions," in *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, ser. COLT'93. ACM, 1993, pp. 370–376. [Online]. Available: http://doi.acm.org/10.1145/168304.168380
68. D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, vol. 100, no. 1, pp. 78–150, sep 1992. [Online]. Available: http://dx.doi.org/10.1016/0890-5401(92)90010-d
69. V. N. Vapnik, "Principles of risk minimization for learning theory," in *Advances in Neural Information Processing Systems (NIPS) 4*, 1992, pp. 831–838. [Online]. Available: http://papers.nips.cc/paper/506-principles-of-risk-minimization-for-learning-theory.pdf
70. S. Shalev-Shwartz, O. Shamir, N. rebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, pp. 2635–2670, dec 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1953019
71. S. Villa, L. Rosasco, and T. Poggio, "On learnability, complexity and stability," in *Empirical Inference*. Springer Berlin Heidelberg, 2013, pp. 59–69. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-41136-6_7
72. D. Pollard, *Convergence of Stochastic Processes*. Springer-Verlag, New York/Berlin, 1984.
73. M. J. Kearns and R. E. Schapire, "Efficient distribution-free learning of probabilistic concepts," in *Proceedings of the third annual conference on Computational learning theory - COLT'90*. Elsevier, 1990, p. 389. [Online]. Available: http://dx.doi.org/10.1016/b978-1-55860-146-8.50035-7
74. S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures on Machine Learning*. Springer, 2003, pp. 1–40. [Online]. Available: http://dx.doi.org/10.1007/3-540-36434-x_1
75. R. M. Dudley, "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes," *Journal of Functional Analysis*, vol. 1, no. 3, pp. 290–330, oct 1967. [Online]. Available: http://dx.doi.org/10.1016/0022-1236(67)90017-1
76. V. N. Sudakov, "Gaussian random processes and measures of solid angles in Hilbert space," *(Russian) Doklady Akademii Nauk SSSR*, vol. 197, pp. 412–415, 1971.
77. S. Mendelson and R. Vershynin, "Entropy and the combinatorial dimension," *Inventiones Mathematicae*, vol. 152, no. 1, pp. 37–55, apr 2003. [Online]. Available: http://dx.doi.org/10.1007/s00222-002-0266-3
78. V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York/Berlin, 1982.
79. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *Journal of ACM*, vol. 36, no. 4, pp. 929–965, oct 1989. [Online]. Available: http://dx.doi.org/10.1145/76359.76371
80. P. L. Bartlett, P. M. Long, and R. C. Williamson, "Fat-shattering and the learnability of

real-valued functions," *Journal of Computer and System Sciences*, vol. 52, no. 3, pp. 434–452, jun 1996. [Online]. Available: http://dx.doi.org/10.1006/jcss.1996.0033
81. R. M. Dudley, E. Giné, and J. Zinn, "Uniform and universal Glivenko-Cantelli classes," *Journal of Theoretical Probability*, vol. 4, no. 3, pp. 485–510, jul 1991. [Online]. Available: http://dx.doi.org/10.1007/bf01210321
82. J. B. Conway, *A Course in Operator Theory*, ser. Graduate Studies in Mathematics (Book 21). American Mathematical Society, 1999.
83. S. Mendelson and G. Schechtman, "The shattering dimension of sets of linear functionals," *The Annals of Probability*, vol. 32, no. 3A, pp. 1746–1770, jul 2004. [Online]. Available: http://dx.doi.org/10.1214/009117904000000388
84. E. Carlen, "Trace inequalities and quantum entropy: An introductory course," in *Entropy and the Quantum, Contemporary Mathematics*, ser. Entropy and Quantum. American Mathematical Society, Providence, RI, 2010, vol. 529, pp. 73–140. [Online]. Available: http://dx.doi.org/10.1090/conm/529/10428
85. T. Heinosaari and M. Ziman, *The Mathematical Language of Quantum Theory: From Uncertainty to Entanglement*. Cambridge University Press, 2012.
86. P. Busch, "Quantum states and generalized observables: A simple proof of Gleason's theorem," *Physical Review Letter*, vol. 91, no. 12, sep 2003. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.91.120403
87. F. Lust-Piquard and G. Pisier, "Non commutative Khintchine and Paley inequalities," *Arkiv för matematik*, vol. 29, no. 1-2, pp. 241–260, dec 1991. [Online]. Available: http://dx.doi.org/10.1007/bf02384340
88. U. Haagerup and M. Musat, "On the best constants in noncommutative Khintchine-type inequalities," *Journal of Functional Analysis*, vol. 250, no. 2, pp. 588–624, sep 2007. [Online]. Available: http://dx.doi.org/10.1016/j.jfa.2007.05.014
89. A. W. van der Vaart and J. A. Wellner, "Weak convergence," in *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996, pp. 16–28. [Online]. Available: http://dx.doi.org/10.1007/978-1-4757-2545-2_3
90. R. M. Dudley, *Uniform Central Limit Theorems*, ser. Cambridge Studies in Advanced Mathematics 63. Cambridge University, 1996.
91. E. Chitambar and M.-H. Hsieh, "Revisiting the optimal detection of quantum information," *Physical Review A*, vol. 88, p. 020302, aug 2013. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevA.88.020302
92. E. Chitambar, R. Duan, and M.-H. Hsieh, "When do local operations and classical communication suffice for two-qubit state discrimination?" *IEEE Transactions on Information Theory*, vol. 60, no. 3, pp. 1549–1561, mar 2014. [Online]. Available: http://dx.doi.org/10.1109/TIT.2013.2295356
93. S. Zhang and M. Ying, "Set discrimination of quantum states," *Physical Review A*, vol. 65, no. 6, jun 2002. [Online]. Available: http://dx.doi.org/10.1103/physreva.65.062322
94. N. Tomczak-Jaegermann, "Banach-Mazur distances and finite-dimensional operator ideals," in *Pitman Monograghs and Surveys in Pure and Applied Mathematics*. Longman Scientic and Technical, 1989, vol. 38.
95. S. J. Szarek, "Volume of separable states is super-doubly-exponentially small in the number of qubits," *Physical Review A*, vol. 72, p. 032304, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1103/physreva.72.032304
96. A. Ambainis, D. Leung, L. Mancinska, and M. Ozols, "Quantum random access codes with shared randomness," 2008. [Online]. Available: http://arxiv.org/abs/0810.2937
97. I. Bengtsson and K. Życzkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge University Press, 2008.
98. E. B. Davies, *Quantum Theory of Open Systems*. Academic Press, London, 1976.

Appendix A. Notation Table

See Table A.1 for the symbols listed in this paper.

Appendix B .The Justification of the Quantum Learning Model

In this section, we address two practical issues that may arise in our quantum learning setting: (1) Only the 'yes' ('1') or 'no' ('0') outcome can be observed rather than the outcome statistics.[‡‡] (2) The measurement apparatus is not perfect (e.g. there are measurement errors in the training data set). However, we will show that the sample complexities of the two scenarios remain the same (up to a Lipschitz constant). We also emphasise that the idea of the quantum learning model comes from Aaronson [14].

The output space consists of binary measurement outcomes rather than measurement statistics. In this case, the training sample (X_i, Y_i) equals to $(X_i, 1)$ with probability $\mathrm{Tr}(\Pi X_i)$, and $(X_i, 0)$ with probability $1 - \mathrm{Tr}(\Pi X_i)$. We show that the covering number remains the same as the training sample $(X_i, \mathrm{Tr}(\Pi X_i))$ considered in the quantum machine learning setting. Other complexity measures easily follow by the same argument. Assume the underlying loss function ℓ_f satisfies the Lipschitz condition, i.e. there exists $L > 0$ such that

$$|\ell_f(X,Y) - \ell_g(X,Y)| \leq \mathsf{L}\,|f(X) - g(X)|\,. \tag{B.1}$$

By denoting $p_X = \mathrm{Tr}(\Pi X)$, then the expected risk can be expressed as follows

$$\begin{aligned} L(f) &= \mathbb{E}_\mu \ell_f(X,Y) \\ &= \mathbb{E}_X \mathbb{E}_{Y|X} \ell_f(X,Y) \\ &= \mathbb{E}_X \left[p_X \ell_f(X,1) + (1-p_X)\ell_f(X,0)\right] \\ &=: \mathbb{E}_X \ell'_f(X,Y). \end{aligned}$$

In the third equality we use the fact that the '1' (resp. '0') outcome occurs with probability $p_X = \mathrm{Tr}(\Pi X)$ (resp. $1 - p_X$). In the last line we introduce the induced loss function $\ell'_f(X,Y) := [p_X \ell_f(X,1) + (1-p_X)\ell_f(X,0)]$. Then for all $X \in \mathcal{X}$, the distance between ℓ'_f and ℓ'_g can be calculated as

$$\begin{aligned} |\ell'_f(X,Y) - \ell'_g(X,Y)| &= |p_X\left(\ell_f(X,1) - \ell_g(X,1)\right) + (1-p_X)\left(\ell_f(X,0) - \ell_g(X,0)\right)| \\ &\leq p_X\,|\ell_f(X,1) - \ell_g(X,1)| + (1-p_X)\,|\ell_f(X,0) - \ell_g(X,0)| \\ &\leq p_X \cdot \mathsf{L}\,|f(X) - g(X)| + (1-p_X)\cdot \mathsf{L}\,|f(X) - g(X)| \\ &= \mathsf{L}\,|f(X) - g(X)|\,. \end{aligned}$$

The second inequality follows from the triangle inequality. The next line is due to the Lipschitz condition. The above relation shows that the distance $|\ell'_f - \ell'_g|$ can be upper bounded by $\mathsf{L}|f - g|$, which is exactly the same as the upper bound for $|\ell_f - \ell_g|$ (see Eq. (B.1)). Recall Definition 6, it is clearly that the covering numbers with respect to the induced loss function and the original loss function are bounded by the same quantity. Therefore, the generalisation error, Eq. (3) and the sample complexity do not change in this scenario.

There is noise involved in the measurement procedure. In this case, we assume that

[‡‡]The situation can also occur when only one measurement is performed.

the training sample is $(X, Y+\mathbf{n})$, where $Y \equiv \mathrm{Tr}(\Pi X)$ and $\mathbf{n}$ is a random variable that models the measurement error. Following the same reasoning, we can calculate the expected risk as follows

$$\begin{aligned} L(f) &= \mathbb{E}_\mu \ell_f(X, Y+\mathbf{n}) \\ &= \mathbb{E}_X \mathbb{E}_\mathbf{n} \ell_f(X, Y+\mathbf{n}) \\ &=: \mathbb{E}_X \ell'_f(X, Y). \end{aligned}$$

In the last line, we let $\ell'_f(X,Y) := \mathbb{E}_\mathbf{n} \ell_f(X, Y+\mathbf{n})$. Thus,

$$\begin{aligned} |\ell'_f(X,Y) - \ell'_g(X,Y)| &= |\mathbb{E}_\mathbf{n} \ell_f(X, Y+\mathbf{n}) - \mathbb{E}_\mathbf{n} \ell_g(X, Y+\mathbf{n})| \\ &\leq \mathsf{L} \mathbb{E}_\mathbf{n} \left[|f(X) - g(X)| \right] \\ &= \mathsf{L} \left| f(X) - g(X) \right|. \end{aligned}$$

Therefore, the original complexity measures (which depends on the distance of the loss function) and the induced sample complexity hold the same.

Appendix C. Learning Framework in Bloch-sphere Representation

When illustrating the state space on a finite dimensional Hilbert space $\mathbb{C}^d$, it is convenient to adopt a geometric parameterisation method called *Bloch-sphere representation* [97, 16, 17]. Here, we provide another point of view on our quantum learning framework. The key idea is to represent the quantum objects in a Euclidean space, wherein classical techniques of traditional ML can be applied. Although the Bloch-sphere representation method may not be as direct as the machinery we used in Sections 4 and 5, it does gain more insights into our quantum ML problems.

Based on the orthogonal basis $\{\mathcal{I}, \Lambda_1, \ldots, \Lambda_{d^2-1}\}$ of $SU(d)$, any state $\rho_\mathbf{r}$ on $\mathbb{C}^d$ can be represented in a *Bloch vector* $\mathbf{r}$ through:

$$\rho_\mathbf{r} = \frac{1}{d} \left(\mathcal{I} + c_d \sum_{i=1}^{d^2-1} r_i \Lambda_i \right) = \frac{1}{d} (\mathcal{I} + c_d \mathbf{r} \cdot \mathbf{\Lambda}), \tag{C.1}$$

where $c_d := \sqrt{\frac{d(d-1)}{2}}$ and the dot product corresponds to the conventional Euclidean inner product, and

$$r_i = \sqrt{\frac{d}{2(d-1)}} \, \mathrm{Tr}\left(\rho_\mathbf{r} \Lambda_i\right) \in \mathbb{R}, \; i = 1, \ldots, d^2 - 1.$$

Define the Bloch vector space as the set of Bloch vectors, which are representative of the valid states on $\mathbb{C}^d$ as

$$\Omega_d := \{\mathbf{r} \in \mathbb{R}^{d^2-1} : \mathbf{r} = \sqrt{\frac{d}{2(d-1)}} \, \mathrm{Tr}\left(\rho_\mathbf{r} \cdot \mathbf{\Lambda}\right)\}.$$

Now we calculate the linear functional of $E_\mathbf{n} \in \mathcal{E}_1$ acting on the state $\rho_\mathbf{r}$ (where $\mathcal{E}_k$ denotes

the convex hull of rank-k projection operators):

$$\begin{aligned}\mathrm{Tr}(P_{\mathbf{n}}\rho_{\mathbf{r}}) &= \mathrm{Tr}\left(\frac{1}{d^2}(\mathcal{I}+c_d\mathbf{r}\cdot\mathbf{\Lambda})(\mathcal{I}+c_d\mathbf{n}\cdot\mathbf{\Lambda})\right)\\ &= \mathrm{Tr}\left(\frac{1}{d^2}[\mathcal{I}+c_d(\mathbf{r}\cdot\mathbf{\Lambda}+\mathbf{n}\cdot\mathbf{\Lambda})+c_d^2(\mathbf{r}\cdot\mathbf{\Lambda})(\mathbf{n}\cdot\mathbf{\Lambda})]\right)\\ &= \frac{1}{d}+\frac{c_d^2}{d^2}\mathrm{Tr}\left((\mathbf{r}\cdot\mathbf{\Lambda})(\mathbf{n}\cdot\mathbf{\Lambda})\right)\\ &= \frac{1}{d}\left(1+(d-1)\mathbf{r}\cdot\mathbf{n}\right).\end{aligned}$$

Consequently, we have the affine functionals with elements in the convex hull of rank-1 projection operators, i.e.

$$\mathcal{F}_1=\{\rho_{\mathbf{r}}\mapsto\frac{1}{d}(1+(d-1)\mathbf{r}\cdot\mathbf{n}):\ \mathbf{n}\in\Omega_d\}.$$

In order to characterise the quantum effects associate with higher dimensional projection operators, it is useful to consider the algebraic properties of the projection operators. The set of projection operators on $\mathbb{C}^d$ is not a vector space but corresponds to an orthocomplemented lattice. Therefore, the sum of two projections, say P and Q, is a projection only when they are orthogonal, i.e. $PQ=QP=\mathcal{O}$. Based on this fact, now let $\{P_{\mathbf{n}_1},\ldots,P_{\mathbf{n}_d}\}$ be arbitrary mutually orthogonal rank-one projections on $\mathbb{C}^d$. To each of them, we associate a unit Bloch vector $\mathbf{n}_i$ such that $P_{\mathbf{n}_i}=\frac{1}{d}(\mathcal{I}+c_d\mathbf{n}_i\cdot\mathbf{\Lambda})$, $i=1,\ldots,d$. It can be verified by Eq. (C.1) that the Bloch vectors $\{\mathbf{n}_1,\ldots,\mathbf{n}_d\}$ form a $(d-1)$-dimensional (regular) simplex since the angle between any two Bloch vectors is $\theta(\mathbf{n}_i,\mathbf{n}_j)=\cos^{-1}(-\frac{1}{d-1})$. With a slight abuse of notation, denote a rank-k projection $P_{\mathbf{n}_{(k)}}$ as the summation of arbitrary k different projections from the set $\{P_{\mathbf{n}_1},\ldots,P_{\mathbf{n}_d}\}$. More formally, we denote an index set $I_k\subseteq\{1,\ldots,d\}$ with cardinality k, and $P_{\mathbf{n}_{(k)}}=\sum_{i\in I_k}P_{\mathbf{n}_i}$, where we adopt the convention that the empty sum is zero. Hence, when a rank-k projection $P_{\mathbf{n}_{(k)}}\in\mathcal{F}_k$ acts on the state $\rho_{\mathbf{r}}$, we have:

$$\mathrm{Tr}(P_{\mathbf{n}_{(k)}}\rho_{\mathbf{r}})=\sum_{i\in I_k}\frac{1}{d}(1+(d-1)\mathbf{r}\cdot\mathbf{n}_i)=k\cdot\frac{1}{d}(1+(d-1)\mathbf{r}\cdot\mathbf{n}_{(k)}),$$

where

$$\mathbf{n}_{(k)}:=\frac{1}{k}\sum_{i\in I_k}\mathbf{n}_i \tag{C.2}$$

is the centroid of the $(k-1)$-face of the simplex Δ_{d-1} subtended by the vectors $\{\mathbf{n}_i\}_{i\in I_k}$.The ℓ_2-norm of $\mathbf{n}_{(k)}$ can be calculated as the Euclidean distance from the center of the simplex Δ_{d-1} to the centroid of $(k-1)$-face; that is

$$\|\mathbf{n}_{(k)}\|_2:=r_{d,k}=\sqrt{\frac{d-k}{k(d-1)}}<1,\ k\in\{1,2,\ldots,d\}. \tag{C.3}$$

Intuitively, we can interpret the value $\mathrm{Tr}(P_{\mathbf{n}_{(k)}}\rho_{\mathbf{r}})$ as an operator $P_{\mathbf{n}_{(k)}}$ acting on the state $\rho_{\mathbf{r}}$, and then scaled by k.

Since every quantum effect can be composed into the extremal effects (i.e. projection operators) of the effect space [98]. We can represent $\mathrm{Tr}(E_{\mathbf{n}}\rho_{\mathbf{r}})$ for all $E_{\mathbf{n}} \in \mathcal{E}(\mathbb{C}^d)$ as:

$$\sum_{k=0}^{d} \frac{w_k \cdot k}{d} \left(1 + (d-1)\mathbf{r} \cdot \mathbf{n}_{(k)}\right) = \frac{1}{d} \left(n_0 + (d-1)\mathbf{r} \cdot \mathbf{n}\right), \tag{C.4}$$

where $\sum_{i=0}^{d} w_k = 1$, $0 \leq n_0 \leq d$ and $\|\mathbf{n}\|_2 \leq \max_{k\in\{0,1,\ldots,d\}} \sqrt{\frac{k(d-k)}{d-1}}$.

By utilising the bijection relationship of quantum state $\rho_{\mathbf{r}}$ and its corresponding Bloch vectors $\mathbf{r}$, we can associate the input space as the Bloch vector space, i.e. $\mathcal{X} = \Omega_d$. Denote the function class $\mathcal{F}_k$ as the linear functionals of $\mathcal{E}_k$ acting on $\rho_{\mathbf{r}}$. According to Eq. (C.2), we have:

$$\mathcal{F}_k = \mathrm{conv}\left(\{\mathbf{r} \mapsto \frac{k}{d}\left(1 + (d-1)\mathbf{r} \cdot \mathbf{n}_{(k)}\right)\}\right). \tag{C.5}$$

For the rank-0 projection operator, the class consists of only one element, i.e. $\mathcal{F}_0 = \{\mathcal{O}\}$. We can see from the above equation that the affine coefficient is fixed such that $\mathcal{F}_k$ consists of linear functionals. For the class of all quantum effects $\mathcal{F} = \mathcal{E}(\mathbb{C}^d)$, by Eq. (C.4) we have a similar result:

$$\mathcal{F} = \{\mathbf{r} \mapsto \frac{1}{d}\left(n_0 + (d-1)\mathbf{r} \cdot \mathbf{n}\right) \,:\, \mathbf{n} \in \mathbb{R}^{d^2-1}\},\ \mathbf{r} \in \Omega_d,$$

where n_0 can be upper bounded by d and $\|\mathbf{n}\|_2$ can be bounded by $k \cdot r_{d,k} = \sqrt{\frac{k(d-k)}{d-1}}$. Clearly, $\mathcal{F} = \mathcal{E}(\mathcal{H})$ is the function class consisting of the affine functionals. However, we can easily convert this formulation into a linear form by letting $\tilde{\mathbf{r}} = [1, \mathbf{r}]$, and $\tilde{\mathbf{n}} = [n_0, \mathbf{n}]$. The intuition behind this is that when characterising the learnability of quantum measurements, all we need is to bound the complexity measures of the class of linear functionals.

Appendix D. Neural Networks

Here we briefly introduce the theory of *Neural Networks*. Readers may refer to Ref. [62] for more details. The basic computing unit in a neural network is the (simple) *perceptron* (see Fig. D.1), which computes a function from $\mathbb{R}^d$ to $\mathbb{R}$:

$$f(\mathbf{r}) = \sigma(\mathbf{v} \cdot \mathbf{r} + v_0),$$

for input vector $\mathbf{r} \in \mathbb{R}^d$, where $\mathbf{v} = (v_1, \ldots, v_d) \in \mathbb{R}^d$ and $v_0 \in \mathbb{R}$ are adjustable parameters, or *weights* (the particular weight v_0 being known as the *threshold*). The function $\sigma : \mathbb{R} \to \mathbb{R}$ is called the *activation function*. In the scenario of binary classification, the activation function may be chosen as the sign function; in the case of real-value outputs, $\sigma(\cdot)$ may satisfy some Lipschitz conditions. Note that the decision boundary of the binary perceptrons is the affine subspace of $\mathbb{R}^d$ defined by the equation $\mathbf{v} \cdot \mathbf{r} + v_0 = 0$.

When using a simple perceptron for a binary classification problem, the *perceptron learning algorithm* (PCA) finds adequate parameters $\mathbf{v}$ and v_0 to well fit the training data set. The algorithm starts from an arbitrary initial parameter and updates the parameter when there are misclassified data. For example, if now the function computes $(\mathbf{r}, y)$ (with $\mathbf{r} \in \mathbb{R}^d$ and $y \in \{0, 1\}$), the algorithm adds $\eta(y - f(\mathbf{r}))[\mathbf{r}, -1]$ element-wise to $[\mathbf{v}, v_0]$, where η is a fixed step constant. PCA iterates until a termination criterion is reached.

The second example is the *two-layer networks* (also called *single-hidden layer nets*) (see Fig. D.2). The network can compute a function of the form

$$f(\mathbf{r}) = \sum_{i=1}^{k} w_k \sigma(\mathbf{v}_i \cdot \mathbf{r} + v_{0i}) + w_0,$$

where $w_i \in \mathbb{R}, i = 0, \ldots, k$, are the output weights, $[\mathbf{v}_i, v_{0i}]$ are the input weights. The positive integer k is the number of hidden units. One can use a 'gradient descent' procedure to adjust the parameters to minimize the squared errors over the training data.

Fig. D.1. Consider a qubit system. A measurement in $\mathcal{F}_1$ can be characterised by a simple perceptron with 3-dimensional input data and the activation function σ. The '1' node is a bias node and v_0 is the corresponding bias weight. The input vector is the Bloch vector $\mathbf{r} \in \Omega_2$. The output variable $y = f(\mathbf{r})$ is computed by the simple perceptron. Hence the problem of learning an unknown measurement $\Pi \in \mathcal{F}_1$ is to infer the simple perceptron, i.e. the values of $\{v_i\}_{i=1}^4$.

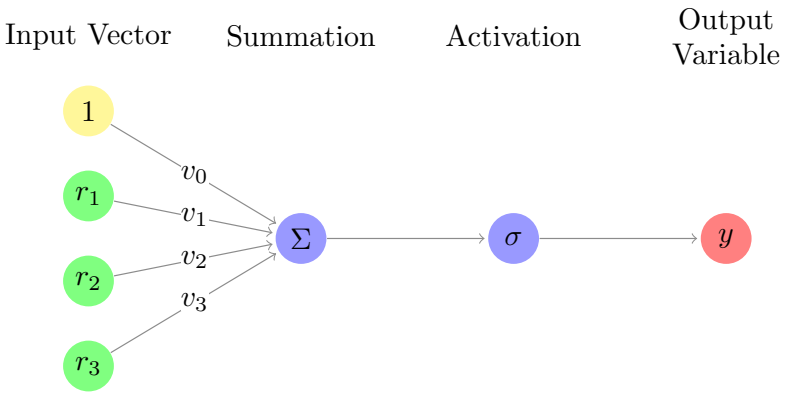


Fig. D.2. Single-hidden layer net computes 3-dimensional input data with activation function σ and three hidden units, which correspond to $\mathcal{F}_i$ for $i = 0, 1, 2$. The value v_{0k} corresponds to the bias weight of the k-th hidden unit. The single-hidden net represents a quantum measurement in $\mathcal{E}(\mathbb{C}^2)$.

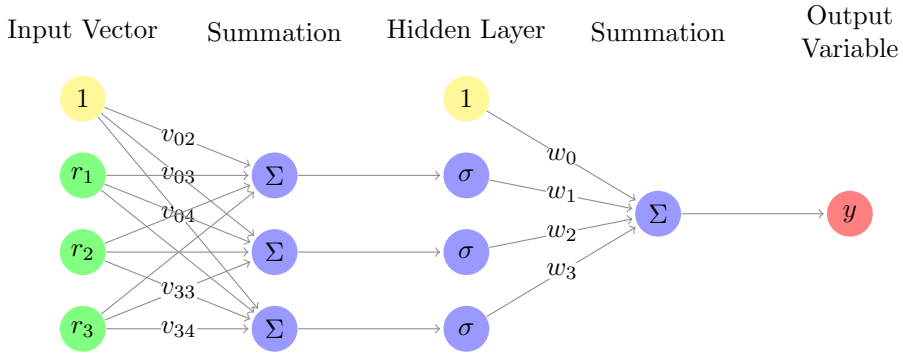


Table A.1. Summary of Notation

Notation	**Mathematical Meaning**
$\mathcal{H}$	the (separable) Hilbert space
d	the dimension of the linear space
$\mathbb{R}$, $\mathbb{N}$	the set of real numbers and positive integers
$\mathbb{C}^d$	the linear space of d-dimensional complex vectors
$\mathbb{M}_d$	the set of all self-adjoint operators on $\mathbb{C}^d$
Tr	the trace function on $\mathbb{M}_d$
$A^\dagger$	the conjugate transpose of A
$\langle A, B\rangle$	$= \mathrm{Tr}(B^\dagger A)$, the Hilbert-Schmidt inner product on $\mathbb{M}_d$; also stands for conventional inner product on $\mathbb{C}^d$
$\mathcal{B}(\mathcal{H})$	the set of bounded operators on $\mathcal{H}$
$\mathcal{T}(\mathcal{H})$	the set of trace class operators (i.e. finite trace) on $\mathcal{H}$
$\mathcal{O}$	the zero operator on $\mathcal{H}$.
$\mathcal{I}$	the identity operator on $\mathcal{H}$.
$A \succeq B$	$= A - B \succeq \mathcal{O}$, the standard partial ordering
$\|M\|_p$	the Schatten p-norm on $\mathbb{M}_d$, which reduces to the ℓ_p norms on $\mathbb{C}^d$.
S_p^d	$=\{M \in \mathbb{M}_d : \|M\|_p \leq 1\}$, the unit ball of Schatten p-class
$\|\varphi\rangle$	the unit vector on $\mathcal{H}$
ρ, σ	the quantum state on $\mathcal{H}$, i.e. $\rho = \rho^\dagger \in \mathcal{T}(\mathcal{H})$, with $\mathrm{Tr}(\rho) = 1$
E, Π	the POVM element on $\mathcal{H}$, i.e. $\mathcal{O}$
$\mathcal{Q}(\mathcal{H})$	state space, the set of all states on $\mathcal{H}$
$\mathcal{E}(\mathcal{H})$	effect space, the set of all POVM elements on $\mathcal{H}$
$\mathcal{X}$	the input space, or called the instances domain (the set)
$\mathcal{Y}$	the output space, or called the labels domain (the set)
$\mathcal{Z}$	$= \mathcal{X} \times \mathcal{Y}$
$\mathcal{F}$	the hypothesis set of functions $f : \mathcal{X} \to \mathcal{Y}$
μ	a distribution on $\mathcal{Z}$
Z_n	a training data set of n elements independently according to μ
$\ell_f : \mathcal{Z} \to (0, \infty)$	loss function
Pr, $\mathbb{E}$	probability and expectation of a random variable
$L(f)$	$= \mathbb{E}_\mu[\ell_f(X, Y)]$, the ensemble error
$\widehat{L}_n(f)$	$= 1/n \sum_{i=1}^n \ell_f(X_i, Y_i)$, the empirical error over the training data set Z_n
$\mathrm{VCdim}(\mathcal{F})$	Vapnik-Chervonenkis dimension of the function class $\mathcal{F}$
$\mathrm{Pdim}(\mathcal{F})$	pseudo dimension of the function class $\mathcal{F}$
$\mathrm{fat}_\mathcal{F}(\epsilon)$	fat-shattering dimension of the function class $\mathcal{F}$ with $\epsilon > 0$
$\underline{\mathrm{fat}}_\mathcal{F}(\epsilon)$	level fat-shattering dimension of the function class $\mathcal{F}$ with $\epsilon > 0$
$\mathcal{N}(\epsilon, \mathcal{F}, \tau)$	covering number of $\mathcal{F}$ with metric τ and $\epsilon > 0$
$\log \mathcal{N}(\epsilon, \mathcal{F}, \tau)$	entropy number
$\mathcal{R}_n(\mathcal{F})$	Rademacher complexity of the function class $\mathcal{F}$ on Z_n
γ_i	uniformly $\{+1, -1\}$-valued random variables or called Rademacher variables
O	the big O notation; $f = O(g)$ means $f(x) \leq cg(x)$ for some positive c, x_0 and all $x \geq x_0$
$A \lesssim B$	$= A \leq cB$ for some constant c
$A \simeq B$	both $A \lesssim B$ and $A \gtrsim B$