A general VC dimension upper bound for quantum circuit learning

Chih-Chieh Chen[1,a)] Masaru Sogabe[1] Kodai Shiba[1,2] Katsuyoshi Sakamoto[2,3]
Tomah Sogabe[2,3,1,b)]

Abstract:
Previously we established a VC dimension upper bound for "encoding-first" quantum circuits, where the input layer is the first layer of the circuit. In this report, we prove a general VC dimension upper bound for quantum circuit learning including "data re-uploading" circuits, where the input gates can be single qubit rotations anywhere in the circuit. We discuss the properties of the bound and some other considerations.

Keywords: Machine learning, quantum circuit, VC dimension

1. Introduction

Due to the difficulty of simulating quantum systems using classical computers, building computing machines using quantum mechanics is suggested as a way toward computational advantages [1], [2], [3], [4], [5], [6]. The computational capability of current Noisy Intermediate-Scale Quantum (NISQ) [7] hardware was experimentally demonstrated [8]. On the other hand, classical machine learning [9], [10], [11], [12] for Artificial Intelligence (AI) shows a wide range of applications [13], [14]. It is reasonable to consider NISQ devices for AI applications [15].

Using variational quantum circuits [16], [17], [18] as prediction models in supervised learning leads to the quantum circuit learning (QCL) method [18], [19], [20], [21]. In this setting, the learning task is similar to classical setting such that the training data set and predictions are restricted to classical data. Only the hypothesis set is constructed using variational quantum circuits. Theoretical efforts toward understanding the expressive power of QCL is conducted by many groups [22], [23], [24], [25], [26].

One important question in supervised learning is the learnability of the hypothesis set being used. If the size of training data set is small but the model complexity is high, a learning machine could overfit to the data noise and hence fail to generalize well for future predictions. Uniform non-asymptotic theory of generalization for supervised machine learning started with Vapnik–Chervonenkis (VC) theory [27] and is generally known as statistical learning theory [28], [29], [30], [31], [32]. Probably Approximately Correct (PAC) framework proposed by Valiant [33] also includes computational requirements in its original form. For binary classification tasks, VC theory can be used to establish the generalization ability by using the VC dimension of the model class [34].

Previous learnability results for quantum machine learning are based on fat-shattering dimension [35], pseudo-dimension [36], or quantum sample complexity [37]. Another VC-dimension upper bound, which is different from our result, is proposed in [38]. Many other recent learnability results based on various measures and settings could be found in literatures [39], [40], [41], [42], [43], [44], [45].

The limitation of expressibility of "encoding-first" quantum circuit was observed by many groups [24], [44], [46], and the "data re-uploading" circuit [46] was proposed to resolve the limitation. The learnability of data re-uploading QCL is shown in [44] by using Rademacher complexity. Our previous study [47] shows that the growth of VC dimension saturates for deep QCL. This is different from classical deep neural networks (number of edges=$|E|$, number of vertices=$|V|$), where the VC dimension grows asymptotically as $O(|E|\log(|E|))$ (for sign activation function) or $O(|V|^2|E|^2)$ (for sigmoid activation function) [29], [31], [48]. In this work, we extend our previous [47] result of VC dimension upper bound to include the data re-uploading scheme.

This report is organized as follows. Section 2 provides brief explanations for quantum circuit learning method and statistical learning theory. Section 3 contains the main result and its proof. Further discussions about the results are presented in Section 4.

2. Preliminaries

Quantum circuit learning and statistical learning theory are introduced in this section.

[1] Grid Inc., 107-0061 Tokyo, Japan
[2] Engineering department, The University of Electro-Communications, 182-8585 Tokyo, Japan
[3] i-PERC, The University of Electro-Communications, 182-8585 Tokyo, Japan
[a)] chen.chih.chieh@gridsolar.jp
[b)] sogabe@uec.ac.jp

2.1 Quantum circuit learning

For a supervised binary classification learning problem, we are given some classical training data set $\{(\vec{x}_i, y_i) : \vec{x}_i \in X, y_i \in Y = \{-1, 1\}, i \in \{1, ..., N\}\}$ drawn from some unknown joint probability distribution $(\vec{x}_i, y_i) \sim P(\vec{x}, y)$ over $X \times Y$. The goal of learning is to obtain a model $h : X \mapsto Y$ such that the prediction error (out-of-sample error) $E_{out} = \mathbb{P}_{(\vec{x},y)\sim P(\vec{x},y)}[h(\vec{x}) \neq y]$ is small.

The QCL considered in this work uses some quantum circuits to construct the hypothesis set H. For some d-dimensional input vector $\vec{x} = (x_0, ..., x_{d-1}) \in [-1, 1]^d = X$, some encoding maps $\vec{\phi} = (\phi_0(x_0), ..., \phi_{d-1}(x_{d-1})) : [-1, 1]^d \mapsto [-\pi, \pi]^d$ and some real variational parameters θ, the circuit gives an unitary evolution $U_\theta(\vec{\phi}(\vec{x}))$ acting on all-zero initial state $|0\rangle^{\otimes n}$. n denotes the number of qubits (circuit width). We do not assume any special structure for variational parameters and entanglers, while the encoding method is specified as follows. For one input vector $\vec{x} = (x_0, ..., x_{d-1}) \in [-1, 1]^d = X$, each dimension $x_i \in [-1, 1]$ is encoded by one encoding mapping $\phi_i : [-1, 1] \mapsto [-\pi, \pi]$ with one single qubit rotation $R_s \in \{R_Y, R_Z\}$. The gate $R_s(\phi_i(x_i))$ is applied to the quantum circuit to upload the data. Data re-uploading means that the gate $R_s(\phi_i(x_i))$ is applied to the circuit several times for an $i \in \{0, ..., d-1\}$. The number n_i denotes the total number of $R_s(\phi_i(x_i))$ gates being applied for an $i \in \{0, ..., d-1\}$. The measurement result is used to compute the expectation value for some fixed observable O.

$$\langle O(\theta, \vec{\phi}(\vec{x}))\rangle = \mathrm{Tr}[OU_\theta(\vec{\phi}(\vec{x}))|0\rangle^{\otimes n}\langle 0|^{\otimes n}U_\theta^\dagger(\vec{\phi}(\vec{x}))]. \tag{1}$$

The expectation value is then thresholded to construct a hypothesis set $H = \{\mathrm{sgn}(f_\theta(\vec{\phi}(\vec{x})) + c) : f_\theta(\vec{\phi}(\vec{x})) = \langle O(\theta, \vec{\phi}(\vec{x}))\rangle = \mathrm{Tr}[OU_\theta(\vec{\phi}(\vec{x}))|0\rangle^{\otimes n}\langle 0|^{\otimes n}U_\theta^\dagger(\vec{\phi}(\vec{x}))], c \in \mathbb{R}\}$ for binary classification.

2.2 Statistical learning theory

VC theory provides a general theory of generalization ability for binary classification tasks. We use the definition that the generalization error is $E_{out} - E_{in}$, where E_{out} is the out-of-sample error (prediction error) and $E_{in} = \frac{1}{N}\sum_{i=1}^N [\![h(\vec{x}_i) \neq f(\vec{x}_i)]\!]$ is the in-sample error. The VC generalization error bound is [27]

$$\mathbb{P}[\sup_{h\in H} |E_{out}(h) - E_{in}(h)| > \epsilon] \leq 4m_H(2N)e^{(-\frac{1}{8}\epsilon^2 N)}, \tag{2}$$

where the randomness is over i.i.d. samples $\{(\vec{x}_i, y_i) \sim P(\vec{x}, y)\ \forall i \in \{1, ..., N\}\}$. N is the sample size. The function $m_H(N) = \max_{\vec{x}_1,...,\vec{x}_N\in X} |\{(h(\vec{x}_1), ..., h(\vec{x}_N)) : h \in H\}|$ could be upper bounded by $m_H(N) \leq \sum_{i=0}^{d_{vc}} \binom{N}{i} \leq N^{d_{VC}} + 1$ for finite VC-dimension $d_{VC} = \max_{N\in\mathbb{N}}\{N : m_H(N) = 2^N\}$. VC dimension is the maximum number of points that can be shattered by the hypothesis set. In general, d_{VC} could be infinite for an uncountable hypothesis set. If d_{VC} is finite, then the generalization ability of the learning machine is guaranteed by the VC bound and the hypothesis set is called "PAC-learnable." Several features of VC theory are worth noting [28]: (1) VC bound is independent of the input distribution. (2) VC bound is non-asymptotic, so it can be applied when the size of training data set is small. (3) VC bound is uniform over the hypothesis set, which means that it is true for all the models in the set.

After VC theory, there are latter developments for the generalization ability of learning machines. For real-valued functions, the pseudo-dimension [49] and the fat-shattering dimension [50], [51] could be used for generalization bounds. VC theory is also extended to real-valued functions [28]. PAC-Bayesian bounds are proposed for Bayesian setting [52], [53], [54]. There are also other generalization bounds which are not VC bound but use VC dimension as a measure [55]. Some introductions and comparative study of these measures could be found in references [29], [31], [32], [55].

3. Main result

The main result is presented here. The proof is an extension of the proof in [47].

Theorem 1 (VC dimension upper bound for quantum circuits)**.** *Assume the input vector $\vec{x} = (x_0, ..., x_{d-1}) \in [-1, 1]^d$. Each dimension $x_i \in [-1, 1]$ is uploaded n_i times using single qubit encoding rotations $R_s(\phi_i(x_i))$ for some fixed encoding mapping $\phi_i : [-1, 1] \mapsto [-\pi, \pi]$ with $s \in \{Y, Z\}$. Then the VC dimension of the hypothesis set $H = \{sgn(f_\theta(\vec{\phi}(\vec{x})) + c) : f_\theta(\vec{\phi}(\vec{x})) = \langle O(\theta, \vec{\phi}(\vec{x}))\rangle = Tr[OU_\theta(\vec{\phi}(\vec{x}))|0\rangle^{\otimes n}\langle 0|^{\otimes n}U_\theta^\dagger(\vec{\phi}(\vec{x}))], c \in \mathbb{R}\}$ for a fixed observable O is upper bounded by*

$$d_{VC} \leq \prod_{i=0}^{d-1}(2n_i + 1). \tag{3}$$

Proof. We claim that $f_\theta(\vec{\phi}(\vec{x}))$ is a real trigonometric polynomial of d variables, and the degree of the polynomial for each variable is at most n_i. Then the theorem is proved by Dudley's theorem for VC dimension of thresholded real vector space function classes [29], [31], [56], [57].

The proof of the claim is as follows. The initial density matrix $\rho_0 = |0\rangle^{\otimes n}\langle 0|^{\otimes n}$ has constant matrix elements. From the assumptions, all the variational unitaries and entanglers do not depend on input vector $\vec{x}$. Consider an input dimension $x_i \in [-1, 1]$ and encoding mapping $\phi_i(x_i) \in [-\pi, \pi]$ where $i \in \{0, ..., d-1\}$. If this dimension is uploaded by R_Y

$$R_Y(\phi_i) = \begin{pmatrix} \cos(\frac{\phi_i}{2}) & -\sin(\frac{\phi_i}{2}) \\ \sin(\frac{\phi_i}{2}) & \cos(\frac{\phi_i}{2}) \end{pmatrix} \tag{4}$$

$$= \cos(\frac{\phi_i}{2})\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin(\frac{\phi_i}{2})\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{5}$$

then the action of this gate on k-th qubit of n-qubit Hilbert space is

$$R_Y(\phi_i)|_k = \cos(\frac{\phi_i}{2})\mathbb{I}_{2^n} + \sin(\frac{\phi_i}{2})\mathbb{I}_{2^k} \otimes \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \otimes \mathbb{I}_{2^{n-k-1}} \tag{6}$$

$$= \cos(\frac{\phi_i}{2})\mathbb{I}_{2^n} + \sin(\frac{\phi_i}{2})\mathbb{A} \tag{7}$$

where $\mathbb{I}_M$ denotes $M \times M$ identity matrix and $\mathbb{A}$ is some constant matrix. The action of this gate on a density matrix ρ is then

$$R_Y(\phi_i)|_k \rho R_Y(\phi_i)|_k^\dagger \tag{8}$$

$$= (\cos(\frac{\phi_i}{2})\mathbb{I}_{2^n} + \sin(\frac{\phi_i}{2})\mathbb{A})\rho(\cos(\frac{\phi_i}{2})\mathbb{I}_{2^n} + \sin(\frac{\phi_i}{2})\mathbb{A}^\dagger) \tag{9}$$

$$= \frac{1}{2}[(1 + \cos(\phi_i))\rho + \sin(\phi_i)(\mathbb{A}\rho + \rho\mathbb{A}^\dagger) \tag{10}$$

$$+(1 - \cos(\phi_i))\mathbb{A}\rho\mathbb{A}^\dagger]. \tag{11}$$

If the matrix elements of ρ are trigonometric polynomials of $\vec{\phi}$, then the matrix elements of the updated density matrix $R_Y(\phi_i)|_k \rho R_Y(\phi_i)|_k^\dagger$ are trigonometric polynomials where the degree for the variable ϕ_i is increased by at most one. Similar argument works if the dimension is uploaded by $R_Z(\phi_i)$. Hence, $f_\theta(\vec{\phi}(\vec{x}))$ is a trigonometric polynomial with the claimed degree upper bound. Let $f_\theta(\vec{\phi}(\vec{x})) = \sum_k a_k(\theta) f_k(\vec{\phi}(\vec{x}))$, where $\{f_k(\vec{\phi})\}$ is the real trigonometric polynomial basis and $\{a_k(\theta)\}$ are the Fourier coefficients. Since $f_\theta(\vec{\phi}(\vec{x}))$ is a real-valued function, the coefficients $a_k(\theta) = \langle f_k | f_\theta \rangle \in \mathbb{R}\ \forall k$. The claim is proved.

□

4. Discussions

In this section, we provide some short discussions regarding the obtained theorem.

4.1 Applicability of the bound

There is no requirement on the structure of variational (trainable) gates and entangling gates of the circuit, except that they should not contain any input data x_i. There is no requirement on the encoding gates $R_s(\phi_i(x_i))$, except that they should not contain any variational parameter.

Notice that in practice, one usually applies some classical post processing techniques to the output expectation values [20]. The VC dimension bound should be adjusted accordingly.

We provide some extensions.

Corollary 1 (Linear combinations of expectations)**.** *If the hypothesis set is the real linear combination of several observables for a fixed circuit such that* $H = \{sgn(f_\theta(\vec{\phi}(\vec{x})) + c_0) : f_\theta(\vec{\phi}(\vec{x})) = \sum_i c_i \langle O_i(\theta, \vec{\phi}(\vec{x})) \rangle = \sum_i c_i Tr[O_i U_\theta(\vec{\phi}(\vec{x}))|0\rangle^{\otimes n}\langle 0|^{\otimes n} U_\theta^\dagger(\vec{\phi}(\vec{x}))], c_i \in \mathbb{R}\}$*, then the bound in Theorem 1 is still true.*

Corollary 2 (Mixed state learnability)**.** *If the initial state is some mixed state* ρ *which does not depend on the input vector* $\vec{x}$ *such that* $H_{QCL} = \{sgn(f_\theta(\vec{\phi}(\vec{x})) + c_0) : f_\theta(\vec{\phi}(\vec{x})) = \sum_i c_i \langle O_i(\theta, \vec{\phi}(\vec{x})) \rangle = \sum_i c_i Tr[O_i U_\theta(\vec{\phi}(\vec{x}))\rho U_\theta^\dagger(\vec{\phi}(\vec{x}))], c_i \in \mathbb{R}\}$*, then the bound in Theorem 1 is still true.*

4.2 Reduction to the previous results

We show how to obtain the special case in our previous work [47] for the ansatz in [19].

$$d_{VC} \leq (2\frac{n}{d} + 1)^{2d}. \tag{12}$$

This bound can be obtained from the general bound in Theorem 1 as follows. The encoding used in [19] can be understood as performing feature maps $x_i \mapsto x_i^2$ to increase the feature dimension from d to $2d$. The encoding maps $\phi_i(x_i) = \arcsin(x_i)$ and $\phi_i'(x_i^2) = \arccos(x_i^2)$ are used, and are uploaded by $R_Y(\phi_i(x_i)) = R_Y(\arcsin(x_i))$ and $R_Z(\phi_i'(x_i^2)) = R_Z(\arccos(x_i^2))$. Each dimension is uploaded $n_i = \frac{n}{d}$ times, and hence we get the bound $(2n_i + 1)^{2d} = (2\frac{n}{d} + 1)^{2d}$. The lightcone bound can be calculated by counting n_i covered by the lightcone for a specific ansatz.

4.3 Looseness of the bound

Notice that our bound is based on counting the number of basis functions, hence the bound does not depend on the number of variational parameters. This suggests that the bound can not be tight in general. For example, if the number of variational parameter is zero, then the VC dimension is zero. Ideally, we also want a scaling with respect to the number of variational gates like the cases in [41], [45].

4.4 Approximation-estimation trade-off considerations

To achieve low prediction error in supervised learning, the approximation-estimation trade-off (also known as bias-variance trade-off) should be considered [31], [32]. The generalization error bound discussed in this work is only for estimation error.

Barron [58] gives the approximation error bound for single layer classical neural network hypothesis set $H_{NN} = \{f(\vec{x}) = \sum_{k=1}^n c_k \phi(\vec{a}_k \cdot \vec{x} + b_k) + c_0 : \vec{a}_k \in \mathbb{R}^d, b_k, c_k \in \mathbb{R}\}$ where ϕ is a sigmoid function and n is the number of nodes. Barron also analyzed the approximation-estimation trade-off of neural networks [59]. It is shown that neural networks have approximation advantage over linear combinations of fixed basis functions in the sense that the approximation has faster convergence rate for high-dimensional inputs.

One attempt to overcome the limitation of fixed basis functions of QCL was actually proposed in [46]: combining neural networks with QCL to construct, for example, the hypothesis set $H_{affineQCL} = \{(f_\theta(\vec{\phi}(\vec{x})) + c_0) : f_\theta(\vec{\phi}(\vec{x})) = \sum_i c_i \langle O_i(\theta, \vec{\phi}(W \cdot \vec{x} + \vec{b})) \rangle = \sum_i c_i \mathrm{Tr}[O_i U_\theta(\vec{\phi}(W \cdot \vec{x} + \vec{b}))\rho U_\theta^\dagger(\vec{\phi}(W \cdot \vec{x} + \vec{b}))], c_i \in \mathbb{R}, W \in \mathbb{R}^{d \times d}, \vec{b} \in \mathbb{R}^d\}$, where the affine transformation $W \cdot \vec{x} + \vec{b}$ is composited with QCL. However, a simple special case $\{\sin(Wx) : W \in \mathbb{R}\}$ has infinite VC dimension, and hence is not PAC-learnable [28], [29], [47]. This is because W provides possibly high-frequency oscillations to shatter arbitrarily many data points. One possible way to resolve this problem could be using a sigmoid activation function ϕ for encoding. For example, uploading the input x_i with $R_s(\pi\phi(W_i x_i + b_i))$ gate. This could be a future direction.

5. Conclusion

In this work, we give a general VC dimension upper bound for quantum circuit learning, and hence establish the PAC learnability of this hypothesis set. While this result provides a basis for quantum circuit supervised learning, many questions remain. For example, we did not address the issues of hardware error and sampling error of quantum machines (due to finite readout samples), which could effect the generalization ability. We did not have a bound which scales with respect to the number of trainable parameters. The approximation-estimation trade-off should also be addressed. These questions are left for future investigations.

Acknowledgments We thank Naoki Yamamoto for valuable discussions. We thank Matthias C. Caro for providing many useful references.

References

[1] Feynman, R. P.: Simulating Physics with Computers, *International Journal of Theoretical Physics*, Vol. 21, No. 6-7, pp. 467–488 (online), DOI: 10.1007/BF02650179 (1982).

[2] Deutsch, D.: Quantum theory, the Church-Turing principle and the universal quantum computer, *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, Vol. 400, No. 1818, pp.

97–117 (online), DOI: 10.1098/rspa.1985.0070 (1985).

[3] Barenco, A., Bennett, C. H., Cleve, R., DiVincenzo, D. P., Margolus, N., Shor, P., Sleator, T., Smolin, J. A. and Weinfurter, H.: Elementary gates for quantum computation, *Phys. Rev. A*, Vol. 52, pp. 3457–3467 (online), DOI: 10.1103/PhysRevA.52.3457 (1995).

[4] Preskill, J.: Lecture notes for physics 229: Quantum information and computation (1998).

[5] Mermin, N. D.: *Quantum Computer Science: An Introduction*, Cambridge University Press (2007).

[6] Nielsen, M. A. and Chuang, I. L.: *Quantum Computation and Quantum Information: 10th Anniversary Edition*, Cambridge University Press (2010).

[7] Preskill, J.: Quantum Computing in the NISQ era and beyond, *Quantum*, Vol. 2, p. 79 (online), DOI: 10.22331/q-2018-08-06-79 (2018).

[8] Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., Biswas, R., Boixo, S., Brandao, F. G. S. L., Buell, D. A., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Collins, R., Courtney, W., Dunsworth, A., Farhi, E., Foxen, B., Fowler, A., Gidney, C., Giustina, M., Graff, R., Guerin, K., Habegger, S., Harrigan, M. P., Hartmann, M. J., Ho, A., Hoffmann, M., Huang, T., Humble, T. S., Isakov, S. V., Jeffrey, E., Jiang, Z., Kafri, D., Kechedzhi, K., Kelly, J., Klimov, P. V., Knysh, S., Korotkov, A., Kostritsa, F., Landhuis, D., Lindmark, M., Lucero, E., Lyakh, D., Mandrà, S., McClean, J. R., McEwen, M., Megrant, A., Mi, X., Michielsen, K., Mohseni, M., Mutus, J., Naaman, O., Neeley, M., Neill, C., Niu, M. Y., Ostby, E., Petukhov, A., Platt, J. C., Quintana, C., Rieffel, E. G., Roushan, P., Rubin, N. C., Sank, D., Satzinger, K. J., Smelyanskiy, V., Sung, K. J., Trevithick, M. D., Vainsencher, A., Villalonga, B., White, T., Yao, Z. J., Yeh, P., Zalcman, A., Neven, H. and Martinis, J. M.: Quantum supremacy using a programmable superconducting processor, *Nature*, Vol. 574, No. 7779, pp. 505–510 (online), DOI: 10.1038/s41586-019-1666-5 (2019).

[9] Bishop, C. M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg (2006).

[10] Goodfellow, I., Bengio, Y. and Courville, A.: *Deep Learning*, The MIT Press (2016).

[11] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media (2009).

[12] Bengio, Y., Lecun, Y. and Hinton, G.: Deep Learning for AI, *Commun. ACM*, Vol. 64, No. 7, p. 58–65 (online), DOI: 10.1145/3448250 (2021).

[13] Russell, S. and Norvig, P.: *Artificial Intelligence: A Modern Approach*, Prentice Hall Press, USA, 4rd edition (2020).

[14] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D.: Mastering the game of Go without human knowledge, *Nature*, Vol. 550, No. 7676, pp. 354–359 (online), DOI: 10.1038/nature24270 (2017).

[15] Dunjko, V. and Briegel, H. J.: Machine learning & artificial intelligence in the quantum domain: a review of recent progress, *Reports on Progress in Physics*, Vol. 81, No. 7, p. 074001 (online), DOI: 10.1088/1361-6633/aab406 (2018).

[16] Peruzzo, A., McClean, J., Shadbolt, P., Yung, M.-H., Zhou, X.-Q., Love, P. J., Aspuru-Guzik, A. and O'Brien, J. L.: A variational eigenvalue solver on a photonic quantum processor, *Nature Communications*, Vol. 5, p. 4213 (online), DOI: 10.1038/ncomms5213 (2014).

[17] McClean, J. R., Romero, J., Babbush, R. and Aspuru-Guzik, A.: The theory of variational hybrid quantum-classical algorithms, *New Journal of Physics*, Vol. 18, No. 2, p. 023023 (online), DOI: 10.1088/1367-2630/18/2/023023 (2016).

[18] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., McClean, J. R., Mitarai, K., Yuan, X., Cincio, L. and et al.: Variational quantum algorithms, *Nature Reviews Physics*, Vol. 3, No. 9, p. 625–644 (online), DOI: 10.1038/s42254-021-00348-9 (2021).

[19] Mitarai, K., Negoro, M., Kitagawa, M. and Fujii, K.: Quantum circuit learning, *Phys. Rev. A*, Vol. 98, p. 032309 (online), DOI: 10.1103/PhysRevA.98.032309 (2018).

[20] Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M. and Gambetta, J. M.: Supervised learning with quantum-enhanced feature spaces, *Nature*, Vol. 567, No. 7747, pp. 209–212 (online), DOI: 10.1038/s41586-019-0980-2 (2019).

[21] Schuld, M., Bergholm, V., Gogolin, C., Izaac, J. and Killoran, N.: Evaluating analytic gradients on quantum hardware, *Phys. Rev. A*, Vol. 99, p. 032331 (online), DOI: 10.1103/PhysRevA.99.032331 (2019).

[22] Sim, S., Johnson, P. D. and Aspuru-Guzik, A.: Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms, *Advanced Quantum Technologies*, Vol. 2, No. 12, p. 1900070 (online), DOI: https://doi.org/10.1002/qute.201900070 (2019).

[23] Nakaji, K. and Yamamoto, N.: Expressibility of the alternating layered ansatz for quantum computation, *Quantum*, Vol. 5, p. 434 (online), DOI: 10.22331/q-2021-04-19-434 (2021).

[24] Schuld, M., Sweke, R. and Meyer, J. J.: Effect of data encoding on the expressive power of variational quantum-machine-learning models, *Phys. Rev. A*, Vol. 103, p. 032430 (online), DOI: 10.1103/PhysRevA.103.032430 (2021).

[25] Hubregtsen, T., Pichlmeier, J., Stecher, P. and Bertels, K.: Evaluation of Parameterized Quantum Circuits: on the relation between classification accuracy, expressibility and entangling capability, *arXiv e-prints*, p. arXiv:2003.09887 (2020).

[26] Goto, T., Tran, Q. H. and Nakajima, K.: Universal Approximation Property of Quantum Machine Learning Models in Quantum-Enhanced Feature Spaces, *Phys. Rev. Lett.*, Vol. 127, p. 090506 (online), DOI: 10.1103/PhysRevLett.127.090506 (2021).

[27] Vapnik, V. N. and Chervonenkis, A. Y.: On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, *Theory of Probability & Its Applications*, Vol. 16, No. 2, pp. 264–280 (online), DOI: 10.1137/1116025 (1971).

[28] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., 2nd edition (1999).

[29] Anthony, M. and Bartlett, P. L.: *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, USA, 1st edition (2009).

[30] Abu-Mostafa, Y. S., Magdon-Ismail, M. and Lin, H.-T.: *Learning From Data*, AMLBook (2012).

[31] Shalev-Shwartz, S. and Ben-David, S.: *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, USA (2014).

[32] Mohri, M., Rostamizadeh, A. and Talwalkar, A.: *Foundations of Machine Learning*, The MIT Press (2018).

[33] Valiant, L. G.: A Theory of the Learnable, *Commun. ACM*, Vol. 27, No. 11, p. 1134–1142 (online), DOI: 10.1145/1968.1972 (1984).

[34] Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K.: Learnability and the Vapnik-Chervonenkis Dimension, *J. ACM*, Vol. 36, No. 4, p. 929–965 (online), DOI: 10.1145/76359.76371 (1989).

[35] Aaronson, S.: The learnability of quantum states, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 463, No. 2088, pp. 3089–3114 (online), DOI: 10.1098/rspa.2007.0113 (2007).

[36] Caro, M. C. and Datta, I.: Pseudo-dimension of quantum circuits, *Quantum Machine Intelligence*, Vol. 2, No. 2 (online), DOI: 10.1007/s42484-020-00027-5 (2020).

[37] Arunachalam, S. and de Wolf, R.: Optimal Quantum Sample Complexity of Learning Algorithms, *Journal of Machine Learning Research*, Vol. 19, No. 71, pp. 1–36 (online), available from ⟨http://jmlr.org/papers/v19/18-195.html⟩ (2018).

[38] Gyurik, C., van Vreumingen, D. and Dunjko, V.: Structural risk minimization for quantum linear classifiers, *arXiv e-prints*, p. arXiv:2105.05566 (2021).

[39] Huang, H.-Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H. and McClean, J. R.: Power of data in quantum machine learning, *Nature Communications*, Vol. 12, p. 2631 (online), DOI: 10.1038/s41467-021-22539-9 (2021).

[40] Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A. and Woerner, S.: The power of quantum neural networks, *Nature Computational Science*, Vol. 1, No. 6, p. 403–409 (online), DOI: 10.1038/s43588-021-00084-1 (2021).

[41] Du, Y., Tu, Z., Yuan, X. and Tao, D.: An efficient measure for the expressivity of variational quantum algorithms, *arXiv e-prints*, p. arXiv:2104.09961 (2021).

[42] Du, Y., Hsieh, M.-H., Liu, T., You, S. and Tao, D.: Learnability of Quantum Neural Networks, *PRX Quantum*, Vol. 2, p. 040337 (online), DOI: 10.1103/PRXQuantum.2.040337 (2021).

[43] Bu, K., Enshan Koh, D., Li, L., Luo, Q. and Zhang, Y.: On the statistical complexity of quantum circuits, *arXiv e-prints*, p. arXiv:2101.06154 (2021).

[44] Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J. and Sweke, R.: Encoding-dependent generalization bounds for parametrized quantum circuits, *Quantum*, Vol. 5, p. 582 (online), DOI: 10.22331/q-2021-11-17-582 (2021).

[45] Caro, M. C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L. and Coles, P. J.: Generalization in quantum machine learning from few training data, *arXiv e-prints*, p. arXiv:2111.05292 (2021).

[46] Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. and Latorre, J. I.: Data re-uploading for a universal quantum classifier, *Quantum*, Vol. 4, p. 226 (online), DOI: 10.22331/q-2020-02-06-226 (2020).

[47] Chen, C.-C., Watabe, M., Shiba, K., Sogabe, M., Sakamoto, K. and Sogabe, T.: On the Expressibility and Overfitting of Quantum Circuit Learning, *ACM Transactions on Quantum Computing*, Vol. 2, No. 2 (online), DOI: 10.1145/3466797 (2021).
[48] Kearns, M. J. and Vazirani, U. V.: *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, USA (1994).
[49] Pollard, D.: *Convergence of Stochastic Processes*, Springer (1984).
[50] Kearns, M. J. and Schapire, R. E.: Efficient distribution-free learning of probabilistic concepts, *Journal of Computer and System Sciences*, Vol. 48, No. 3, pp. 464–497 (online), DOI: https://doi.org/10.1016/S0022-0000(05)80062-5 (1994).
[51] Bartlett, P. L., Long, P. M. and Williamson, R. C.: Fat-Shattering and the Learnability of Real-Valued Functions, *Journal of Computer and System Sciences*, Vol. 52, No. 3, pp. 434–452 (online), DOI: https://doi.org/10.1006/jcss.1996.0033 (1996).
[52] McAllester, D. A.: Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, New York, NY, USA, Association for Computing Machinery, p. 230–234 (online), DOI: 10.1145/279943.279989 (1998).
[53] McAllester, D. A.: Some PAC-Bayesian Theorems, *Mach. Learn.*, Vol. 37, No. 3, p. 355–363 (online), DOI: 10.1023/A:1007618624809 (1999).
[54] McAllester, D. A.: PAC-Bayesian Stochastic Model Selection, *Mach. Learn.*, Vol. 51, No. 1, p. 5–21 (online), DOI: 10.1023/A:1021840411064 (2003).
[55] Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. and Bengio, S.: Fantastic Generalization Measures and Where to Find Them, *arXiv e-prints*, p. arXiv:1912.02178 (2019).
[56] Dudley, R. M.: Central Limit Theorems for Empirical Measures, *The Annals of Probability*, Vol. 6, No. 6, pp. 899 – 929 (online), DOI: 10.1214/aop/1176995384 (1978).
[57] Sontag, E.: VC dimension of neural networks, *NATO ASI series. Series F: computer and system sciences*, pp. 69–95 (1998).
[58] Barron, A.: Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Transactions on Information Theory*, Vol. 39, No. 3, pp. 930–945 (online), DOI: 10.1109/18.256500 (1993).
[59] Barron, A. R.: Approximation and Estimation Bounds for Artificial Neural Networks, *Mach. Learn.*, Vol. 14, No. 1, p. 115–133 (online), DOI: 10.1023/A:1022650905902 (1994).

