THE LEARNABILITY OF UNKNOWN QUANTUM MEASUREMENTS

HAO-CHUNG CHENG, MIN-HSIU HSIEH, AND PING-CHENG YEH

Abstract

Machine Learning (ML) [1–3] aims to systematically devise algorithms for machines to infer the input-output relationship of an unknown functional from historic data. More precisely, the key elements in ML are an *input space* $\mathcal{X}$ (a measurable space); an *output space* $\mathcal{Y}$ (a closed subset of real line), and a collection of functions, denoted as the *hypothesis set* $\mathcal{F}$. The learning machine seeks to construct a function $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{F}$ that approximates the unknown functional from the training data set $\{(X_i, Y_i)\}_{i=1}^n$, which are randomly and independently drawn from some measure μ on $\mathcal{X} \times \mathcal{Y}$. The main focuses of ML are: (i) *computational complexity* which measures the 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 [4]. In other words, a sample complexity problem investigates how many samples (e.g. the size of the training data set) are required to bound the *generalization error*. Denote a loss function $l : \mathbb{R} \times \mathcal{Y} \to [0, +\infty)$. The generalization error, defined as the difference between the *in-sample error* $R_n(f) := \frac{1}{n}\sum_{i=1}^n l(f(X_i), Y_i)$ and the *out-of-sample error* $R(f) := \mathbb{E}_\mu l(f(X), Y)$, is one of the most popular figures of merit in ML because it indicates how well the training set can approximate the input space under the function f. Eventually, we are interested in whether a quantity, denoted as $m_{\mathcal{F}}(\epsilon, \delta)$, exists such that given $n \geq m_{\mathcal{F}}(\epsilon, \delta)$, for every $0 < \epsilon, \delta < 1$ and any probability measure μ,

$$\Pr\left\{\sup_{f\in\mathcal{F}} \left|R(f) - R_n(f)\right| \geq \epsilon\right\} \leq \delta. \tag{1}$$

This quantity $m_{\mathcal{F}}(\epsilon, \delta)$ is called the sample complexity of the hypothesis set $\mathcal{F}$ with accuracy ϵ and confidence δ. One of the biggest achievements in ML [5–7] shows that the sample complexity can be characterised by a complexity measure—*fat-shattering dimension*, denoted as $\mathrm{fat}_{\mathcal{F}}(\epsilon)$, which quantifies the "effective size" of the hypothesis set[1]. Intuitively, a large $\mathrm{fat}_{\mathcal{F}}(\epsilon)$ implies a "richer" hypothesis set which is capable of approximating the target function well but it requires more samples to reduce the generalization error. As a result, the fat-shattering dimension effectively determines the information-theoretic efficiency of a learning algorithm and the rate of the uniform convergence in Eq. (1).

Quantum Information Processing (QIP) has achieved significant breakthroughs recently [8], and researchers have begun to explore whether QIP can advance other subjects of classical computer science. Consequently, the interdisciplinary area of quantum machine learning has emerged and attracted substantial interests lately. The central problems are two-fold. The first kind of problems investigates how quantum machines can serve to accelerate the classical ML processes to improve the computational efficiency, or to reduce the sample complexity by transforming classical training data into special sets of quantum states. We call this line of research as *Quantum Computational Learning* [9–20]. On the other hand, certain fundamental quantum problems, such as inference of unknown quantum states or operations (also known as state/process tomography) or the hidden structure of the underlining quantum system, fits well into the setting of statistical learning theory. We term this line of research as *Quantum Statistical Learning* [21–29]. The majority of previous works in quantum machine learning focused on computational issues of a learning algorithm. The issue of sample complexity exhibited in original quantum learning setting, e.g. state/process tomography, was rarely touched [22].

Any quantum statistical learning problem can be similarly formulated as that in the first paragraph; namely, the input and output space $\mathcal{X}$ and $\mathcal{Y}$ could be any subset of $\mathbb{C}^{d\times d}$ (instead of $\mathbb{R}^d$ in classical ML), and the hypothesis set $\mathcal{F}$ could contain matrix-valued functions. Such a generalization encompasses all system models in previous works [9–29]. Under this framework, we consider the problem of *learning an unknown quantum measurement*, and we mainly focus on learning a two-outcome measurement. For multi-outcome POVMs, the results can be easily

[1]Given $\epsilon > 0$, we say a set $\mathcal{S} = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ is ϵ-shattered by $\mathcal{F}$ if for every subset $B \subseteq \mathcal{S}$ there is some function $f \in \mathcal{F}$ and constant c_i for which $f(x_i) \geq c_i + \epsilon$ if $x_i \in B$, and $f(x_i) \leq c_i - \epsilon$ if $x_i \notin B$. The fat-shattering dimension is defined to be the largest cardinality of a ϵ-shattered set.

generalized[2]. Note that each two-outcome measurement is completely characterized by a positive semi-definite matrix $0 \leq E \leq \mathcal{I} \in \mathbb{C}^{d\times d}$ (will be denoted as $E \in [0, \mathcal{I}]$). Equivalently, we can consider the hypothesis set to be $\mathcal{F} = \{E : E \in [0, \mathcal{I}]\}$. We can apply a sequence of quantum states $\{\rho_1, \ldots, \rho_n\}$ (input variables in $\mathcal{X}$) through the measurement apparatus. Suppose that the output statistics $\{\mathrm{Tr}(E\rho_1), \ldots, \mathrm{Tr}(E\rho_n)\}$ (output variables in $\mathcal{Y}$) is also available. With the given set of training data $\{(\rho_i, \mathrm{Tr}(E\rho_i))\}_{i=1}^n$, our goal is to examine the sample complexity of this problem, i.e., what is the fat-shattering dimension. Our main result is the following quantification of the fat-shattering dimension for learning an unknown quantum measurement.

Theorem 1 (Main Result: Fat-shattering Dimension of Learning Quantum Measurements)**.** *Assume the hypothesis set $\mathcal{F}$ consists of all possible two-outcome measurements on $\mathbb{C}^d$, i.e., $E \in [0, \mathcal{I}]$. For all $0 < \epsilon < 1/2$, and $d \geq 2$, we have*

$$fat_{\mathcal{F}}(\epsilon) = \min\{O(d/\epsilon^2), d^2\}.$$

The key ingredients used in the proof are the powerful classical ML results from Banach space theory [30] (see Lemma 1) and the *noncommutative Khintchine inequalities* [31] in Random Matrix Theory (Lemma 2).

Lemma 1 (Mendelson and Schechtman [30])**.** *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 $a_1, \ldots, a_n \in \mathbb{R}$,*

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

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

Lemma 2 (Noncommutative Khintchine inequalities [31])**.** *Let $\{x_i\}_{i=1}^n$ be deterministic $d \times d$ matrices, each ϵ_i takes $\{+1, -1\}$-value uniformly and independently (called Rademacher variable). Then*

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

where $\approx_p$ means that the equality holds up to an absolute constant depending on p.

For technical simplicity, we consider "symmetrization" of the state space and the effect space, which will be convenient for our derivations in Banach space theory and convex analysis. In other words, the unit ball of Schatten 1-class is the symmetric convex hull of the state space, i.e., $S_1^d = \mathrm{conv}(-\mathcal{S}(\mathbb{C}^d) \cup \mathcal{S}(\mathbb{C}^d))$, where $\mathcal{S}(\mathbb{C}^d) = \{\rho : \rho \geq 0, \mathrm{Tr}(\rho) = 1\}$ and $\mathrm{conv}(\cdot)$ denotes the convex hull operation. Similarly, we have $S_\infty^d = \mathrm{conv}(-\mathcal{E}(\mathbb{C}^d) \cup \mathcal{E}(\mathbb{C}^d))$, where $\mathcal{E}(\mathbb{C}^d) = \{E : 0 \leq E \leq \mathcal{I}\}$, and is a dual unit ball of S_1^d. Thus the input space $\mathcal{X} \subset S_1^d$ and the hypothesis set $\mathcal{F}$ consists of linear functionals with elements in S_∞^d. Choose the constants $\{a_i\}$ in Lemma 1 to be $\{\pm 1\}$-value random variables (the Rademacher variables $\{\epsilon_i\}$), then Lemma 2 is applicable. Theorem 1 follows with some direct algebra. (See supplemental material, Theorem 3.1, for a detailed proof.) We also showed the bound is tight by the following proposition.

Proposition 1. *There exists a set of d quantum states on $\mathbb{C}^d$ that can be $1/2$-shattered by two-outcome POVMs.*

Our investigation of learning an unknown measurement is largely motivated by the problem of learning an unknown quantum state, pioneered by Aaronson [22], where the training data set is a collection of two-outcome measurements and the corresponding statistics; namely, $\{(E_i, \mathrm{Tr}(E_i\rho))\}_{i=1}^n$. In fact, learning an unknown quantum state is a *dual problem* to learning an unknown measurement, rooted from the fact that S_1^d and S_∞^d are dual unit balls in Banach Space.

Proposition 2 (Duality)**.** *Learning quantum measurements and learning quantum states are dual to each other.*

Proposition 2 together with Theorem 1 hinted a new proof, completely from known results in classical ML, of Aaronson's bound on the sample complexity of learning an unknown quantum state. By using Lemma 1 and the matrix concentration inequality in Lemma 3, we can re-derive the fat-shattering dimension, stated in Theorem 2 below, for learning quantum states in [22]. Note that Lemma 3 ($p = \infty$) and Lemma 2 play similar roles in their respective proofs .

[2]For the scenario of learning multi-outcome measurements, each POVM element can be considered as a two-outcome POVM. Therefore, every POVM element can be learned independently.

Lemma 3 (Rademacher Series [32]). *Consider a finite sequence $\{x_i\}$ of deterministic Hermitian matrices with dimension d, and let $\{\epsilon_i\}$ be independent Rademacher variables. Form the matrix Rademacher series $Y = \sum_i \epsilon_i x_i$. Then*

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

where $\sigma^2 = \sigma^2(Y) = \|\mathbb{E}\left(Y^2\right)\|_\infty$ is the variance parameter.

Theorem 2 (Fat-shattering Dimension of Learning Quantum States). *Assume the hypothesis set $\mathcal{F}$ consists of all possible quantum states on $\mathbb{C}^d$. For all $0 < \epsilon < 1/2$, and $d \geq 2$, we have*

$$fat_{\mathcal{F}}(\epsilon) = \min\{O(\log d/\epsilon^2), d^2 - 1\}.$$

Note that Aaronson's original proof [22] used the entropic inequality from Quantum Random Access (QRA) Codes [33] to prove $\text{fat}_{\mathcal{F}}(\epsilon) = O(m/\epsilon^2)$ for learning m qubits.

Applications of our results:

- Existence of QRA codes: An (n, m, p)-QRA coding is a mapping that encodes each n-bit string $y = y_1 \cdots y_n$ into an m-qubit ρ_y such that there exist two-outcome POVM $\{E_i\}_{i=1}^n$ satisfying $\text{Tr}(E_i\rho_y) \geq p$ if $y_i = 1$; $\text{Tr}(E_i\rho_y) \leq p$ if $y_i = 0$ (i.e. $\{E_i\}_{i=1}^n$ is $(p - 1/2)$-shattered by ρ_y). However, the result in Theorem 2 (i.e. $\text{fat}_{\mathcal{F}}(\epsilon) \leq d^2 - 1$) shows that there is no d^2 quantum measurements that can be shattered (by the hypothesis of all states). In other words, there exists no $(d^2 = 2^{2m}, m, p)$-QRA coding for $p > 1/2$, which coincides Hayashi *et al.*'s result [34].
- Quantum state discrimination: There is an operational interpretation of the fat-shattering dimension in terms of ambiguous set discrimination [35] as follows. Define a 2ϵ-separable ensemble, if two arbitrary subsets of the ensemble can be discriminated with the error probability no greater than $1/2 - \epsilon$ (e. g. $\text{Tr}(E\rho) \leq 1/2 - \epsilon$ for a misclassified state ρ by the discriminator E). We show that the fat-shattering dimension $\text{fat}_{\mathcal{F}}(\epsilon)$ for learning quantum measurements is equivalent to the maximum cardinality of a 2ϵ-separable ensemble. In other words, the fat-shattering dimension guarantees a set of quantum states that can be discriminated into two subsets with the worst error probability no greater than $1/2 - \epsilon$.

Efficient implementation of the learning algorithm: Finally, in addition to the theoretical bounds on sample complexity, we also proposed an efficient implementation of the learning algorithm that works for both learning an unknown state and measurement. The idea comes from the observation that, when formulating our problems using general Bloch-sphere representation [36–38], the hypothesis set of rank-k projection-valued measures (PVMs) $\mathcal{F}_k$ can be written as a set of linear functionals, and every POVM element can be decomposed as the convex combination of rank-k PVMs, $k = 1, \cdots, d$, [39]:

$$\mathcal{F} = \text{conv}\left(\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_d\right) = \{f : \mathbf{r} \mapsto \frac{1}{d}(n_0 + (d-1)\mathbf{r} \cdot \mathbf{n})\},$$

where $\mathbf{r}$ is the Bloch vector of a quantum state; n_0 and $\mathbf{n}$ parameterises the functions in the hypothesis set $\mathcal{F}$ (each $[n_0, \mathbf{n}]$ corresponds to a two-outcome quantum measurement). Using the language from the theory of *neural network* [40], each rank-k PVMs $\mathcal{F}_k$ is called a *linear perceptron* or a *single-layer neural network*, and $\mathcal{F}$ is thus a *two-layer neural network* (the two-layer means a linear combination of a set of single-layer networks). Since the operations of a neural network depend on each computing unit (i.e. linear perceptron), it suffices to tune all linear coefficients of the network to complete the learning process. As a result, by means of general Bloch-sphere representation, classical neural network algorithms can be easily applied to perform the quantum ML procedures.

Contributions of our work. We summarize the contributions of this work. First, we start from the standpoint of statistical learning theory and show that any quantum statistical learning problem can also be formulated using the standard language of ML. Then, we proved an upper bound for the sample complexity problems for learning an unknown quantum measurement. The upper bound, given by the fat-shattering dimension, is linearly proportional to the dimension of the underlining Hilbert space. Second, for learning an unknown quantum state, we show that it becomes a dual problem to learning an unknown quantum measurement. We provide an alternative proof solely from classical statistical learning theory, which reproduces Aaronson's work. Third, 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 performing this particular quantum ML task. Our work could provide a new viewpoint to the study of measurement or process tomography. Finally, we discuss connections between the quantum learning problems and other fields in QIP such as existence of QRA Codes and quantum state discrimination. We hope that the development of our results would stimulate more theoretical studies in quantum statistical learning and find more applications in quantum information processing and related areas.

References

[1] T. Mitchell, *Machine Learning*. McGraw-Hill Education, 1997.

[2] V. N. Vpnik, *Statistical Learning Theory*. Wiley, 1998.

[3] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from Data*. AMLBook, 2012.

[4] L. G. Valiant, "A theory of the learnable," *Comm. ACM*, vol. 27, pp. 1134–1142, 1984.

[5] P. L. Bartlett, P. M. Long, and R. C. Williamson, "Fat-shattering and the learnability of real-valued functions," *J. Comput. System Sci.*, vol. 52, no. 3, pp. 434–452, 1996.

[6] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *J. ACM*, vol. 44, no. 4, pp. 616–631, 1997.

[7] R. V. S. Mendelson, "Entropy and the combinatorial dimension," *Inventiones Mathematicae*, vol. 152, pp. 37–55, 2003.

[8] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[9] R. Servedio and S. J. Gortler, "Quantum versus classical learnability," in *16th IEEE Conference on Computational Complexity (CCC 2001)*, 2001.

[10] R. Servedio, "Separating quantum and classical learning," in *28th International Conference on Automata, Language and Programming (ICALP 2001)*, 2001.

[11] R. Servedio and S. Gortler, "Equivalences and separations between quantum and classical learnability," *SIAM Journal on Computing*, vol. 31, no. 5, 2004.

[12] D. Anguita, S. Ridella, F. Riviecco, and R. Zunino, "Quantum optimization for training support vector machines," *Neural Networks*, vol. 16, no. 1, pp. 763–770, 2003.

[13] E. Aïmeur, G. Brassard, and S. Gambs, "Quantum clustering algorithms," in *24th Annual International Conference on Machine Learning (ICML)*, 2007.

[14] Aïmeur, G. Brassard, and S. Gambs, "Quantum speed-up for unsupervised learning," *Machine Learning*, vol. 90, pp. 261–287, 2013.

[15] S. Lloyd, M. Mohseni, and P. Rebentrost. (2013) Quantum algorithms for supervised and unsupervised machine learning. [Online]. Available: http://arxiv.org/abs/1307.0411

[16] P. Rebentrost, M. Mohseni, and S. Lloyd. (2013) Quantum support vector machine for big feature and big data classification. [Online]. Available: http://arxiv.org/abs/1307.0471

[17] S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum principal component analysis," *Nature Physics*, arXiv: quant-ph/1307.0401.

[18] K. L. Pudenz and D. A. Lidar, "Quantum adiabatic machine learning," *Quantum Inf Process*, vol. 12, pp. 2027–2070, 2013.

[19] N. Wiebe, A. Kapoor, and K. Svore. (2014) Quantum nearest-neighbor algorithms for machine learning. arXiv: quant-ph/1401.2142.

[20] G. Wang. (2014) Quantum algorithms for curve fitting. arXiv: quant-ph/1402.0660.

[21] E. Aïmeur, G. Brassard, and S. Gambs, "Machine learning in a quantum world," in *Can. AI 2006*, 2006, pp. 431–442.

[22] S. Aaronson, "The learnability of quantum states," *Proc. R. Soc. A*, vol. 463, no. 2088, pp. 3089–3144, 2007.

[23] S. Gambs. (2008) Quantum classification. arXiv: quant-ph/0809.0444.

[24] M. Guta and W. Kotlowski, "Quantum learning: optimal classification of qubit stats," *New J. Phys*, vol. 12, 2010.

[25] A. Bisio, G. Chiribella, G. M. D'Ariano, S. Facchini, and P. Perinotti, "Optimal quantum learning of a unitary transformation," *Phys. Rev. A*, vol. 81, 2010.

[26] A. Bisio, G. M. D'ariano, P. Perinotti, and M. Sedlak, "Quantum learning algorithms for quantum measurements," *Phys. lett. A*, vol. 375, pp. 3425–3434, 2011.

[27] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, 2010.

[28] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, "Quantum tomography via compressed sensing: Error bounds, sample complexity, and efficient estimators," *New J. Phys.*, vol. 14, 2012.

[29] S. Lu and S. Braunstein, "Quantum decision tree classifier," *Quantum Inf. Process*, vol. 13, pp. 757–770, 2014.

[30] S. Mendelson and G. Schechtman, "The shattering dimension of sets of linear functionals," *Annals of Probability*, vol. 32, pp. 1746–1770, 2004.

[31] F. Lust-Piquard and G. Pisier, "Non commutative Khintchine and Paley inequalities," *Arkiv för Matematik*, vol. 29, pp. 241–260, 91.

[32] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[33] A. Ambainis, A. Nayak, A. Ta-Shma, and U. V. Vazirani, "Quantum dense coding and quantum finite qutomata," *J. ACM*, vol. 49, pp. 496–511, 2002.

[34] 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 J. Phys.*, vol. 8, no. 129, 2006.

[35] S. Zhang and M. Ying, "Set discrimination of quantum states," *Phys. Lett. A*, vol. 65, p. 062322, 2002.

[36] I. Bengtsson and K. Życzkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge University Press, 2008.

[37] G. Kimura, "The Bloch vector for n-level systems," *Phys. Lett. A*, vol. 314, no. 56, pp. 339 – 349, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0375960103009411

[38] 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

[39] E. B. Davies, *Quantum Theory of Open Systems*. Academic Press, London, 1976.

[40] M. Anthony and P. L. Barlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.