Universal Approximation Property of Quantum Machine Learning Models in Quantum-Enhanced Feature Spaces

Takahiro Goto,[1,†] Quoc Hoan Tran,[1,2,*] and Kohei Nakajima[1,2,3,‡]
[1]Reservoir Computing Seminar Group, Nagase Hongo Building F8, 5-24-5, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan
[3]Next Generation Artificial Intelligence Research Center, The University of Tokyo, Tokyo 113-8656, Japan

(Received 4 January 2021; accepted 21 July 2021; published 27 August 2021)

Encoding classical data into quantum states is considered a quantum feature map to map classical data into a quantum Hilbert space. This feature map provides opportunities to incorporate quantum advantages into machine learning algorithms to be performed on near-term intermediate-scale quantum computers. The crucial idea is using the quantum Hilbert space as a quantum-enhanced feature space in machine learning models. Although the quantum feature map has demonstrated its capability when combined with linear classification models in some specific applications, its expressive power from the theoretical perspective remains unknown. We prove that the machine learning models induced from the quantum-enhanced feature space are universal approximators of continuous functions under typical quantum feature maps. We also study the capability of quantum feature maps in the classification of disjoint regions. Our work enables an important theoretical analysis to ensure that machine learning algorithms based on quantum feature maps can handle a broad class of machine learning tasks. In light of this, one can design a quantum machine learning model with more powerful expressivity.

DOI: 10.1103/PhysRevLett.127.090506

The rapidly increasing volume and complexity of data have led to the notable progress of machine learning (ML) techniques to build sophisticated models to find patterns in data. The main interest lies in the ability to recognize the patterns these techniques can produce. If a physical computation model can produce atypical patterns that cannot be generated by a classical computer, it may reveal patterns that are difficult to recognize in the classical regime [1]. This expectation has led to the advent of quantum machine learning (QML): a field that takes advantage of quantum effects to surpass the classical ML techniques. QML is currently benefiting from the arrival of noisy intermediate-scale quantum devices that may include a few tens to hundreds of qubits with no error correction capability [2,3]. Such devices have ushered in the era of hybrid quantum-classical algorithms [4–9].

Because a quantum computer can efficiently access and manipulate quantum states, the quantum Hilbert space can be used as a quantum-enhanced feature space for classical data. The motivation is that quantum systems can explore a larger class of features than can classical systems. The input data are encoded in a quantum state via a quantum feature map: a nonlinear feature map that maps data to the quantum Hilbert space (Fig. 1). The quantum feature map is first proposed and implemented as a fixed quantum circuit, followed by a variational circuit that adapts the measurement basis with trainable parameters [8,9]. Such QML models can be rephrased as quantum kernel methods induced from feature maps [10–14]. Quantum feature maps underscore the QML advantage; there may be a provable exponential speedup due to the classical intractability of generating correlations for a particular learning problem. For example, under the widely known hardness assumption of the discrete logarithm problem, the first probable exponential QML advantage was demonstrated via the estimation of a support vector machine kernel matrix on a fault-tolerant quantum computer [15]. Furthermore, one can construct engineered datasets to demonstrate the most significant separation between quantum and classical models from a learning-theoretic sense to yield the quantum

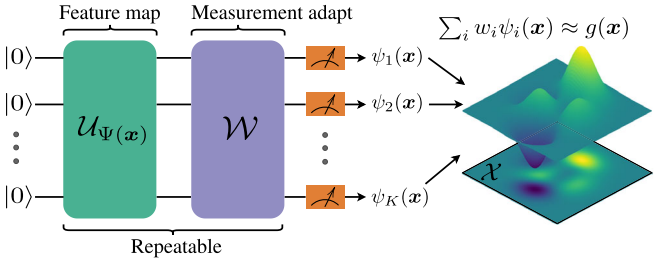

FIG. 1. Quantum feature framework consisting of a feature map circuit $\mathcal{U}_{\Psi(\boldsymbol{x})}$ realizing $\Psi(\boldsymbol{x})$ to map classical data $\boldsymbol{x} \in \mathcal{X}$ to a quantum state in Hilbert space and quantum circuit $\mathcal{W}$ to adapt measurement basis. Combination of $\mathcal{U}_{\Psi(\boldsymbol{x})}$ and $\mathcal{W}$ can be repeated as a sequence with different parameters. This framework has the universal approximation property if the linear combining of measurement results can approximate any continuous function $g: \mathcal{X} \to \mathbb{R}$.

advantage in ML problems [16]. Still, little is known about the relation between the classical intractability of quantum feature maps and the generalization learning performance.

An interesting research question is whether a QML model based on a quantum feature map can obtain expressivity that is as powerful as, or is more powerful than, classical ML schemes. The answer can determine whether QML models can handle a broad class of ML tasks in general. This can be investigated from the perspective of the universal approximation property (UAP) and the classification capability, which have been extensively explored in feedforward classical neural networks [17–19]. Here, the UAP refers to the ability to approximate any continuous function [20,21]. The classification capability implies that the function constructed from quantum feature maps can form disjoint decision regions [22]. Quantum neural networks, which employ qubits as quantum perceptrons with nonlinear excitation responses [23], can be emulated on a photonic quantum computer to obtain the UAP [24]. It is conjectured that under a special kind of classical data preprocessing, sequentially repeated quantum feature maps can become universal function approximators [25]. In Ref. [26], the expressivity of a quantum model with a variational circuit is characterized in terms of a partial Fourier series in the data. However, the study of the UAP and classification capability of QML models with quantum feature maps still remains challenging.

In this Letter, we formulate the universal approximation problem of QML models in terms of quantum feature maps. We present a provable UAP and classification capability in two typical scenarios when setting the quantum feature map. In the first scenario, which is defined as the *parallel scenario*, the quantum feature map is a tensor product of multiple quantum circuits; each circuit acts on a subsystem, and the number of qubits can be set freely. In the second scenario, which is defined as the *sequential scenario*, the quantum feature map is the repetition of a simple fixed quantum circuit, and the number of qubits is fixed. We obtain the UAP in the first scenario and prove the UAP for the second in single-qubit circuits of the finite input space. Both scenarios have been mentioned in prior proposals via short circuit sequences in realistic near-term settings [8,26,27]. We therefore focus on the extent to which these abstract setups can influence the approximating power of QML models in future implementations with wider and deeper quantum circuits.

Quantum feature maps.—We will now define the quantum feature map mentioned in Refs. [8,9]. Let $\mathcal{H}$ be a Hilbert space and $\mathcal{X} \subset \mathbb{R}^d$ be an input set. The quantum feature map $\Psi: \mathcal{X} \to \mathcal{H}$ is a procedure of input encoding that encodes some input $\boldsymbol{x} \in \mathcal{X}$ into a quantum feature state $|\Psi(\boldsymbol{x})\rangle \in \mathcal{H}$. This mapping action is equivalent to applying the quantum circuit $\mathcal{V}(\boldsymbol{x}) = \mathcal{U}_{\Psi(\boldsymbol{x})}$ to the initial state $|0\rangle^{\otimes N}$, where N is the number of qubits. A quantum classifier can be constructed from the quantum feature map using two approaches: the variational circuit approach, and the kernel-induced approach. In the variational circuit approach, a short-depth quantum circuit $\mathcal{W}$ is applied to the quantum feature state to adapt the measurement basis [8,9] (Fig. 1). The parameters of circuit $\mathcal{W}$ are optimized during the training, and the quantum measurement is performed to obtain a complex nonlinear output. This output can be represented as a linear combination of exponentially many nonlinear functions. In the kernel-induced approach, the quantum computer estimates the inner product between quantum feature states, giving rise to a kernel

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \Psi(\boldsymbol{x}) | \Psi(\boldsymbol{x}') \rangle = \langle 0 \ldots 0 | \mathcal{V}^\dagger(\boldsymbol{x}) \mathcal{V}(\boldsymbol{x}') | 0 \ldots 0 \rangle$$

to feed into classical kernel methods [9].

Quantum feature framework.—We unify the two above approaches into a quantum feature framework, combining quantum feature maps with an appropriate possible set of observables. We introduce observables $O_1, O_2, \ldots, O_K$, which are Hermitian operators applied to the state $|\Psi(\boldsymbol{x})\rangle$. If we measure O_i, we can obtain the expectation value of this observable and consider it as the basis function $\psi_i(\boldsymbol{x}): \mathcal{X} \to \mathbb{R}$, which is defined as

$$\psi_i(\boldsymbol{x}) = \langle \Psi(\boldsymbol{x}) | O_i | \Psi(\boldsymbol{x}) \rangle = \mathrm{Tr}[O_i | \Psi(\boldsymbol{x}) \rangle \langle \Psi(\boldsymbol{x}) |]. \quad (1)$$

If these basis functions have nonlinearity properties with sufficiently high dimension, we can solve a complex task by the linear regression on the output function $f: \mathcal{X} \to \mathbb{R}$, which is the linear combination of the basis functions $\psi_i(x)$ with the weights $w_i \in \mathbb{R}$ $(i = 1, \ldots, K)$ [28]:

$$f(\boldsymbol{x}) = \sum_{i=1}^{K} w_i \psi_i(\boldsymbol{x}). \quad (2)$$

The observables $\{O_i\}$ should be chosen for easy physical implementation but can produce nonlinearity with sufficient high-dimensional basis functions [29].

Universal approximation property and classification capability.—A quantum feature framework $\mathcal{F}$ based on a set of quantum feature maps and a set of observables on the Hilbert space is defined as the collection of function $f: \mathcal{X} \to \mathbb{R}$, where each f has the form in Eq. (2). We define the UAP and classification capability of $\mathcal{F}$. Let $\mathcal{G}$ be a space of continuous functions $g: \mathcal{X} \to \mathbb{R}$. The framework $\mathcal{F}$ has the UAP with respect to $\mathcal{G}$ and a norm $\|\cdot\|$ if given any function $g \in \mathcal{G}$; then, for any $\varepsilon > 0$, there exists $f \in \mathcal{F}$ such that $\|f - g\| < \varepsilon$. This f is called an approximator of g with ε error. Furthermore, $\mathcal{F}$ has the classification capability if, for arbitrary disjoint regions (i.e., closed sets) $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$ in $\mathcal{X}$, there exists $f \in \mathcal{F}$ such that f can separate these regions [17]. We investigate the UAP and the classification capability in two typical scenarios in setting the quantum feature map. We assume that $\mathcal{X}$ is a compact set. For the sake of readability, we present some definitions

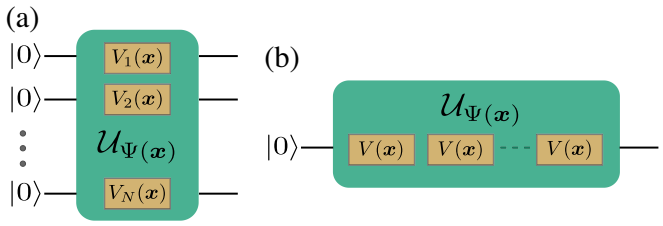

FIG. 2. Quantum circuit $U_{\Psi(\boldsymbol{x})}$ for quantum feature map. (a) Circuit is tensor product of multiple circuits, where each circuit $V_i(\boldsymbol{x})$ acts on a subsystem. (b) Circuit is repetition of simple circuit $V(\boldsymbol{x})$ (for example, a single Pauli-Y rotation) acting on the same qubits.

for notations used in this study. A supremum norm of a function $h: \mathcal{X} \to \mathbb{R}$ is defined as $\|h\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}}|h(\boldsymbol{x})|$. Let $L^2(\mathcal{X})$ be a space of functions $h: \mathcal{X} \to \mathbb{R}$ that is square integrable; that is,

$$\int_{\mathcal{X}} |h(\boldsymbol{x})|^2 d\boldsymbol{x} < \infty.$$

The norm of function h in $L^2(\mathcal{X})$ space is defined as

$$\|h\|_{L^2(\mathcal{X})} = \left[\int_{\mathcal{X}} |h(\boldsymbol{x})|^2 d\boldsymbol{x}\right]^{1/2}.$$

Parallel scenario.—We examine the first scenario where the quantum feature map is a tensor product of multiple quantum circuits acting on subsystems where the number of qubits can be set freely [Fig. 2(a)]. We consider a typical feature map $\Psi_N^{\mathcal{V}}$ represented by the following circuit applied to $|0\rangle^{\otimes N}$:

$$\mathcal{V}_N(\boldsymbol{x}) = V_1(\boldsymbol{x}) \otimes V_2(\boldsymbol{x}) \otimes \ldots \otimes V_N(\boldsymbol{x}), \tag{3}$$

where $V_j(\boldsymbol{x})$ is a single-qubit Pauli rotation: for example, a Y-basis rotation $e^{-i\theta_j(\boldsymbol{x})Y}$ applied to the jth qubit with the function $\theta_j: \mathcal{X} \to \mathbb{R}$. Here,

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},$$

and

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

are the Pauli matrices. We show that the UAP can be obtained via the nonlinearity of the basis functions. This nonlinearity can be introduced by an appropriate selection of observables or by a classical preprocessing, such as using a nonlinear pretransformation for the input.

To begin, we propose a popular setting of $\theta_j(\boldsymbol{x})$ and observables to produce the nonlinearity in the quantum feature framework. Because $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, without a loss of generality, we assume that $\mathcal{X} = [0, 1]^d$. Given the input data $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X}$ and $N \geq d$, we consider the circuits in Eq. (3) with $V_j(\boldsymbol{x}) = e^{-i \arccos(\sqrt{x_k})Y}$, where $1 \leq k \leq d$ and $k \equiv j (\mathrm{mod}\, d)$, $(1 \leq j \leq N)$. The observables are $O_{\boldsymbol{\alpha}} = Z^{\alpha_1} \otimes Z^{\alpha_2} \otimes \ldots \otimes Z^{\alpha_N}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2 \ldots \alpha_N) \in \{0, 1\}^N$. The basis functions are calculated as

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \langle 0|^{\otimes N} \mathcal{V}_N^\dagger(\boldsymbol{x}) O_{\boldsymbol{\alpha}} \mathcal{V}_N(\boldsymbol{x}) |0\rangle^{\otimes N}. \tag{4}$$

From $\{\psi_{\boldsymbol{\alpha}}\}$, we can construct any polynomial function on $\mathcal{X}$ [31]. Due to a special case of the Stone–Weierstrass theorem [36], any continuous function on $\mathcal{X}$ can be approximated by polynomial functions with arbitrary precision in terms of the supremum norm. Therefore, we obtain the following UAP (see proof in the Supplemental Material [31]).

Result 1 (UAP in the parallel scenario).—For any continuous function, $g: \mathcal{X} \to \mathbb{R}$; then, for any $\varepsilon > 0$, there exist N and a collection of output weights $w_{\boldsymbol{\alpha}}$ and observables $O_{\boldsymbol{\alpha}} = Z^{\alpha_1} \otimes Z^{\alpha_2} \otimes \ldots \otimes Z^{\alpha_N}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \{0, 1\}^N$ such that

$$\left|\sum_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - g(\boldsymbol{x})\right| < \varepsilon$$

for all $\boldsymbol{x}$ in $\mathcal{X}$. Here, the basis function $\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is defined as that in Eq. (4).

Result 1 implies that the induced quantum feature framework has the UAP with respect to the supremum norm. Furthermore, we prove the classification capability of this framework. We consider m disjoint regions $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$ in $\mathcal{X}$ and their corresponding m distinct real values as labels $c_1, c_2, \ldots, c_m$. According to lemma 2.1 in Ref. [17], there exists a continuous function h_c such that $h_c(\boldsymbol{x}) = c_i$ if $\boldsymbol{x}$ in $\mathcal{K}_i$. We say that a function $h: \mathcal{X} \to \mathbb{R}$ can separate m disjoint regions $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$ at $\boldsymbol{x}_0$ if

$$|h_c(\boldsymbol{x}_0) - h(\boldsymbol{x}_0)| < \delta = \frac{1}{2}\min\{|c_i - c_j| \,|\, \forall\, i \neq j\}.$$

From result 1, we can obtain a function $f: \mathcal{X} \to \mathbb{R}$ in the form

$$\sum_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}),$$

such that

$$|h_c(\boldsymbol{x}_0) - f(\boldsymbol{x}_0)| < \delta = \frac{1}{2}\min\{|c_i - c_j| \,|\, \forall\, i \neq j\}$$

for all $\boldsymbol{x}_0$ in $\mathcal{X}$. Therefore, f can separate $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$.

We note that the number of observables $O_{\boldsymbol{\alpha}}$ in the parallel scenario does not need to scale exponentially with respect to the number of qubits N. From the construction of the

circuits, for each k $(1 \le k \le d)$, any combination of $\alpha_k, \alpha_{k+d}, \alpha_{k+2d}, \ldots$ with p nonzero elements gives the same terms in the basis functions $\psi_{\boldsymbol{\alpha}}$. Hence, for each p, we only need to choose one combination to construct the observable $O_{\boldsymbol{\alpha}}$. Let $q(k)$ denote the number of values that p can take for each k. Then, the number of observables $O_{\boldsymbol{\alpha}}$ does not need to be larger than $q(1)q(2)\ldots q(d)$. Because the number of elements in $\alpha_k, \alpha_{k+d}, \alpha_{k+2d}, \ldots$ does not exceed $1 + \lfloor (N-1)/d \rfloor$, the value of p is taken in $0, 1, \ldots, 1 + \lfloor (N-1)/d \rfloor$, where $\lfloor r \rfloor$ denotes the greatest integer less than or equal to r. Therefore, $q(k) \le 2 + \lfloor (N-1)/d \rfloor$ for each k; thus, the number of observables does not exceed $(2 + \lfloor (N-1)/d \rfloor)^d$.

Next, we show that the nonlinearity to establish the UAP can be implemented by a special kind of data preprocessing with an activation function incorporated into $\theta_j(\boldsymbol{x})$. The activation function can be computed by a classical algorithm on the level of logical gates and then translated into a reversible routine to be used as a quantum algorithm [37]. Given an activation function $\sigma: \mathbb{R} \to [-1, 1]$, we further assume two conditions for σ. First, σ is nonconstant and piecewise continuous. Here, σ is said to be piecewise continuous if it has a finite number of discontinuities in any interval, and its left and right limits are defined (not necessarily equal) at each discontinuity. Second, $\sigma_{\boldsymbol{a},b}(\boldsymbol{x}) = \sigma(\boldsymbol{a} \cdot \boldsymbol{x} + b)$ is dense in $L^2(\mathcal{X})$, where $\boldsymbol{a} \cdot \boldsymbol{x}$ denotes the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{x}$ in $\mathbb{R}^d$. This means that for any $\varepsilon > 0$ and $g \in L^2(\mathcal{X})$, there exist $\boldsymbol{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $\|g - \sigma_{\boldsymbol{a},b}\|_{L^2(\mathcal{X})} < \varepsilon$. We apply $\mathcal{V}_N(\boldsymbol{x})$ in Eq. (3) with

$$\theta_j(\boldsymbol{x}) = \arccos\left(\sqrt{\frac{1 + \sigma_{\boldsymbol{a}_j,b_j}(\boldsymbol{x})}{2}}\right),$$

where $\boldsymbol{a}_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$ are randomly generated from any continuous sampling distribution for each j. In this scheme, the number of observables can be reduced to N. We consider the observables

$$O_j = I \otimes \ldots \otimes \underbrace{Z}_{j\,\text{index}} \otimes \ldots \otimes I$$

$(1 \le j \le N)$ with the corresponding basis functions

$$\begin{aligned}\psi_j(\boldsymbol{x}) &= \langle 0|^{\otimes N} \mathcal{V}_N^\dagger(\boldsymbol{x}) O_j \mathcal{V}_N(x) |0\rangle^{\otimes N} \\ &= \langle 0| e^{i\theta_j Y} Z e^{-i\theta_j Y} |0\rangle = \sigma_{\boldsymbol{a}_j,b_j}(\boldsymbol{x}). \end{aligned} \tag{5}$$

Result 2 is obtained from the main result in the UAP of the classical framework in Ref. [19] (theorem 2.3), which states that for any $\varepsilon > 0$, there exist N and

$$\{w_j\}_{j=1}^N (w_j \in \mathbb{R})$$

such that

$$\left\| \sum_{j=1}^N w_j \sigma_{\boldsymbol{a}_j,b_j} - g \right\|_{L^2(\mathcal{X})} < \varepsilon.$$

Result 2 (UAP when implementing activation functions in preprocessing).—For any continuous function, $g: \mathcal{X} \to \mathbb{R}$ and the construction of basis functions ψ_j is in Eq. (5); then, for any $\varepsilon > 0$, there exist N and

$$\{w_j\}_{j=1}^N (w_j \in \mathbb{R})$$

such that

$$\left\| \sum_{j=1}^N w_j \psi_j - g \right\|_{L^2(\mathcal{X})} < \varepsilon.$$

Result 2 implies that with a sufficient number of qubits, the framework induced from the nonlinear activation function with the selected observables can work as a universal approximator to any continuous function $g: \mathcal{X} \to \mathbb{R}$ in $L^2(\mathcal{X})$ with any arbitrary precision. Similar to the analysis from result 1, we consider the function h_c to investigate the classification capability in this setting. From result 2, for $\varepsilon > 0$, there exists $f: \mathcal{X} \to \mathbb{R}$ in the form of Eq. (2) such that

$$\|h_c - f\|_{L^2(\mathcal{X})} < \varepsilon.$$

Let $\mathcal{Y} = \{\boldsymbol{y} \in \mathcal{X} | |h_c(\boldsymbol{y}) - f(\boldsymbol{y})| \ge \delta\}$ and $V_{\mathcal{Y}}$ be the volume of $\mathcal{Y}$; we then have $V_{\mathcal{Y}}^{1/2}\delta < \varepsilon$ or $V_{\mathcal{Y}} < (\varepsilon/\delta)^2$. Therefore, by selecting sufficiently small ε, we can reduce $V_{\mathcal{Y}}$ as small as possible to increase the classification capability.

Sequential scenario.—In the parallel scenario, it is assumed that we can increase the number of qubits to approximate the output function to a target continuous function with arbitrary precision. However, there is a limitation in the current realistic model with a large number of qubits. We investigate whether the UAP can be obtained by constructing the quantum feature map with only a single qubit by repeating a simple quantum circuit $V(\boldsymbol{x})$ [Fig. 2(b)]. Unlike the parallel scenario, the quantum feature map described in the following paragraph is not capable of approximating a function whose domain is an infinite set (see the Supplemental Material [31]). We restrict the input set to a finite set $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$. For example, in a real-world application, $\mathcal{X}$ can be the set of RGB fixed-size images.

To obtain the UAP, it is important to set the appropriate form of $V(\boldsymbol{x})$. In the Supplemental Material [31], we present a counterexample of $V(\boldsymbol{x})$ in which we cannot obtain the UAP. Here, we consider the unitary operator $V(\boldsymbol{x}) = e^{-\pi i \theta(\boldsymbol{x}) Y}$ applied to the single qubit and establish the condition of $\theta(\boldsymbol{x}_1), \theta(\boldsymbol{x}_2), \ldots, \theta(\boldsymbol{x}_M)$ to obtain the UAP. The quantum feature map is constructed by repeating $V(\boldsymbol{x})$: that is, applying $V^n(\boldsymbol{x}) = e^{-n\pi i \theta(\boldsymbol{x}) Y}$ $(n \in \mathbb{N})$ to $|0\rangle$, where

$\theta: \mathcal{X} \to \mathbb{R}$. The corresponding basis function with the observable Z (Pauli-Z) becomes

$$\psi_n(\boldsymbol{x}) = \langle 0|(V^n)^\dagger(\boldsymbol{x}) Z V^n(\boldsymbol{x})|0\rangle = 2\cos^2(\pi n\theta(\boldsymbol{x})) - 1 = \cos(2\pi n\theta(\boldsymbol{x})) = \cos(2\pi\{n\theta(\boldsymbol{x})\}), \quad (6)$$

where $\{n\theta(\boldsymbol{x})\} = n\theta(\boldsymbol{x}) - \lfloor n\theta(\boldsymbol{x})\rfloor$ is the fractional part of $n\theta(\boldsymbol{x})$. The UAP is studied via the Kronecker–Weyl theorem [38,39] on the density of the fractional parts $(\{n\theta(\boldsymbol{x}_1)\}, \ldots, \{n\theta(\boldsymbol{x}_M)\})_{n\in\mathbb{N}}$. In the Supplemental Material [31], we prove the following result, which states that with the condition of the linear independence for $1, \theta(\boldsymbol{x}_1), \ldots, \theta(\boldsymbol{x}_M)$, any function in $\mathcal{X}$ can be approximated by repeatedly applying $V(\boldsymbol{x})$ with an appropriate iteration number n. Here, real numbers $b_1, b_2, \ldots, b_L$ are linearly independent over the set of rational numbers $\mathbb{Q}$ if the only integral solution to $z_1 b_1 + z_2 b_2 + \cdots + z_L b_L = 0$ is the all zero $z_1 = z_2 = \ldots = z_L = 0$.

Result 3(UAP in the sequential scenario).—If $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\} \subset \mathbb{R}^d$ and $1, \theta(\boldsymbol{x}_1), \ldots, \theta(\boldsymbol{x}_M)$ are linearly independent over $\mathbb{Q}$, then for any function $g: \mathcal{X} \to \mathbb{R}$ and for any $\varepsilon > 0$, there exist $n \in \mathbb{N}$ and $w \in \mathbb{R}$ such that $|w\psi_n(\boldsymbol{x}) - g(\boldsymbol{x})| < \varepsilon$ for all $\boldsymbol{x}$ in $\mathcal{X}$. Here, the basis function $\psi_n(\boldsymbol{x}) = \cos(2\pi\{n\theta(\boldsymbol{x})\})$ is defined as in Eq. (6).

Similar to the analysis from result 1, we can also obtain the classification capability via result 3.

Approximation rate.—An interesting theoretical question is how to describe relative goodness or badness in a universal approximation. The approximation rate can be used here, which is the decay rate of the approximation error. This rate refers to the speed at which the approximation error decreases when the parameters, such as the number of qubits N and the input dimension d, are increased. The approximation rate strongly depends on the nature of the target function g to be approximated and the type of the input set $\mathcal{X}$. In the Supplemental Material [31], we prove the following result, which describes the approximation rate in the parallel scenario.

Result 4 (approximation rate).—If $\mathcal{X} = [0, 1]^d$ and the target function g is Lipschitz continuous with respect to the Euclidean norm, we can construct an explicit form of the approximator to g in the parallel scenario by N qubits with the error $\varepsilon = O(d^{7/6} N^{-1/3})$. Furthermore, we can achieve an approximation error with a better approximation rate in terms of N as $\varepsilon = O(d^{3/2} N^{-1})$.

The approximation error $\varepsilon = O(d^{3/2} N^{-1})$ can be obtained by using the Jackson theorem of the quantitative information on the degree of polynomial approximation to a continuous function [40]. It implies that $O(d^{3/2}\varepsilon^{-1})$ qubits are enough to obtain an approximation with ε error. However, the explicit form of this approximator remains for future work.

The approximation rate provides a method to compare the asymptotic universality between our quantum feature framework and the classical neural networks. The number of observables K in our framework corresponds with the number of parameters in the classical neural networks. Since $K = O(N^d)$ in the parallel scenario, we can write our best approximation error as $\varepsilon = O(K^{-1/d})$ if we fix d and focus on K. Interestingly, this is also the best approximation when using a classical neural network to approximate a Lipschitz continuous function [41,42]. This result suggests a strong guarantee that the QML models in quantum-enhanced feature spaces can exhibit at least the same expressivity as the classical ML models.

Conclusion.—We present a comprehensive understanding of the UAP of quantum feature frameworks induced from quantum-enhanced feature spaces. This research lays a foundation for further theoretical analysis of the expressivity of these frameworks and provides insights into the design of a good expressive model in QML applications. Our proposal addresses the theoretical research question about whether QML models in quantum-enhanced feature spaces can solve the tasks that conventional ML models can in classical settings. We obtain the results that under typical quantum feature map settings, the QML models can achieve both the UAP and classification capability, and can thus handle a wide class of ML tasks. The suggestions in practical applications are left for future works, such as finding an efficient scheme with the lowest implementation cost to obtain the necessary approximation accuracy.

K. N. and Q. H. T. were supported by MEXT Quantum Leap Flagship Program Grants No. JPMXS0118067394 and No. JPMXS0120319794.

T. G. and Q. H. T. contributed equally to this work.

*Corresponding author.
tran_qh@ai.u-tokyo.ac.jp
†goto.takahiro.2020@gmail.com
‡k_nakajima@mech.t.u-tokyo.ac.jp

[1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).
[2] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).
[3] G. Torlai and R. G. Melko, Machine-learning quantum states in the NISQ era, Annu. Rev. Condens. Matter Phys. **11**, 325 (2020).
[4] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).
[5] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.
[6] K. Fujii and K. Nakajima, Harnessing Disordered-Ensemble Quantum Dynamics for Machine Learning, Phys. Rev. Applied **8**, 024030 (2017).
[7] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[8] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature (London) **567**, 209 (2019).

[9] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[10] D. K. Park, C. Blank, and F. Petruccione, The theory of the quantum kernel-based binary classifier, Phys. Lett. A **384**, 126422 (2020).

[11] C. Blank, D. K. Park, J.-K. K. Rhee, and F. Petruccione, Quantum classifier with tailored quantum kernel, npj Quantum Inf. **6**, 41 (2020).

[12] R. LaRose and B. Coyle, Robust data encodings for quantum classifiers, Phys. Rev. A **102**, 032420 (2020).

[13] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, Quantum embeddings for machine learning, arXiv:2001.03622.

[14] M. Schuld, Supervised quantum machine learning models are kernel methods, arXiv:2101.11020.

[15] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. (2021), https://doi.org/10.1038/s41567-021-01287-z.

[16] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 2631 (2021).

[17] G.-B. Huang, Y.-Q. Chen, and H. Babri, Classification ability of single hidden layer feedforward neural networks, IEEE Trans. Neural Networks **11**, 799 (2000).

[18] G.-B. Huang, L. Chen, C. K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Networks **17**, 879 (2006).

[19] G.-B. Huang and L. Chen, Convex incremental extreme learning machine, Neurocomputing **70**, 3056 (2007).

[20] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. **2**, 303 (1989).

[21] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks **4**, 251 (1991).

[22] W. Y. Huang and R. P. Lippmann, Neural net and traditional classifiers, in *Neural Information Processing Systems*, edited by D. Z. Anderson (American Institute of Physics, New York, 1988), p. 387.

[23] E. Torrontegui and J. J. García-Ripoll, Unitary quantum perceptron as efficient universal approximator, Europhys. Lett. **125**, 30004 (2019).

[24] N. Killoran, T. R. Bromley, J. M. Arrazola, M. Schuld, N. Quesada, and S. Lloyd, Continuous-variable quantum neural networks, Phys. Rev. Research **1**, 033063 (2019).

[25] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, Quantum **4**, 226 (2020).

[26] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, Phys. Rev. A **103**, 032430 (2021).

[27] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).

[28] K. Fujii and K. Nakajima, Quantum reservoir computing: A reservoir approach toward quantum machine learning on near-term quantum devices, in *Reservoir Computing: Theory, Physical Implementations, and Applications*, edited by K. Nakajima and I. Fischer (Springer Singapore, Singapore, 2021), pp. 423–450, https://doi.org/10.1007/978-981-13-1687-6_18.

[29] This scheme is analogous with the classical extreme learning machine (ELM) framework [17,30]. In the ELM, the input data $\boldsymbol{x}$ are fed into a single- or multilayer perceptron where all weights between layers are fixed. The states of hidden nodes at some layers are regarded as basis functions that play a similar role as $\psi_i(\boldsymbol{x})$.

[30] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomputing **70**, 489 (2006).

[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.090506 for proofs of results 1–4, which include Refs. [32–35].

[32] C. Heitzinger, Simulation and inverse modeling of semiconductor manufacturing, Ph.D. thesis, Technische Universität Wien, 2002.

[33] M. M. Wouodjié, On multivariate Bernstein polynomials, Master's thesis, University of Yaoundé I, 2014.

[34] M. Foupouagnigni and M. M. Wouodjié, On multivariate Bernstein polynomials, Mathematics **8**, 1397 (2020).

[35] P. Davis, *Interpolation and Approximation*, Dover Books on Mathematics (Dover, New York, 1975).

[36] K. Yoshida, *Functional Analysis* (Springer-Verlag, Berlin, 1980), p. 9.

[37] M. Schuld and F. Petruccione, *Supervised Learning with Quantum Computers*, 1st ed. (Springer, New York, 2018), p. 184.

[38] J. L. King, Three problems in search of a measure, Am. Math. Mon. **101**, 609 (1994).

[39] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction* (Princeton University, Princeton, NJ, 2003), Vol. 1, p. 105.

[40] D. Newman and H. Shapiro, Jackson's theorem in higher dimensions, in *Proceedings of Conference on Approximation Theory in Oberwolfach* (Springer-Birkhäuser, Basel, Switzerland, 1964), p. 208.

[41] H. N. Mhaskar, Neural networks for optimal approximation of smooth and analytic functions, Neural Comput. **8**, 164 (1996).

[42] D. Yarotsky, Error bounds for approximations with deep ReLU networks, Neural Networks **94**, 103 (2017).