Quantum Fisher kernel for mitigating the vanishing similarity issue

Yudai Suzuki1,∗, Hideaki Kawaguchi2, and Naoki Yamamoto2,3

1 Department of Mechanical Engineering, Keio University, Hiyoshi 3-14-1, Kohoku, Yokohama 223-8522, Japan
2 Quantum Computing Center, Keio University, Hiyoshi 3-14-1, Kohoku, Yokohama 223-8522, Japan
3 Department of Applied Physics and Physico-Informatics, Keio University, Hiyoshi 3-14-1, Kohoku, Yokohama 223-8522, Japan

∗ Author to whom any correspondence should be addressed.

E-mail: yudai.suzuki.sh@gmail.com

Keywords: quantum machine learning, quantum kernel methods, quantum information, quantum fisher kernel

Abstract
Quantum kernel (QK) methods exploit quantum computers to calculate QKs for the use of kernel-based learning models. Despite a potential quantum advantage of the method, the commonly used fidelity-based QK suffers from a detrimental issue, which we call the vanishing similarity issue; the exponential decay of the expectation value and the variance of the QK deteriorates implementation feasibility and trainability of the model with the increase of the number of qubits. This implies the need to design QKs alternative to the fidelity-based one. In this work, we propose a new class of QKs called the quantum Fisher kernels (QFKs) that take into account the geometric structure of the data source. We analytically and numerically demonstrate that the QFK can avoid the issue when shallow alternating layered ansätze are used. In addition, the Fourier analysis numerically elucidates that the QFK can have the expressivity comparable to the fidelity-based QK. Moreover, we demonstrate synthetic classification tasks where QFK outperforms the fidelity-based QK in performance due to the absence of vanishing similarity. These results indicate that QFK paves the way for practical applications of quantum machine learning toward possible quantum advantages.

1. Introduction
Quantum computers have the potential to enhance existing machine learning models in terms of performance and computational speed. Thus far, there have been several proposals of quantum machine learning (QML) algorithms that outperform the classical counterparts for certain classes of problems [1–4]. An example of QML with possible quantum advantage is the quantum kernel method that utilizes quantum computing for classical kernel methods [5, 6]. It has been demonstrated that quantum kernel methods in combination with classical linear classifiers such as support vector machines, successfully classify some data that classical models cannot efficiently separate. For instance, a synthesized dataset inspired by the discrete logarithmic problem (DLP) has been proposed [4]. In addition, a recently proposed procedure that could screen intrinsic quantum advantages of the method [7] has led to explorations of real-world datasets [8, 9]. Also, the relationship between quantum kernel methods and the so-called quantum neural networks has been discussed, e.g. in reference [10], emphasizing the importance of the method in supervised QML frameworks.

Quantum kernel methods potentially have quantum advantages because the corresponding Hilbert space, which is considered hard for classical computers to access efficiently, is used as a feature space for machine learning tasks. On the other hand, the use of the large Hilbert space hinders the performance and implementation of the method. Quantum kernel methods measure the similarity between a pair of data $x$ and $x'$ using a function called the quantum kernel (QK), defined as the fidelity between data-dependent quantum states [5]:

$$\text{Fidelity} = \langle \rho_x | \rho_{x'} \rangle$$

where $\rho_x$ and $\rho_{x'}$ are the quantum states corresponding to data $x$ and $x'$, respectively. However, the fidelity-based QK suffers from the vanishing similarity issue, where the expectation value and variance of the fidelity decay exponentially with the number of qubits, making it impractical for large-scale applications. To mitigate this issue, the quantum Fisher kernel (QFK) is proposed in this work, which takes into account the geometric structure of the data source.

Supplementary material for this article is available online.
Here, \( \rho_{x, \theta} = U(x, \theta)\rho_0 U(x, \theta)^\dagger \) is the density operator representation of the quantum state generated by the input- and parameter-dependent unitary \( U(x, \theta) \) with initial state \( \rho_0 \). Then, the Gram matrix composed of QKs given all data pairs is applied to machine learning tasks such as regression and classification. However, the fidelity-based QK in equation (1) has a detrimental issue, which we call the vanishing similarity issue, stating that all off-diagonal elements of the Gram matrix (similarity between different data pairs) significantly vanish as the number of qubits increases. More precisely, the expectation value and the variance of those elements decrease exponentially with respect to the number of qubits. This means that an exponential number of measurement shots is required to precisely estimate the QK, which erases the possible quantum advantage. In other words, a realistic number of measurements on real quantum hardware yields a Gram matrix close to the identity matrix; consequently, overfitting happens and the generalization performance of classifiers or regressors using the Gram matrix could be poor.

A concept equivalent to the vanishing similarity issue was first introduced in [7], followed by some attempts to analytically understand the phenomenon [11–13]. However, this issue has not been resolved yet. On the other hand, an analogy of the vanishing similarity issue in variational quantum algorithms—the barren plateau problem [14]—can be mitigated by considering cost function designs [15–17] and the structure of the parameterized quantum circuits (PQCs) such as the so-called alternating layered ansatz (ALA) [15]. This gives us insight into a circumventing approach for the vanishing similarity issue; that is, we should design a QK that takes into account the data source structure through a feature map \( U(x, \theta) \), instead of the fidelity-based one.

In this work, we propose a novel QK called the quantum Fisher kernel (QFK), as a quantum extension of the classical Fisher kernel [18]. The Fisher kernel is constructed using the information-geometric quantity of the data source (i.e. the logarithmic derivatives of the generative model), which, as a result, incorporates data structures into kernel designs [18–20]. Thus we derive QFKs that utilize \( \rho_{x, \theta} = U(x, \theta)\rho_0 U(x, \theta)^\dagger \) as a generative model constituting a set of density-operator-valued data \( \{\rho_{x, \theta}\} \). Specifically, we examine the symmetric logarithmic derivative (SLD) [21, 22] and the anti-symmetric logarithmic derivative (ALD) [23].

Then, with a focus on the vanishing similarity issue, we calculated expectation values and variances of the fidelity-based QK and QFK, assuming quantum circuits satisfy the property of a 2-design [24–26]. To be more specific, we work on two types of quantum circuits: (1) random quantum circuits acting on all qubits and (2) ALAs. We found that the variance of QFK does not depend on the number of qubits but on the size of unitary blocks in ALAs and the depth of the corresponding unitary block, while the same issue arises in the case of random quantum circuits. That is, QFK can avoid the issue when shallow ALAs are used. However, the fidelity-based QK suffers from vanishing similarity for both cases, regardless of circuit depth. Numerical simulations also verify these results. Figure 1 summarizes the aforementioned results.

Moreover, we numerically show the Fourier representation of the QFK and the fidelity-based QK to demonstrate that they have comparable expressivity. We then performed classification tasks using one-dimensional synthesized datasets to validate the Fourier analysis. In addition, we used the synthesized dataset to demonstrate an example where QFK performs well, but the performance of the fidelity-based QK deteriorates due to the vanishing similarity issue. Hence, our proposed QFK with ALAs can avoid the vanishing similarity issue and possibly show better performance than the fidelity-based QK when a large number of qubits are used, which could help to find a practical advantage of QFKs in machine learning tasks.

Lastly, we remark that an approach to avoid the vanishing similarity issue is to use projected QKs [7] that reduce the effective dimension by projecting data-embedded quantum states onto a low-dimensional space. However, we will not go deep into the approach in this study, as our motivation is to exploit the large Hilbert space using the method inspired by the established classical Fisher kernel. We also mention that even projected QKs may suffer from the vanishing similarity issue [11, 13], although there is more room for further exploration.

2. Results

2.1. Motivating examples

To begin with, we numerically demonstrate the vanishing similarity issue for the fidelity-based QK in equation (1) configured with two types of quantum circuits \( U(x, \theta) \): tensor-product quantum circuits and instantaneous quantum polynomial (IQP) type quantum circuits [5] with depth \( L = 2 \). Here, each layer comprises an input-embedded circuit and a PQC, and the data re-uploading technique is employed [27].

The details are provided in section V.A of supplementary information (SI section V.A). Figure 2 shows expectation values and variances of the fidelity-based QK where each element of input data \( x \) and parameters in PQC \( \theta \) is randomly chosen from the range \( [-\pi, \pi] \). As shown in figure 2, they both decrease exponentially.
fast with respect to the number of qubits. We note that the IQP-type quantum circuit was first applied to quantum kernel methods in reference [5] with the motivation that this circuit is conjectured to be hard for classical computers to simulate efficiently [28, 29]. Hence, the presence of vanishing similarity in this case is critical, as such possible quantum advantage can be erased. Even worse, the tensor-product quantum circuit, which is efficiently simulatable by classical means, witnesses the vanishing similarity issue, indicating that even less-expressive quantum circuits with hopeless quantum advantage can suffer from the same issue.

2.2. Vanishing similarity issue in the fidelity-based QK
We state the vanishing similarity issue in the fidelity-based QK in general settings. We recall that vanishing similarity is the phenomenon where the expectation value and the variance of the QK taken over the ensemble of a pair of data-dependent unitary operators, \(\{U(x, \theta), U(x', \theta)\}\), decrease exponentially in the
Figure 3. Quantum circuits used in our analysis and quantum circuit representations of the structure used in QFK of equation (13), $\hat{B}_{k,0} = U_{1,i}^\dagger(x, \theta) B_k U_{1,i}(x, \theta)$. Panels (a) and (b) show the random quantum circuit acting on all qubits and the ALA, respectively. Similarly, panels (c) and (d) represent $\hat{B}_{x,\theta}^i$ for the random quantum circuit and the ALA. For the ALA in panel (d), the thick gray unitary block adjacent to $B_k$ is represented by $\hat{W}_{k,d}(x, \theta)$ and the shaded region represent $V_{r}(x, \theta)$. Reproduced with permission from [30].

number of qubits. Note that the variance plays a more critical role than the expectation in the vanishing similarity issue; that is, even if the mean value is large, exponentially small variance results in QKs that output almost the same value regardless of a pair of inputs. As a result, infeasible implementation and untrainability arise.

Here, we analytically obtain the expectation value and the variance of the QK for two types of quantum circuits (i.e. unitary operators): (1) random quantum circuits acting on all qubits and (2) ALAs, as shown in figures 3(a) and (b), respectively. The ALA is a brick-like quantum circuit with alternating layers of $m$-qubits local unitary blocks, which could mitigate the barren plateau issue thanks to the circuit configuration [15]. We here suppose that the total depth is $L$ and the number of unitary blocks in each layer is $k$, satisfying $n = mk$ for the total number of qubits $n$. Also, the $k$th unitary block in the $d$th layer is denoted as $W_{k,d}(x, \theta_{k,d})$. The quantum circuits can be composed of data-dependent gates, parameter-dependent gates, and data- and parameter-independent gates, where each gate can be expressed as entangling gates and single-qubit rotation gates, i.e. $R_{\sigma}(\theta) = \exp(-i\theta\sigma/2)$. See SI section I.B for the details of the circuit settings.

Our analysis is based on the assumption that the random quantum circuits and local unitary blocks in ALAs are 2-designs. Recall that the $t$-design is an ensemble of unitaries that have the same statistical properties as the unitary group with respect to the Haar measure up to the $t$th moment [24–26]. The assumption has been used to analyze vanishing gradients in variational quantum algorithms [14, 15, 31, 32] and explore the expressivity of quantum circuits [33, 34]. The details are given in SI section I.A. Then, we have the following result (the proof is provided in SI section II).

**Proposition 1.** Let the expectation value and the variance of the $n$-qubit fidelity-based QK defined in equation (1) be $\langle k_0 \rangle$ and $\text{Var}[k_0]$, respectively. Also, let the initial state $\rho_0$ be an arbitrary pure state.

1. When $U(x, \theta)$ and $U(x', \theta)$ are the independent random quantum circuits acting on all qubits, and at least either $U(x, \theta)$ or $U(x', \theta)$ is a $t$-design with $t \geq 2$, the expectation value and the variance are given by

\[ \langle k_0 \rangle \quad \text{and} \quad \text{Var}[k_0], \]
\begin{align}
\langle k_{Q} \rangle &= \frac{1}{2^n}, \\
\text{Var}[k_{Q}] &= \frac{2^n - 1}{2^{2n} (2^n + 1)} \approx \frac{1}{2^n}.
\end{align}

(2) Let \( m \)-qubit local unitary blocks in ALAs, \( U(x, \theta) \) and \( U(x', \theta) \), be independent and \( t \)-designs with \( t \geq 2 \). Then, the expectation value and the upper bound of the variance are given by

\begin{align}
\langle k_{Q} \rangle &= \frac{1}{2^n}, \\
\text{Var}[k_{Q}] &\leq \frac{2^\kappa}{(2^{2m} - 1)^\kappa} - \frac{1}{2^{2m}} \approx \frac{1}{2^n (1 - \frac{1}{n})}.
\end{align}

Proposition 1 states that, for both types of circuits, all off-diagonal elements of the Gram matrix given by the fidelity-based QK take almost zeros if the number of qubits \( n \) is large, because of exponential decay of (the upper bound) of the variance. Note that every diagonal element, i.e. \( k_{Q}(x, x) \), is \( 1 \) for any input \( x \) regardless of the number of qubits used. Also, the implication of proposition 1 is that devising the circuit structure will not circumvent the issue, as long as the fidelity is used as the metric. We remark that the class of ALAs includes the tensor-product quantum circuit studied in the motivating example, as a special case (i.e. \( L = 1, m = 1 \) and \( \kappa = n \)). Moreover, the initial state \( \rho_{0} \) can be extended to a mixed state with slight modification; see SI section II. We should mention that the similar result for the case (1) can be seen in [11, 13] and the one for the case (2) could be related to equation (170) in [15]. However, proposition 1 (2) still has an important implication, since this is the first to show the vanishing similarity issue for the case of ALAs from the perspective of the quantum kernel method.

Lastly, let us note that we derive the analytical results in proposition 1 by integrating the unitary operators \( U(x, \theta) \) and \( U(x', \theta) \) independently while there could be a correlation because of the shared parameters \( \theta \). However, our analytical results can be valid even in the case where some dependent unitary operators are included in the quantum circuits and local unitary blocks in the ALAs. We elaborate the details in SI section IV.A.

2.3. QFKs

We define QFKs by taking into account the structure of the quantum data \( \rho_{x, \theta} \) used for QML frameworks. As described in Introduction, QFKs are based on the idea of classical Fisher kernel [18] defined as

\[ k_{f}(x, x') = \langle g(x, \theta), g(x', \theta) \rangle_{\mathcal{I}^{-1}} = g^\top(x, \theta) \mathcal{I}^{-1} g(x', \theta), \]

where \( g(x, \theta) = \nabla_{\theta} \log P(x, \theta) \) is the logarithmic derivative of the generative model \( P(x, \theta) \) (called the Fisher score) and \( \mathcal{I} = \mathbb{E}_{x}[g(x, \theta) g^\top(x, \theta)] \) is the Fisher information matrix. The motivation behind the Fisher kernel is to construct a powerful classifier from the probabilistic generative model, utilizing the knowledge of data structure via the Fisher score, a natural feature vector in the space of probability distributions. The classical Fisher kernel has been applied in several areas such as computer vision, thanks to its specialized expressivity to data and the performance for some tasks [35–38].

To define QFKs, we should first note that there are multiple definitions for the quantum versions of the Fisher score [39]. In this work, we focus on the SLD [21, 22] and the ALD [23]. The SLD \( \mathcal{L}_{x, \theta}^{S} \) and the ALD \( \mathcal{L}_{x, \theta}^{A} \) with respect to the \( l \)-th parameter \( \theta_{l} \) for \( \rho_{x, \theta} = U(x, \theta) \rho_{0} U^\dagger(x, \theta) \) are defined as solutions of the following equations, respectively;

\begin{align}
\partial_{\theta_{l}} \rho_{x, \theta} &= \frac{1}{2} \left( \rho_{x, \theta} \mathcal{L}_{x, \theta}^{S} + \mathcal{L}_{x, \theta}^{S} \rho_{x, \theta} \right), \\
\partial_{\theta_{l}} \rho_{x, \theta} &= \frac{1}{2} \left( \rho_{x, \theta} \mathcal{L}_{x, \theta}^{A} - \mathcal{L}_{x, \theta}^{A} \rho_{x, \theta} \right).
\end{align}

Here, \( \partial_{\theta_{l}} \) represents the partial derivative with respect to the parameter \( \theta_{l} \), i.e. \( \partial_{\theta_{l}} \equiv \partial / \partial \theta_{l} \). It is known that these quantities are not uniquely determined. However, if the initial state is pure, one of the solutions for the SLD equation can be expressed as

\[ \mathcal{L}_{x, \theta}^{S} = 2 \partial_{\theta_{l}} \rho_{x, \theta}. \]

Also, a solution of the ALD equation can be obtained as

\[ \mathcal{L}_{x, \theta}^{A} = i \left( B_{x, \theta} - \text{Tr}[\rho_{x, \theta} B_{x, \theta}] \right), \]

where \( B_{x, \theta} \) is a Hermitian operator.
with $B_{x,θ} = 2i(∂_θ U(x, θ))U^†(x, θ)$. Then, as in the definition of classical Fisher kernel in equation (6), we define the QFK as an inner product of these logarithmic derivatives; namely, using $L_{x,θ}^γ = [L_{x,θ}^γ, L_{x,θ}^{γ',...}]$ with $γ = \{S,A\}$,

$$
k_{QF}(x,x') ≡ \left< L_{x,θ}^γ, L_{x',θ}^{γ'} \right> = \frac{1}{2} \sum_{ij} F_{γ,ij}^{-1} \left\{ \rho_0 \left< L_{x,θ}^γ, L_{x',θ}^{γ'} \right> \right\} ,
$$

(11)

In the second line of equation (11), we introduce a pre-inner product for operators $[23]$, $(A,A')_ρ = \frac{1}{2} \text{Tr} [ρ(A′ A^† + A A^†)]$ using certain quantum state $ρ$. We note that the pre-inner product works as the inner product in our case due to the property of the SLD and the ALD operators, i.e. $S_{x,θ}^γ = (S_{x,θ}^{γ})^†$ and $L_{x,θ}^{A,θ} = -(L_{x,θ}^{A,θ})^†$. Here, $F_{γ}(S_{x,θ})$ is the SLD-based (ALD-based) quantum Fisher information matrix, where $(i,j)$-element is given by $F_{γ,ij} = (S_{x,θ}^{i,θ} S_{x,θ}^{j,θ})_{ρ}$. Let us also notice that the Hilbert-Schmidt inner product can be used as the inner product of operators to define QFKs. For more details on the QFK using the Hilbert-Schmidt inner product, see SI section VI.

Interestingly, QFK in equation (11) can further be expressed in the same form for the SLD derived from pure states and the ALD. The details of the derivation for these cases are presented in SI section III. Then, the QFK can be rewritten as

$$
k_{QF}(x,x') = -\frac{1}{2} \sum_{ij} F_{γ,ij}^{-1} \text{Tr} \left[ ρ_0 \left< L_{x,θ}^{A,θ}, L_{x',θ}^{A,θ} \right> \right] ,
$$

(12)

where $\{·,·\}$ is the anti-commutator and $L_{x,θ}^{A,θ} = U^†(x, θ)U_{x,θ}^{A,θ}$. Henceforth, the superscript $γ \in \{A,S\}$ of QFK is omitted for the sake of clarity.

In the main text below, the quantum Fisher information matrix in QFK of equation (12) is set as the identity matrix, i.e. $F = I$, as in the classical case [18]; this is because reference [18] demonstrated that the Fisher information matrix in classical Fisher kernel is insignificant with respect to the performance, and previous works have practically used the identity matrix [18, 38, 40–42] or the diagonal matrix approximating the Fisher information matrix [36, 43] due to the computational efficiency. The impact of the quantum Fisher information matrix on the variance and the computational cost is discussed in SI section IV.B. Moreover, without loss of generality, we do not take into account the term $\text{Tr}[ρ_0 B_{x,θ}]$ in the ALD of equation (10) to simplify the discussion on the vanishing similarity issue. We remark that the same scalings in variance hold in the theorem shown below, even for QFK with the term $\text{Tr}[ρ_0 B_{x,θ}]$; the proof is provided in SI section IV. In that case, QFK can be expressed as

$$
k_{QF}(x,x') = \frac{1}{2} \sum_{i} \text{Tr} \left[ ρ_0 \left< \tilde{B}_{x,θ}, \tilde{B}_{x',θ} \right> \right] ,
$$

(13)

where $\tilde{B}_{x,θ} = U_{i,j}^†(x, θ)B_{θ} U_{i,j}(x, θ)$. Here, $U_{i,j}(x, θ)$ denotes a bunch of unitary gates from $U_i(x, θ)$ to $U_j(x, θ)$, in the circuit representation $U(x, θ) = U_{i,θ} U_{j,θ} \cdots U_2(x, θ) U_1(x, θ)$. We remind that $θ_i$ is the rotation angle of single-qubit rotation gates.

2.4. Main results

Here we examine if QFK could avoid the vanishing similarity issue, under the assumption that $U(x, θ)$ is (1) random quantum circuits acting on all qubits or (2) ALAs. In this section, we consider the $i$th term $k_{QF}^{(i)} ≡ \text{Tr}[ρ_0 \left< \tilde{B}_{x,θ}^i, \tilde{B}_{x',θ}^i \right>] / 2$ of equation (13). The variance of $k_{QF}^{(i)}$ is different from the actual variance of QFK in equation (13). However, we rather focus on its variance to understand how deep quantum circuits can be exploited via QFK for machine learning tasks. Recall that the fidelity-based QK is subjected to the vanishing similarity issue for both cases and even single-layer quantum circuits lead to the exponential decay, as shown in proposition 1.

For ease of analysis, we assume that the random quantum circuit $U_{i,j}(x, θ)$ is a 2-design for arbitrary $i$. For ALAs, we assume that the $i$th parameter $θ_i$ is located in the $k$th unitary block in the $d$th layer of the circuits, $W_{k,d}(x, θ_{k,d})$. In addition, we decompose the ALA as $U_{i,j}(x, θ) = W_{k,d}(x, θ_l)V_{r}(x, θ)$, where $W_{k,d}(x, θ_l)$ denotes a sequence of gates that includes the first gate in $W_{k,d}(x, θ_{k,d})$ through the gate with $i$th parameter. Also, $V_{r}(x, θ)$ denotes a sequence of the unitary blocks in the light-cone of $W_{k,d}(x, θ_{k,d})$. Then, we assume unitary blocks in $V_{r}(x, θ)$ and $W_{k,d}(x, θ_l)$ for any $k$ and $d$ are 2-designs. Figures 3(c) and (d) show quantum circuit representations of $\tilde{B}_{x,θ}$ for the cases of random quantum circuits and ALAs. Under the above setting, we have the following result; the proof is given in SI section III.
Theorem 1. Let the expectation value and the variance of \( k^{(i)}_{QF} = \text{Tr}[\rho_{0} \{ \tilde{B}_{x, \theta}, \tilde{B}_{x', \theta'} \}] / 2 \) in the n-qubit QFK defined in equation (13) be \( \langle k^{(i)}_{QF} \rangle \) and \( \text{Var}[k^{(i)}_{QF}] \), respectively. Also, let the initial state \( \rho_{0} \) be a pure state.

1. When \( U(x, \theta) \) and \( U(x', \theta') \) are the independent random quantum circuits acting on all qubits, and both \( U_{1, i}(x, \theta) \) and \( U_{1, i}(x', \theta') \) are t-designs with \( t \geq 2 \), then we have

\[
\langle k^{(i)}_{QF} \rangle = 0, \quad \text{Var}[k^{(i)}_{QF}] = \frac{2^n}{2^{(2n-1)}} \left( 1 + \frac{2^n - 2}{2^{(2n+1)}} \right) \approx \frac{1}{2^{n+1}}. \tag{14}
\]

2. Let \( \tilde{W}_{i, a}(x, \theta) \), \( \tilde{W}_{i, a}(x', \theta) \) and unitary blocks in \( V_{i}(x, \theta) \) and \( V_{i}(x', \theta) \) for ALAs \( U(x, \theta) \) and \( U(x', \theta) \) be independent and t-designs with \( t \geq 2 \). Then, the expectation value is given by

\[
\langle k^{(i)}_{QF} \rangle = 0. \tag{16}
\]

Additionally, we assume the initial state \( \rho_{0} \) is represented as the tensor product of arbitrary single-qubit pure states \( \rho_{0} = \rho_{0,1} \otimes \rho_{0,2} \otimes \ldots \otimes \rho_{0,j} \otimes \ldots \otimes \rho_{0,n} \). Then, the lower bound of the variance is given by

\[
\text{Var}[k^{(i)}_{QF}] \geq \frac{2^{md}(2^{md} - 1)}{2^{(2^{m} - 1)}(2^{m} + 1)^{4(d-1)}}. \tag{17}
\]

We first remark that the assumption on the initial state for the variance calculation in case (2) is moderate from the practical perspective, since the tensor product state is a common choice for initial state preparation. Further, we can derive the lower bound of the variance of the QFK for a larger class of initial states. The details are provided in SI section III.B.

Theorem 1 together with proposition 1 implies that the variance of the QFK is less likely to decrease with respect to the number of qubits than that of the fidelity-based QK. We recall that the large variance can prevent the QK from outputting almost the same value regardless of a pair of inputs, suggesting the absence of the vanishing similarity issue. In case (1), although the QFK shows an exponential decrease as well, there is a nearly quadratic difference in the scaling compared to the fidelity-based QK. Moreover, in case (2), where ALAs are used, the lower bound of the variance does not depend on the total number of qubits, but on the size of unitary block \( m \) and the depth \( d \) of the local unitary block \( W_{i, a}(x, \theta_{k, a}) \). In other words, each term \( k^{(i)}_{QF} \) of QFK for shallow ALAs can avoid the vanishing similarity issue; this is because the light-cone structure prevents its expressivity from increasing too much with respect to the circuit depth.

Importantly, QFK has the potential to exploit quantum circuits whose depth is possibly \( O(\text{polylog}(n)) \) for machine learning tasks. This scaling is according to [15], stating that the transition point between exponential and polynomial decay would lie in the region of depth \( d \in O(\text{polylog}(n)) \); thus, 3th term of QFK can be obtained with an affordable number of measurement shots up to certain point of the region. We remind that QFK is the summation of \( k^{(i)}_{QF} \) over all \( i \), and we show in SI section IV.C that the variance of QFK in equation (13) follow the same scaling as that shown in theorem 1. Note that the variance of \( k^{(i)}_{QF} \) scales exponentially in the depth of the local unitary block \( d \), but terms \( k^{(i)}_{QF} \) lying in the depth region \( d \in O(\text{polylog}(n)) \) can still be exploited; that is, the weighted sum representation of QFKs, i.e.

\[
k_{QF}(x, x') = \frac{1}{p} \sum_{i} w_{i} \text{Tr}[\rho_{0} \{ \tilde{B}_{x, \theta}, \tilde{B}_{x', \theta'} \}] \text{ with properly chosen weights } \{ w \} \text{ can utilize such small components. This would help to seek out a practical advantage for real-world datasets due to its flexibility; however, investigating the weighted-sum QFK will be left for our future work. We also note that the same results can be obtained in a situation where unitary operators } U(x, \theta) \text{ and } U(x', \theta) \text{ include dependent unitary operators, which is shown in SI section IV.A.}

2.5. Numerical experiments

We performed numerical simulations to verify proposition 1 and theorem 1. Here, the variance of the fidelity-based QK and QFK are numerically computed for three types of circuits: tensor-product quantum circuits, ALAs with 2-qubit local unitary blocks, and hardware efficient ansatzes (HEAs). We used fixed entangling gates and single-qubit rotation gates whose rotation axes, i.e. the Pauli operators \{ X, Y, Z \}, are randomly chosen, to make HEAs serve as the random circuits. However, we mention that HEAs and unitary blocks in ALAs may not form t-designs with \( t \geq 2 \). See SI section V.B for the details.

In the experiments, we randomly generated 100 data points \( \{ x \}_{i=1}^{100} \) and a set of PQC parameters \( \theta \). We then calculated the QKs for all different pairs of data, \( x \neq x' \). We repeated this process 25 times with different sets of parameters and inputs to obtain the variance of each QK. As for QFK, we use a normalized version, i.e.

\[
\bar{k}_{QF}(x, x') = k_{QF}(x, x') / p. \tag{18}
\]
Figure 4. Variance of the fidelity-based QK and the QFK with respect to the number of qubits, \(n = 2, 4, 6, 8, 10, 12, 14\). Three types of quantum circuits (tensor-product quantum circuits, ALAs with 2-qubit local unitary blocks and HEAs) with depth \(L = 3\) are used. Reproduced with permission from [30].

where \(p\) is the number of parameters. We divide QFK by \(p\) so that the trace of its Gram matrix coincides with the number of data points as in the fidelity-based QK. For numerical simulations, we used the state vector simulator provided by Cirq, a software library for quantum computing [44]. The detailed settings of the numerical experiments are provided in SI section V.B.

Figure 4 shows a semi-log plot of the variance of QKs against the number of qubits. As expected from proposition 1, the variance of the fidelity-based QK vanishes exponentially fast, regardless of the types of quantum circuits. On the other hand, QFK does not show such an exponential decrease of variance for every circuit. Indeed, the variance of QFK for the ALA decreases with respect to the number of qubits in figure 4, contrary to theorem 1 (2). However, this is attributed to the fact that a normalization factor for QFK, \(p\), linearly depends on the number of qubits in our setting. We also remark that the variance of QFK for HEAs does not decrease exponentially in the number of qubits, which looks inconsistent with theorem 1 (1). This is because the assumption on the 2-design property of quantum circuits is not satisfied due to the insufficient expressivity of the quantum circuits \(U_{1,\ell}(x, \theta)\) for any \(i\). Furthermore, in SI section V.B, we numerically checked the variance of the \(i\)th term \(k^{(i)}_Q\) for different layers to support our analytical result in theorem 1. Then, we found that these results are consistent with the analytical results. Overall, the numerical experiments confirm the vanishing similarity issue in the fidelity-based QK in proposition 1, and the validity of theorem 1 in the main result.

2.6. Expressivity analysis and performance comparison

We have addressed that QFK can be free from the vanishing similarity issue due to the structure of ALAs. However, QFK may lose the expressivity, another important property that any machine learning model should fulfill. That is, it needs to be clarified how large class of functions can be approximated by the QFK. Thus, we performed numerical analysis to show that QFK has almost the same level of expressivity as the fidelity-based QK.

In general, QK can be represented as an inner product of two Fourier series [10];

\[
k(x, x') = \sum_{\omega, \omega' \in \Omega} e^{i\omega x} e^{i\omega' x'} c_{\omega, \omega'},
\]

where \(c_{\omega, \omega'} \in \mathbb{C}\) is the Fourier coefficient satisfying \(c_{\omega, \omega'} = c_{-\omega, -\omega'}^{*}\), and \(\Omega\) is the set of integer-valued frequencies. Hence, we can gauge the expressivity by numerically calculating the magnitude of \(c_{\omega, \omega'}\) over the effective frequency set. Since the Fourier decomposition becomes computationally challenging for the case of high-dimensional data, we used 100 one-dimensional data points and the ALAs with \(n = 1, 2, 3\) and depth \(L = 2, 3, 4\) in the experiments. Also, we truncate the set of frequency to \(\tilde{\Omega} \in \{-12, -11, \ldots, 11, 12\}\). The details are given in SI section V.C.

Figure 5 shows the amplitudes of Fourier coefficients \(c_{\omega, \omega'}\) for the fidelity-based QK and the normalized QFK, in settings of the number of qubits \(n = 1, 2, 3\) and the circuit depth \(L = 3\). Due to the symmetry in Fourier coefficients, we just focus on half of them; see SI section V.C for the details and the results of the rest. Here, we use the 'curve_fit' function in Scipy [45] to obtain the coefficients by fitting each QK to its Fourier representation. As shown in figure 5, QFK has almost the same non-zero Fourier coefficients as those of the
The quantum circuit used in this experiment consists of four layers, where we employed the data re-uploading technique \([27]\). We choose a tensor-product quantum circuit for input embedding and the ALA with 2-qubits unitary blocks as the PQC for each layer. When embedding the one-dimensional data into the quantum circuit, the data is rescaled according to the position of the qubit on which the rotation gate acts. Precisely, we rescale the inputs as \(x \rightarrow (n_{\text{pos}}[i] + 1)x\) with the qubit position the gate containing ith parameter acts on, \(n_{\text{pos}}[i] \in \{0, 1, \ldots, n - 1\}\). We remark that rescaling the input means the shift of the power spectrum of the PQC and is thus equivalent to artificially adjusting the expressivity of the circuit \([47, 48]\). Hence, such rescaling is a natural choice especially when we do not know in advance how difficult the task is. As we did for the classification tasks to validate the Fourier analysis and performance of QKs. In summary, from the results shown in figure 5, our conclusion is that QFK has the expressivity comparable to the fidelity-based QK.

Finally, we demonstrate an example where QFK can outperform the fidelity-based QK in performance due to the absence of vanishing similarity. We again consider the aforementioned synthesized dataset with \(w = 2\) and \(b = 0.3\) and examine the performance of each QK using the different numbers of qubits. The quantum circuit used in this experiment consists of four layers, where we employed the data re-uploading technique \([27]\). We choose a tensor-product quantum circuit for input embedding and the ALA with 2-qubits unitary blocks as the PQC for each layer. When embedding the one-dimensional data into the quantum circuit, the data is rescaled according to the position of the qubit on which the rotation gate acts. Precisely, we rescale the inputs as \(x \rightarrow (n_{\text{pos}}[i] + 1)x\) with the qubit position the gate containing ith parameter acts on, \(n_{\text{pos}}[i] \in \{0, 1, \ldots, n - 1\}\). We remark that rescaling the input means the shift of the power spectrum of the PQC and is thus equivalent to artificially adjusting the expressivity of the circuit \([47, 48]\). Hence, such rescaling is a natural choice especially when we do not know in advance how difficult the task is. As we did for the classification tasks to validate the Fourier analysis, we implemented the support vector machines as classifiers. The detailed experimental setup is shown in SI section V.C.

Figure 7(a) shows the accuracy of the fidelity-based QK and QFK with the different numbers of qubits on the synthesized dataset. We note that accuracy is defined as the number of correct predictions divided by the total number of test data. From figure 7(a), the performance of the fidelity-based QK clearly gets worse with the increase of qubit numbers, while QFK performs well consistently. This difference is caused by the vanishing similarity issue, which can be confirmed by the Gram matrices shown in figure 7(b). Namely, off-diagonal elements in the Gram matrix of the fidelity-based QK are vanishing as the number of qubits increases; on the other hand, off-diagonal elements for QFK remain sufficiently large even for 14 qubits used.
Figure 6. Comparison between accuracy of 2-qubit QFK for the synthesized datasets and its expressivity given by Fourier analysis. (a) The synthesized datasets with \((w, b) = (2, 0.3)\) and \((4, 0.3)\) are illustrated. (b) Upper panel shows accuracy of QFKs using ALAs with depth 3 (red) and 4 (blue) for the dataset with \(w \in \{2, \ldots, 12\}\), while lower panel exhibits amplitudes of the corresponding QFKs’ Fourier coefficients \(c_{\omega, -\omega}\) against the frequency \(\omega \in \{2, \ldots, 12\}\). Note that learning synthesized dataset with \(w\) requires non-zero Fourier coefficients \(c_{\omega, -\omega}\) with \(\omega = w\).

The results indicate the potential of QFK to have better performance than the fidelity-based QK when a large number of qubits are used. Besides the synthesized datasets, QFKs could outperform fidelity-based QKs, e.g. on parity-based classification tasks used in reference [49], because of the absence of the vanishing similarity issue.

3. Discussion

To demonstrate the potential of quantum advantage in machine learning tasks, it has been well recognized that the data structure should be incorporated into the models [4, 50]. From this perspective, it is important to design a QK that can take into account the data structure, instead of the fidelity-based one that would suffer from the vanishing similarity issue. In this work, we propose QFK as a quantum extension of the classical Fisher kernel, which has already been developed in the classical regime to incorporate the structure of data (generative models) into kernel designs. We showed that QFK with shallow ALAs may avoid the vanishing similarity issue in quantum kernel methods, even for large quantum systems. Also, the Fourier analysis indicates that QFK has almost the same expressivity as the fidelity-based QK. Furthermore, we demonstrated a situation where QFK can achieve high accuracy even with a large number of qubits, while the performance of the fidelity-based QK deteriorates with the increase of the number of qubits.

Although the classical Fisher kernel is not as well-known as the Gaussian kernel, QFK might hold an essentially important position in QML field due to the aforementioned desirable features. In addition, QFK
can be related to quantum dynamics. The structure in QFK, i.e. $UBU^\dagger$, can be regarded as the extent to which the unitary cancellation process is affected by $B$. Indeed, the similar structures appear in quantities such as the out-of-time-ordered correlator function [51, 52] and Loschmidt Echo [53, 54], both of which are used to investigate quantum chaos and quantum information scrambling. Therefore, the performance of QFK can be investigated from the viewpoint of quantum dynamics, which hopefully suggests us some pathway toward quantum advantage.

Lastly, we discuss the practical usability of QFK. First, we note that some previous works have shown the potential of quantum kernel methods to outperform classical models on practical datasets such as electric health records [9], weather data [8], high energy physics related data [55], the modified Fashion-MNIST [7] and graph-structured data [56, 57] using some types of QKs. Moreover, [58] demonstrates that the $k$-Forrelation problem, which is Promise BQP-complete and considered hard for classical models to learn efficiently, can be solved by QKs. These examples rely on the clever choice of feature maps suitable for the data source of given tasks. Importantly, QFK intrinsically incorporates the data structure into kernel designs through the quantum version of the Fisher score and also enjoys the trainability property guaranteed by theorem 1. Additionally, theorem 1 implies that the weighted-sum QFK is more beneficial because quantum circuits with depth in $d \in \mathcal{O}(\text{polylog}(n))$ could be exploited, while the fidelity-based quantum kernel cannot utilize even shallow quantum circuits. Therefore, our view is that QFK will work better than the fidelity-based QK and thereby serve as a promising QML method for the practical use. As a future work, it would be worth exploring the performance of the QFK for practical tasks, e.g. where high-dimensional data is handled.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This work was supported by Grant-in-Aid for JSPS Fellows 22J14183, and MEXT Quantum Leap Flagship Program Grants Nos. JPMXS0118067285 and JPMXS0120319794.
Author contributions

The idea of this project was conceived by Y S and N Y, Y S conducted analytical and numerical calculations in this work. All authors contributed to the analysis of the results, the writing of the manuscript and the continuous improvement of the manuscript.

ORCID iD

Yudai Suzuki  https://orcid.org/0000-0003-1444-8124

References

[1] Farhi E et al 2001 A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem Science 292 472–5
[2] Rebentrost P, Mohseni M and Lloyd S 2014 Quantum support vector machine for big data classification Phys. Rev. Lett. 113 130503
[3] Biamonte J et al 2017 Quantum machine learning Nature 549 195–202
[4] Liu Y, Arunachalam S and Temme K 2021 A rigorous and robust quantum speed-up in supervised machine learning Nat. Phys. 17 1015–1017
[5] Havlíček V et al 2019 Supervised learning with quantum-enhanced feature spaces Nature 567 209–12
[6] Schuld M and Killoran N 2019 Quantum machine learning in feature Hilbert spaces Phys. Rev. Lett. 122 040504
[7] Huang H-Y et al 2021 Power of data in quantum machine learning Nat. Commun. 12 2631
[8] Enos G R et al 2021 Synthetic weather radar using hybrid quantum-classical machine learning (arXiv:2111.15605)
[9] Krunic Z, Flöther F, Seeger G, Earnest-Noble N D and Shehab O 2022 Quantum kernels for real-world predictions based on electronic health records IEEE Trans. Quantum Eng. 3 1–11
[10] Schuld M 2021 Supervised quantum machine learning models are kernel methods (arXiv:2101.11020)
[11] Kübler J, Buchholz S and Schölkopf B 2021 The inductive bias of quantum kernels Advances in Neural Information Processing Systems vol 34 pp 12661–73
[12] Canatar A, Peters E, Pelelevan C, Wild S M and Shaydulin R 2022 Bandwidth enables generalization in quantum kernel models (arXiv:2206.06868)
[13] Thanasilp S, Wang X, Cerezo M and Holmes Z 2022 Exponential concentration and untrainability in quantum kernel methods (arXiv:2208.11060)
[14] McClean J R, Boixo S, Smelyanskiy V N, Babbush R and Neven H 2018 Barren plateaus in quantum neural network training landscapes Nat. Commun. 9 4812
[15] Cerezo M, Sone A, Voldkoff T, Cincio L and Coles P J 2021 Cost function dependent barren plateaus in shallow parametrized quantum circuits Nat. Commun. 12 1791
[16] Khatri S et al 2019 Quantum-assisted quantum compiling Quantum 3 140
[17] LaRose R, Tikku A, O’Neel-Judy E, Cincio L and Coles P J 2019 Variational quantum state diagonalization npj Quantum Inf. 5 57
[18] Jaakola T and Haussler D 1998 Exploiting generative models in discriminative classifiers in Advances in Neural Information Processing Systems vol 11
[19] Tsuda K, Akaho S, Kawanabe M and Müller K-R 2004 Asymptotic properties of the fisher kernel Neural Comput. 16 115–37
[20] Hofmann T, Schölkopf B and Smola A J 2008 Kernel methods in machine learning Annu. Stat. 36 1171–220
[21] Helstrom C W 1967 Minimum mean-squared error of estimates in quantum statistics Phys. Lett. A 25 101–2
[22] Liu J, Yuan H, Lu X-M and Wang X 2019 Quantum fisher information matrix and multiparameter estimation J. Phys. A: Math. Theor. 53 023001
[23] Fujiwara A and Nagaoka H 1995 Quantum fisher metric and estimation for pure state models Phys. Lett. A 201 119–24
[24] Harrow A and Mehraban S 2023 Approximate unitary r-designs by short random quantum circuits using nearest-neighbor and long-range gates Commun. Math. Phys. 401 1531–626
[25] Renes J M, Blume-Kohout R, Scott A J and Caves C M 2004 Symmetric informationally complete quantum measurements J. Math. Phys. 45 2171–80
[26] Klappenecker A and Rotteler M 2005 Mutually unbiased bases are complex projective 2-designs Proc. Int. Symp. on Information Theory 2005, IST 2005 (IEEE) pp 1740–4
[27] Pérez-Salinas A, Cervera-Lierta A, Gil-Fuster E and Latorre J I 2020 Data re-uploading for a universal quantum classifier Quantum 4 226
[28] Bremer M J, Montanaro A and Shepherd D J 2016 Average-case complexity versus approximate simulation of commuting quantum computations Phys. Rev. Lett. 117 080501
[29] Goldberg L A and Guo H 2017 The complexity of approximating complex-valued ising and Tutte partition functions Comput. Complexity 26 765–833
[30] Suzuki Y 2024 Applicability of quantum-enhanced machine learning PhD Thesis Keio University
[31] Holmes Z et al 2021 Barren plateaus preclude learning scramblers Phys. Rev. Lett. 126 190501
[32] Holmes Z, Sharma K, Cerezo M and Coles P J 2022 Connecting ansatz expressibility to gradient magnitudes and barren plateaus PRX Quantum 3 010313
[33] Sim S, Johnson P D and Aspuru-Guzik A 2019 Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms Adv. Quantum Technol. 2 1900070
[34] Nakaji K and Yamamoto N 2021 Expressibility of the alternating layered ansatz for quantum computation Quantum 5 434
[35] Perornini F, Sánchez J and Mensink T 2010 Improving the Fisher kernel for large-scale image classification European Conf. on Computer Vision (Springer) pp 143–56
[36] Sánchez J, Perornini F, Mensink T and Verbeek J 2013 Image classification with the Fisher vector: Theory and practice Int. J. Comput. Vis. 105 222–45
[37] Sydorov V, Sakurada M and Lampert C H 2014 Deep Fisher kernels - end to end learning of the Fisher kernel gmm parameters Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) pp 1402–9
[38] Gudovskiy D, Hodgkinson A, Yamaguchi T and Tsukizawa S 2020 Deep active learning for biased datasets via fisher kernel self-supervision Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition pp 9041–9
[39] Petz D 1996 Monotone metrics on matrix spaces Linear Algebr. Appl. 244 81–96
[40] Hofmann T 1999 Learning the similarity of documents: An information-geometric approach to document retrieval and categorization Advances in Neural Information Processing Systems vol 12
[41] Van Der Maaten I 2011 Learning discriminative Fisher kernels Int. Conf. Machine Learning vol 11 pp 217–24
[42] Bishop C M and Nasrabadi N M 2006 Pattern Recognition and Machine Learning vol 4 (Springer)
[43] Perronnin F and Dance C 2007 Fisher kernels on visual vocabularies for image categorization 2007 IEEE Conf. Computer Vision and Pattern Recognition (IEEE) pp 1–8
[44] Cirq-Developers 2022 Cirq (v0.14.1) Zenodo https://doi.org/10.5281/zenodo.6599601
See full list of authors on Github: https://github.com/quantumlib/Cirq/graphs/contributors
[45] Virtanen P et al 2020 SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python Nat. Methods 17 261–72
[46] Buitinck L et al 2013 API design for machine learning software: experiences from the scikit-learn project (arXiv:1309.0238)
[47] Schuld M, Swayne R and Meyer J J 2021 Effect of data encoding on the expressive power of variational quantum-machine-learning models Phys. Rev. A 103 032430
[48] Wen J, Huang Z, Cai D and Qian L 2024 Enhancing the expressivity of quantum neural networks with residual connections (arXiv:2401.15871)
[49] Jerbi S et al 2023 Quantum machine learning beyond kernel methods Nat. Commun. 14 517
[50] Glick J R et al 2024 Covariant quantum kernels for data with group structure Nat. Phys. 20 479–83
[51] Hashimoto K, Murata K and Yoshii R 2017 Out-of-time-order correlators in quantum mechanics J. High Energy Phys. 2017 138
[52] Swingle B 2018 Unscrambling the physics of out-of-time-order correlators Nat. Phys. 14 988–90
[53] Gorin T, Prosen T, Seligman T H and Žnidarič M 2006 Dynamics of Loschmidt echoes and fidelity decay Phys. Rep. 435 33–156
[54] Goussev A, Jalabert R A, Pastawski H M and Wisniacki D 2012 loschmidt echo (arXiv:1206.6348)
[55] Wu S L et al 2021 Application of quantum machine learning using the quantum kernel algorithm on high energy physics analysis at the LHC Phys. Rev. Res. 3 033221
[56] Henry L-P, Thabet S, Dalyac C and Henriot L 2021 Quantum evolution kernel: Machine learning on graphs with programmable arrays of qubits Phys. Rev. A 104 032416
[57] Albrecht B et al 2023 Quantum feature maps for graph machine learning on a neutral atom quantum processor Phys. Rev. A 107 042615
[58] Jäger J and Krems R V 2023 universal expressiveness of variational quantum classifiers and quantum kernels for support vector machines Nat. Commun. 14 576