Numerical evidence against advantage with quantum fidelity kernels on classical data

Lucas Slattery, Ruslan Shaydulin, Shouvanik Chakrabarti, Marco Pistoia, Sami Khairy, and Stefan M. Wild

Numerical evidence against advantage with quantum fidelity kernels on classical data

Lucas Slattery,[1, 2] Ruslan Shaydulin,[3] Shouvanik Chakrabarti,[3]
Marco Pistoia,[3] Sami Khairy,[2, *] and Stefan M. Wild[2, 4]

[1]*Department of Physics, University of Illinois Urbana-Champaign, Champaign, IL 61820 USA*
[2]*Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439 USA*
[3]*Global Technology Applied Research, JPMorgan Chase, New York, NY 10017 USA*
[4]*Applied Mathematics & Computational Research Division,*
Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA
(Dated: April 17, 2023)

Quantum machine learning techniques are commonly considered one of the most promising candidates for demonstrating practical quantum advantage. In particular, quantum kernel methods have been demonstrated to be able to learn certain classically intractable functions efficiently if the kernel is well-aligned with the target function. In the more general case, quantum kernels are known to suffer from exponential "flattening" of the spectrum as the number of qubits grows, preventing generalization and necessitating the control of the inductive bias by hyperparameters. We show that the general-purpose hyperparameter tuning techniques proposed to improve the generalization of quantum kernels lead to the kernel becoming well-approximated by a classical kernel, removing the possibility of quantum advantage. We provide extensive numerical evidence for this phenomenon utilizing multiple previously studied quantum feature maps and both synthetic and real data. Our results show that unless novel techniques are developed to control the inductive bias of quantum kernels, they are unlikely to provide a quantum advantage on classical data that lacks special structure.

I. INTRODUCTION

Quantum computers are known to have the theoretical potential to solve certain problems faster than their classical counterparts [1]. Machine learning problems are of particular interest due to their wide applicability in science [2–5] and industry [6, 7], and the possibility of large quantum speedups [8]. Quantum machine learning methods applied to quantum data (i.e., data given as quantum states) are known to provide an exponential advantage over classical machine learning methods [9, 10]. On the other hand, if the input data is classical (such as vectors of finite-precision numbers), exponential end-to-end quantum speedup for machine learning is only known for very specifically crafted data [11, 12]. The lack of provable guarantees for classical data motivates the development of techniques that improve the performance of quantum machine learning methods with the goal of empirically demonstrating quantum advantage.

A particular class of quantum machine learning methods we study in this work is quantum fidelity kernel methods applied to classical data. In quantum fidelity kernel methods, a classical data point is embedded in a quantum state using a parameterized quantum evolution (typically, a parameterized quantum circuit) called a quantum feature map [13–18]. As with classical kernel methods, the goal of this embedding is to make the data easier to analyze. The kernel function is given by a pairwise similarity measure on the embedded data points (e.g., the fidelity of two embedded quantum states). The

kernel matrix is a matrix of all such pairwise similarities between input data points. The quantum kernel can then be used with classical kernel machines, such as the support vector machine (SVM), in order to perform clustering, classification, or regression [10, 14, 18–26].

Because most learning properties can be inferred from the properties of the kernel matrix alone, quantum kernel methods are particularly amenable to theoretical analysis. In fact, many of the known speed-ups and improved generalization with fewer data points for quantum machine learning have been demonstrated using the quantum kernel approach [9, 11, 27]. The task-model alignment framework [28] quantifies the number of samples a model needs to learn a target function. For each eigenmode of the kernel, the number of samples required to learn the mode depends on the corresponding eigenvalue. Therefore a "flat" spectrum implies a large sample complexity of learning each mode, making generalization near impossible for any target.

The task-model alignment perspective brings to the fore the central challenge of quantum machine learning, namely the need to introduce inductive bias into the quantum model to enable it to learn. Kübler et al. [10] observe that without an inductive bias, the quantum model is guaranteed to be unable to generalize. Specifically, Kübler et al. show that the spectrum of the kernel becomes "flat", and the largest eigenvalue becomes exponentially small with the number of qubits. Therefore the model requires exponentially many (with number of qubits) training points to learn even the first mode. As a consequence, for a quantum model to generalize efficiently, inductive bias must be introduced into the model [2, 10, 18–21, 29–34].

Recently proposed approaches for introducing induc-

---

* Now at Microsoft.

tive bias into quantum machine learning models include projected kernels [22], group-invariant machine learning [35, 36], and quantum kernel bandwidth [21, 25]. All three approaches prevent exponential "flattening" of the kernel spectrum and enable provable generalization in some settings. In projected or biased kernels, the quantum state generated by a feature map is projected onto a lower dimensional space of the full Hilbert space of the quantum state [22]. However, appropriately selecting the feature map and projection for the kernel requires additional information about the target function that is, in most cases, unavailable. Previous results suggest that projected kernels are unlikely to provide advantage on classical data [22]. The group-invariant quantum machine learning approach has been proposed for quantum neural networks [37], but applies trivially to quantum kernels as well [38]. Specifically, problem-dependent symmetries are introduced into the feature map [35]. However, this is only possible if such symmetries are present in the data. Finally, bandwidth tuning enables generalization by tuning a hyperparameter controlling the expressiveness of the model. This general approach has been shown to enable generalization on both synthetic and real data [21] and does not depend on knowing any special property of the problem. In this paper, we focus on bandwidth tuning for improving the performance of quantum kernels.

Our main contribution is to provide strong numerical evidence that when quantum fidelity kernels generalize well, there are classical kernels that perform similarly well. Specifically, we tune the bandwidth of the quantum kernel so that the first eigenvalue of the kernel does not decay with the number of qubits, thereby ensuring that there exists at least one mode that can be learned with a number of training points that does not grow with the number of qubits. We then observe that the geometric difference between the quantum kernel and an appropriately constructed classical kernel decays rapidly as the number of qubits grows. Since a large geometric difference between quantum and classical kernels is required for quantum advantage [22], our result suggests that quantum fidelity kernels are unlikely to provide an advantage on common classical datasets without novel techniques for controlling the inductive bias.

The rest of the paper is organized as follows. In Section II, we introduce the theoretical background underpinning our numerical experiments. We introduce our numerical experiment design as well as the feature maps and data distributions that we study in Section III. In Section IV, we present the results of our numerical experiments. In Section V, we discuss the context and the limitations of the numerical experiments presented in this work. Finally, in Section VI we discuss the potential of other quantum machine learning methods such as quantum neural networks to improve upon our results and suggest future research directions.

II. BACKGROUND

We begin by briefly reviewing the relevant concepts. We study the problem of classification. Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^P$ with $\mathbf{x}$ drawn from some marginal probability distribution on $\mathcal{X} \subset \mathbb{R}^n$ and $\bar{f}(x_i) = y_i$, the goal is to learn the target function $\bar{f}(x)$. In our numerical experiments with real data, we focus specifically on the problem of binary classification, where $y_i \in \{-1, +1\}$, though our results apply to quantum supervised learning in general.

For an n-qubit quantum system, we consider quantum feature maps ψ that map a data point $\mathbf{x}$ drawn from some distribution defined by a probability density function $p : \mathcal{X} \to \mathbb{R}$ to an n-qubit quantum state given by a density matrix $|\psi(\mathbf{x})\rangle \langle \psi(\mathbf{x})| = \rho(\mathbf{x})$. We denote by $\mathcal{H}$ the quantum Hilbert space of all density matrices encoding n-qubit quantum states endowed with inner product $\langle \rho, \rho' \rangle_{\mathcal{H}} = \text{Tr}(\rho \rho')$. We can interpret $\mathcal{H}$ as a finite-dimensional reproducing kernel Hilbert space (RKHS) associated with the kernel $k(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle_{\mathcal{H}} = \text{Tr}(\rho(\mathbf{x})\rho(\mathbf{x}'))$. Then for a set of data points $\{\mathbf{x}_j\}_{j=1}^P$ the kernel matrix is defined element-wise as $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Its continuous analogue is the integral kernel operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ defined according to its action:

$$(T_k f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') p(\mathbf{x}) d\mathbf{x}, \qquad (1)$$

with orthonormal eigenfunctions $T_k \phi_k = \eta_k \phi_k$ from Mercer's theorem [39]. While many theoretical results are formulated in terms of the eigenvalues of the integral operator (e.g., [10]), the eigenvalues γ_j of the kernel matrix K approach the corresponding eigenvalues η_j of the integral operator in the limit of infinite number of data points. Therefore in numerical experiments we focus on studying the spectrum of the kernel matrix K.

For quantum advantage in machine learning, two necessary (but not sufficient) conditions must hold. First, the quantum model must generalize well. The framework of task-model alignment [28], which is based on the replica method of statistical physics [40–43], allows us to connect the spectral properties of the kernel to model generalization. Informally, the target function can be decomposed in the eigenbasis of the integral operator T_k. Then the model needs to "learn" a sufficient number of eigenmodes of T_k so that the cumulative weight of the target on these modes is close to one. The cost of learning each mode is dependent on the corresponding eigenvalue η. As a consequence, if the largest eigenvalue $\eta_{\max}$ is small, even the first mode cannot be learned, and generalization is impossible. A simplified version of this observation is given in Ref. [10, Theorem 3], wherein for a fixed target function the number of data points needs to scale as $\frac{1}{\sqrt{\gamma_{\max}}}$ to keep the kernel ridge regression error constant. Interested readers are referred to Refs. [10, 21, 28] for a more detailed discussion.

Second condition is that the quantum model must be "sufficiently different" from *all* classical models. Con-

sider a quantum model that is linear in the (quantum) feature space (i.e., $f(x) = Tr(O\rho(x))$). This class of models includes quantum kernel methods. In Huang et al. [22], the authors demonstrate that the prediction error for the quantum model and trained quantum/classical kernel model $h(x)$ can be upper bounded by the trained model's complexity $s_K(N)$. The simplified prediction error bound for a new test sample x is given by:

$$\mathbb{E}_x = |h(x) - f(x)| \leq c\sqrt{\frac{s_K(N)}{N}} \sim \mathcal{O}\left(\sqrt{\frac{s_K(N)}{N}}\right) \quad (2)$$

for $c > 0$ and $s_K(N)$ defined as:

$$s_K(N) = \sum_i^N \sum_j^N (K^{-1})_{ij} Tr(O\rho(x_i)) Tr(O\rho(x_j)) \quad (3)$$

Furthermore, the authors introduce the asymmetric quantity, the geometric difference g_d:

$$g_d(K_1 \| K_2) = \sqrt{\|\sqrt{K_2} K_1^{-1} \sqrt{K_2}\|_\infty} \quad (4)$$

which quantifies the difference between two kernel matrices and is independent of data labels. The trained model complexities can then be shown to satisfy the inequality $s_{K_1} \leq g(K_1 \| K_2)^2 s_{K_2}$ when $s_{K_1} \leq s_{K_2}$. Thus, for a quantum kernel, K_Q to significantly outperform a classical kernel K_C on a dataset it is necessary but not sufficient that $g_d(K_C \| K_Q) \ll \sqrt{N}$. To satisfy the necessary condition, the geometric difference needs to be minimized over many optimized classical machine learning models. However, to show that this condition does not hold, it is sufficient to present one classical model with a small geometric difference, which is the approach we take in this work.

To solve the classification problem and obtain generalization error on real data, we use the quantum kernel matrix in a support vector machine (SVM). SVM aims to find a separating hyperplane maximizing the distance to any data point in either class (margin) [44]. The hyperplane is obtained by solving the following convex optimization problem

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \quad (5)$$

subject to $\sum_{i=1}^N \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$, $i = 1, \ldots, N$, where C is a regularization hyperparameter.

III. METHODS

To numerically investigate the possibility of quantum advantage with quantum kernels, we study the scaling behavior with the number of qubits for kernels that are guaranteed to generalize on at least some target with a fixed number of training points. To enforce this condition, we follow Ref. [25] and perform hyperparameter tuning on the quantum feature maps. For the real datasets, we tune the hyperparameters in order to maximize the model test score. For the unlabeled synthetic data sets, the test score is not available. We therefore tune the hyperparameters in order to keep the maximum kernel matrix eigenvalue fixed at a target eigenvalue with increasing system size in our simulations, which guarantees that the model is able to generalize with a constant (with number of qubits) number of training data points on at least one target function. Next, we obtain the geometric difference g_d between the quantum fidelity kernel matrix and a classical kernel matrix computed on the same data distribution. Specifically, we perform a grid search on the hyperparameter of our classical kernel in order to minimize geometric difference between the quantum and classical kernel matrices. We now discuss our choice of datasets and quantum feature maps.

A. Feature Maps

We consider two distinct families of feature maps in our numerical experiments. We consider a feature map in Eq. 6 inspired by the instantaneous quantum polynomial (IQP) feature map studied in Ref. [15].

$$|\psi(x_i)\rangle = \hat{U}_Z(x_i) \hat{H}^{\otimes d} \hat{U}_Z(x_i) \hat{H}^{\otimes d} |0\rangle \quad (6)$$

where

$$\hat{U}_Z(x_i) = \exp\left(\sum_{j=1}^d \lambda x_{ij} \hat{\sigma}_j^z + \sum_{j,k=1}^d \lambda^\alpha x_{ij} x_{ik} \hat{\sigma}_j^z \hat{\sigma}_k^z\right). \quad (7)$$

Our IQP style feature map has a single hyperparameter, the scaling factor λ. In our numerical experiments, the scaling factor λ will be tuned to control the eigenvalue spectrum of our kernel matrix. The feature map also has the parameter α which we keep fixed during hyperparameter tuning. We study feature maps with $\alpha \in \{0.5, 1.0, 2.0\}$. When $\alpha = 2.0$, the feature map is identical to that studied in Eq. 4 of Ref. [25]. In this case, the scaling factor λ can be seen as re-scaling the data points $x_i \rightarrow \lambda x_i$ and the scaling factor behaves like bandwidth in classical kernel methods [25].

We also consider a feature map using 1D Heisenberg interactions

$$|\psi(x_i)\rangle = \left(\prod_{j=1}^d \exp\left(-i\hat{H}(x_{ij})\right)\right)^{n_l} |0\rangle, \quad (8)$$

where

$$\hat{H}(x_{ij}) = \vec{S}_j \cdot \vec{S}_{j+1} + \lambda x_{ij} \sigma_j^z. \quad (9)$$

In the Heisenberg feature map, we again have a single hyperparameter, the scaling factor λ. Throughout our numerical experiments, we fix the number of layers in Eq. 8, n_l, to 4.

In our numerical experiments, we choose our classical kernel to be a 'classical' version of our quantum fidelity kernel. In the case of the IQP style feature map, we compare our quantum fidelity kernel with a fidelity kernel produced by the feature map

$$|\psi(x_i)\rangle = \hat{U}_Z(x_i)\hat{H}^{\otimes d}|0\rangle. \qquad (10)$$

Note that while the output distribution of this feature map is believed to be hard to reproduce classically the fidelity kernel produced by the inner product of the feature map is classically simulatable in polynomial time [15, 32]. In the case of Heisenberg feature map, we compare to a product state feature map wherein we have removed the Heisenberg interaction term:

$$|\psi(x_i)\rangle = \prod_{j=1}^{d} \exp\left(-i\lambda x_{ij}\sigma_j^z\right)|0\rangle \qquad (11)$$

We choose these "classical versions" of the kernels (6), (8) because they are reasonable ansatze for minimizing the geometric difference between the quantum fidelity kernel and a classical kernel. However, other classical kernels, including more sophisticated neural network architectures could also have been chosen that may have recovered the same or smaller geometric differences. For all feature maps studied the number of qubits is equal to the data point dimension.

B. Data Distributions

As outlined in Sec. II, a large task-model alignment indicates that a kernel method will be able to generalize efficiently to new data [28]. A prerequisite for this is that the eigenvalue spectrum of the kernel must not be "flat", that is the largest eigenvalue must not be "too small". Concretely, for the number of training data points required to obtain constant generalization error with the number of qubits, the largest eigenvalue must not decrease. In our analysis, we examine the largest eigenvalue of our kernel matrix and the geometric difference between a quantum and classical kernel. Our analysis is agnostic to any accompanying labels to data points. We examine both real data distributions and synthetically generated unlabeled data distributions. For real datasets, we use the Plasticc dataset and the Fashion-Mnist dataset. Our synthetic data distributions are generated by randomly sampling data elements x_{ij} in our data points x_i from the symmetric generalized normalized distribution

$$x_{ij} \sim \frac{\beta}{2\sigma\Gamma(1/\beta)} \exp\left(-\left|\frac{x-\mu}{\sigma}\right|^\beta\right), \qquad (12)$$

where $\mu = 0$ and $\sigma = 1$. After sampling all data points for a data distribution, the data elements are standardized and normalized. In our experiments, we study distributions with $\beta \in \{0.1, 1.0, 2.0\}$. In the case of $\beta = 1.0$ and 2.0 the distribution reduces to the laplacian and the normal distribution respectively. The effect of β on the distribution is that smaller values of β lead to larger kurtosis of the distribution (kurtosis $= \frac{\Gamma(5/\beta)\Gamma(1/\beta)}{\Gamma(3/\beta)^2} - 3$).

We also consider a correlated data distribution using the symmetric generalized normalized distribution. In this case, the correlated data elements are generated using

$$x_{ij} \sim r \cdot x_{i(j-1)} + $$
$$\sqrt{1-r^2} \cdot \frac{\beta}{2\sigma\Gamma(1/\beta)} \exp\left(-\left|\frac{x-\mu}{\sigma}\right|^\beta\right), \qquad (13)$$

where the expected correlation between the data element x_{ij} and $x_{i(j-1)}$ is r.

IV. NUMERICAL EXPERIMENTS

We now present the numerical evidence that quantum advantage with quantum fidelity kernel methods is unlikely on classical data. We observe that as the number of qubits grows, the quantum model behaves "more classically" as evidenced by the decreasing geometric difference with classical kernels. For the Plasticc and Fashion-Mnist data sets, the decrease holds both for fixed classical kernel and the best-performing one. We denote the geometric difference between the classical and quantum kernels, $g_d(K_C \| K_Q)$, as simply g_d.

A. IQP Style Feature Map

We first present the numerical results for the IQP style feature map (Eq. 6) and the classical fidelity kernel generated by the feature map in (Eq. 10) for $\alpha \in \{0.5, 1.0, 2.0\}$. We present results for the Plasticc and Fashion-Mnist data sets and for synthetic data sets. All kernel matrices use N=500 randomly sampled data points from the uncorrelated data distribution (Eq. 12) unless explicitly stated otherwise.

1. $\alpha = 2.0$, Real Datasets

In this section, we present results for the IQP style feature map with $\alpha = 2.0$ applied to the Plasticc data set and the Fashion-Mnist data. We sample from the Plasticc dataset 700 training vectors and 200 test vectors. For the Fashion-Mnist data set, we sample 700 training vectors and 200 test vectors in addition to varying the number of training vectors in Fig. 2c. For number of qubits 4 to 20, we perform principal component analysis
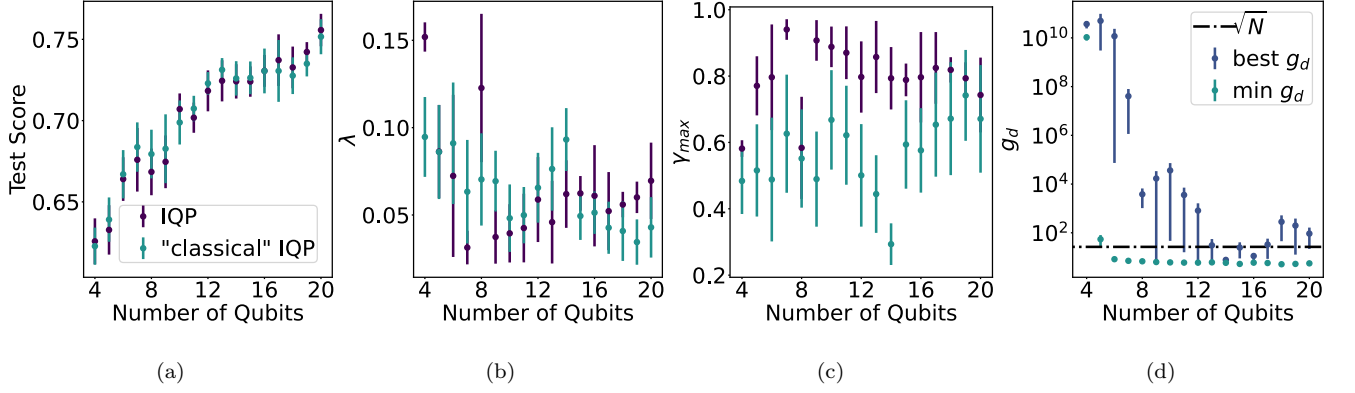
FIG. 1. For the Plasticc dataset, we plot the test score (a), scaling factor λ (b), maximum eigenvalue γ_{max} (c), and geometric difference (d) vs. the number of qubits for both the quantum and 'classical' IQP style feature map with $\alpha = 2.0$. For all for plots, we show results for the highest test score kernels. In (d), we also include the minimum g_d found between the highest test score quantum kernel and a grid search over classical kernels with varying λ.

(PCA) on the data points and reduce the data set dimension to the number of qubits in our feature map before performing standardization and normalization. For both the IQP style feature map (Eq. 6) and the 'classical' IQP feature map (Eq. 10), we tune the scaling factor, λ, and the SVM regularization parameter, C in order to find the best performing SVM on the test data for each feature map. In Fig. 1, we show the results for the best performing feature maps for each system size studied.

For the Plasticc data set, we plot four key quantities as a function of the number of qubits: test score, scaling factor λ, maximum kernel eigenvalue γ_{max} and the geometric difference, g_d. We observe that including more principal components increases the test score performance of both the quantum and classical fidelity kernel (Fig. 1(a)). The IQP style feature map may perform marginally better than the specific classical model we compare it to and have larger γ_{max} (Fig. 1(c)). In Fig. 1d, we plot g_d against the number of qubits for between the highest test score kernels for both the classical and quantum IQP style feature map. In addition, we plot the minimum g_d between the highest test score quantum kernel and a grid search over the classical feature map scaling factor. We observe a decay in the geometric difference g_d with increasing number of qubits (Fig. 1(d)). While we make no definitive claims about the future trend of g_d, for the system sizes we study, the g_d between the two top-performing kernels approaches $\sqrt{N}$. This indicates, for the data distribution (independent of the labels), the two models can still differ in their predictions. However, the minimum g_d found between the models indicates that the potential for quantum advantage decreases with system size.

For the Fashion-Mnist data set, we plot two key quantities vs. the number of qubits. As in the Plasticc data set, we plot the test score and g_d vs. the number of qubits. In addition, we plot g_d vs. the number of training vectors. Again, we observe that including more principal components increases the test score of both the quantum and

classical fidelity kernel (Fig. 2a). For this data set, the "classical" IQP fidelity kernel performs better than the quantum kernel on the test set. As before, we observe the decay of g_d with increasing number of qubits (Fig. 2b). For this data set, the g_d between the top kernels and the minimum g_d found via grid search over the classical feature map scaling factor both decay below $\sqrt{N}$. The decay in g_d indicates the potential for the compared models to make different predictions decreases with system size. In addition to decay with the number of qubits, we observe that the minimum g_d grows slower than $\sqrt{N}$, suggesting that quantum advantage is unlikely. Fig. 2c plots g_d against the number of training vectors for a 16 qubit system.

2. $\alpha = 2.0$, $\beta = 1.0$ uncorrelated data

First, for $\alpha = 2.0$ and $\beta = 1.0$, we perform a grid search with our IQP feature map for $10^{-1.5} \leq \lambda \leq 10^{0.5}$ and number of qubits 4 to 18. In Fig. 3(a), we plot the maximum kernel eigenvalue as a function of the number of qubits for varying λ. The solid lines represent interpolations between the mean values of the maximum kernel eigenvalue. The maximum kernel eigenvalue rapidly decays for fixed λ. For the largest λs we observe the maximum kernel eigenvalue decay to $1/N = 1/500$ indicating a completely flat eigenvalue spectrum. In Fig. 3(b), we plot λ vs. number of qubits for fixed maximum kernel eigenvalues. The points for this plot are extracted from the grid-search interpolations in Fig. 3(a).

Using the interpolation in Fig. 3(b), we hyperparameter tune λ in order to keep γ_{max} approximately constant with increasing number of qubits. In practice, we find that γ_{max} remains within 0.08 and 20% of our target for all kernel matrices studied in this paper. We compare the quantum fidelity kernel matrices to classical fidelity kernel matrices using the feature map in Eq. 10 with the same number of qubits and the same randomly sampled
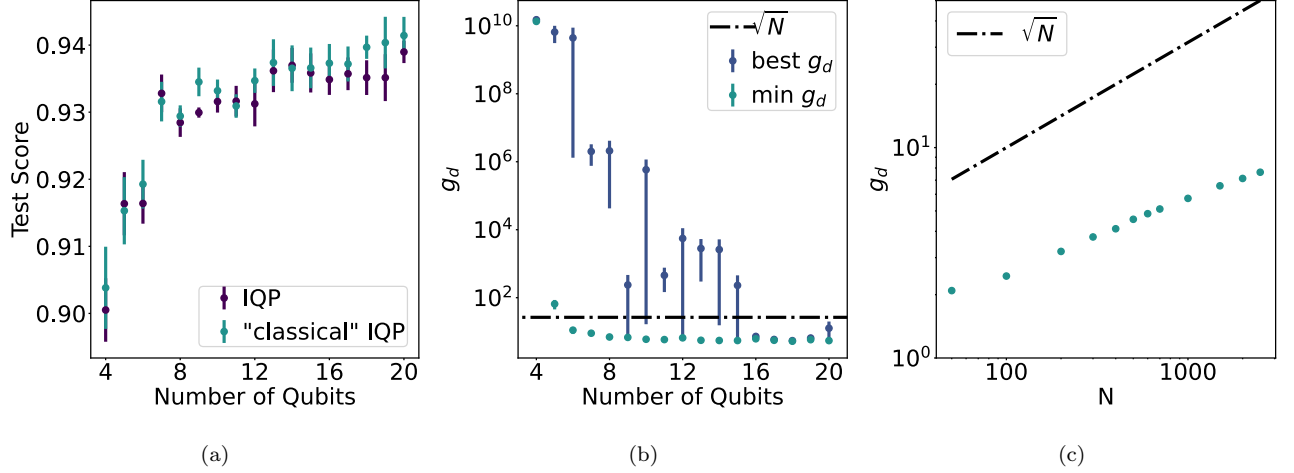
FIG. 2. For the Fashion-Mnist dataset, we plot the test score (a), geometric difference vs. the number of qubits (b) and vs. the number of data points (c) for both the quantum and 'classical' IQP style feature map with $\alpha = 2.0$. In (a), plot the test scores of the two top performing models. In (b), plot the g_d between the two high test score kernels as well as the minimum g_d found between the highest test score quantum kernel and a grid search over classical kernels with varying λ. In (c), we plot g_d vs. the number of data points for 16 qubit feature maps. The feature maps use the λs that minimize g_d for N=700.

data points. For the classical fidelity kernels, we perform a grid search over $10^{-1.5} \leq \lambda \leq 10^{0.5}$ in an attempt to minimize the distance between the classical fidelity kernel matrix and our quantum fidelity kernel matrix. In Fig. 3(c), we plot the results of the grid searches. We plot the minimum g_d found vs. number of qubits for $\gamma_{max} \in \{0.01, 0.05, 0.2, 0.3, 0.4, 0.5\}$. For all γ_{max}, we observe a rapid decay in g_d with increasing number of qubits to well below $\sqrt{N}$ indicating that there can be no quantum advantage for this data point distribution and feature map combination. This is further supported by the observation that the geometric difference grows slower than $\sqrt{N}$. Specifically, in Fig. 3(d), we repeat our analysis using $\gamma_{max} = 0.3$ and a 12 qubit system while varying the number of data points in our distribution (and thus the kernel matrix size) N. The geometric difference g_d remains significantly below the $\sqrt{N}$ line and grows at a slower trend, indicating that no quantum advantage can exist.

3. $\alpha \in \{0.5, 1.0, 2.0\}$, $\beta \in \{0.1, 1.0, 2.0\}$ uncorrelated data

In Sec. IV A 2, we presented numerical results for the IQP style feature map with $\alpha = 2.0$ and data distribution with $\beta = 1.0$. In this section, we will expand our results to include IQP style feature maps with $\alpha \in \{0.5, 1.0, 2.0\}$ as well as uncorrelated data distributions with $\beta \in \{0.1, 1.0, 2.0\}$. For each α and β pair, we repeat the analysis done in Fig. 3 in order to extract the appropriate λs to tune our IQP style feature map and keep γ_{max} fixed. In Fig. 4(a), We plot the minimum g_d found vs. number of qubits for $\beta = 1.0$, $\gamma_{max} = 0.3$ and $\alpha \in \{0.5, 1.0, 2.0\}$. All three IQP style feature maps exhibit the same qualitative behavior where

g_d rapidly decays with increasing number of qubits to below the $\sqrt{N}$ line. In Fig. 4(b), We plot the minimum g_d found vs. number of qubits for $\alpha = 2.0$, $\gamma_{max} = 0.3$ and $\beta \in \{0.1, 1.0, 2.0\}$. Again, g_d rapidly decays to below the $\sqrt{N}$ line with increasing number of qubits. While the higher kurtosis distribution with $\beta = 0.1$ initially starts with much higher g_d, it rapidly decays to below $\sqrt{N}$. In both Fig. 4(a) and Fig. 4(b), we observe a rapid decay in g_d with increasing number of qubits to well below $\sqrt{N}$ indicating that the potential for a quantum advantage is diminishing with system size for this data point distribution and feature map combinations.

4. $\alpha = 2$, $\beta = 2$ correlated data

In addition to uncorrelated data distributions, we present numerical results for correlated data distributions. We use data points sampled from the correlated distribution given by Eq. 13 with $\beta = 2.0$ and varying r. For our quantum kernel, we embed the data points in the IQP feature map with $\alpha = 2.0$ while for our classical kernel we use the 'classical' IQP feature map with $\alpha = 2.0$. In Fig 4(c)(d), we plot the minimum g_d found between the quantum kernel and the classical kernel vs. number of qubits for $r \in \{0.0, 0.33, 0.66, 0.9, 0.95\}$ and for $\gamma_{max} = 0.1$ and $\gamma_{max} = 0.4$. For both γ_{max}s and all r, the minimum g_d starts below $\sqrt{N}$ and only decays further indicating that the possibility of an advantage for the quantum kernel decreases with an increasing number of qubits.
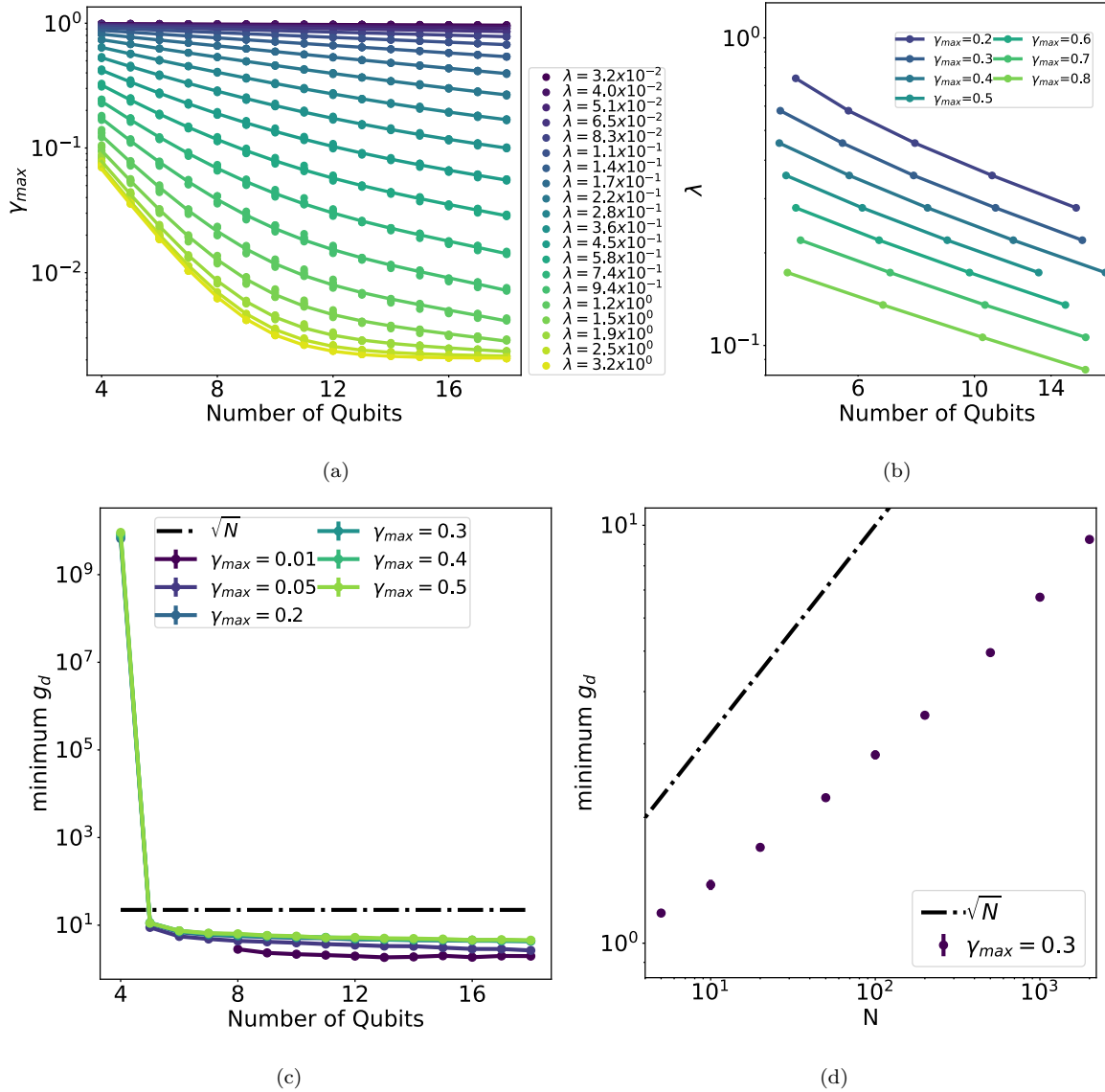
(a)

(b)

(c)

(d)

FIG. 3. Plots for $\alpha = 2.0$ and $\beta = 1.0$ using the IQP style feature map. (a) Maximum kernel matrix eigenvalue vs. number of qubits for varying λ. (b) λ vs number of qubits for a target γ_{max}. (c) Minimum geometric difference, g_d, vs. number of qubits for varying γ_{max}. (d) Minimum geometric difference, g_d, vs. number of data points, N, in distribution for $\gamma_{max} = 0.3$ and a 12 qubit feature map. A geometric difference of $O(\sqrt{N})$ (the black dashed line) is necessary for any potential quantum advantage to exist.

B. Heisenberg Feature Map

In this section, we present numerical results for the Heisenberg feature map (Eq. 8) and the classical fidelity kernel generated by the feature map in (Eq. 11). All kernel matrices use N=500 randomly sampled data points. We perform the same analysis as described in Sec. IV A 2 and shown in Fig. 3 for the Heisenberg feature map with $n_l = 4$. In Fig. 5, we plot the minimum g_d vs. number of qubits for uncorrelated data distributions with $\beta = 1.0$. We observe the same qualitative behavior as the IQP feature maps. The minimum g_d found between quantum and the classical kernel decays rapidly with an increasing

number of qubits to below $O(\sqrt{N})$.

V. LIMITATIONS

The central observation we make is that numerically, the geometric difference between "well-behaved" quantum models and classical ones is small and goes down with the number of qubits. Here "well-behaved" is operationalized as "able to generalize with a fixed number of training points". There are three main potential limitations of the current study that we now discuss.

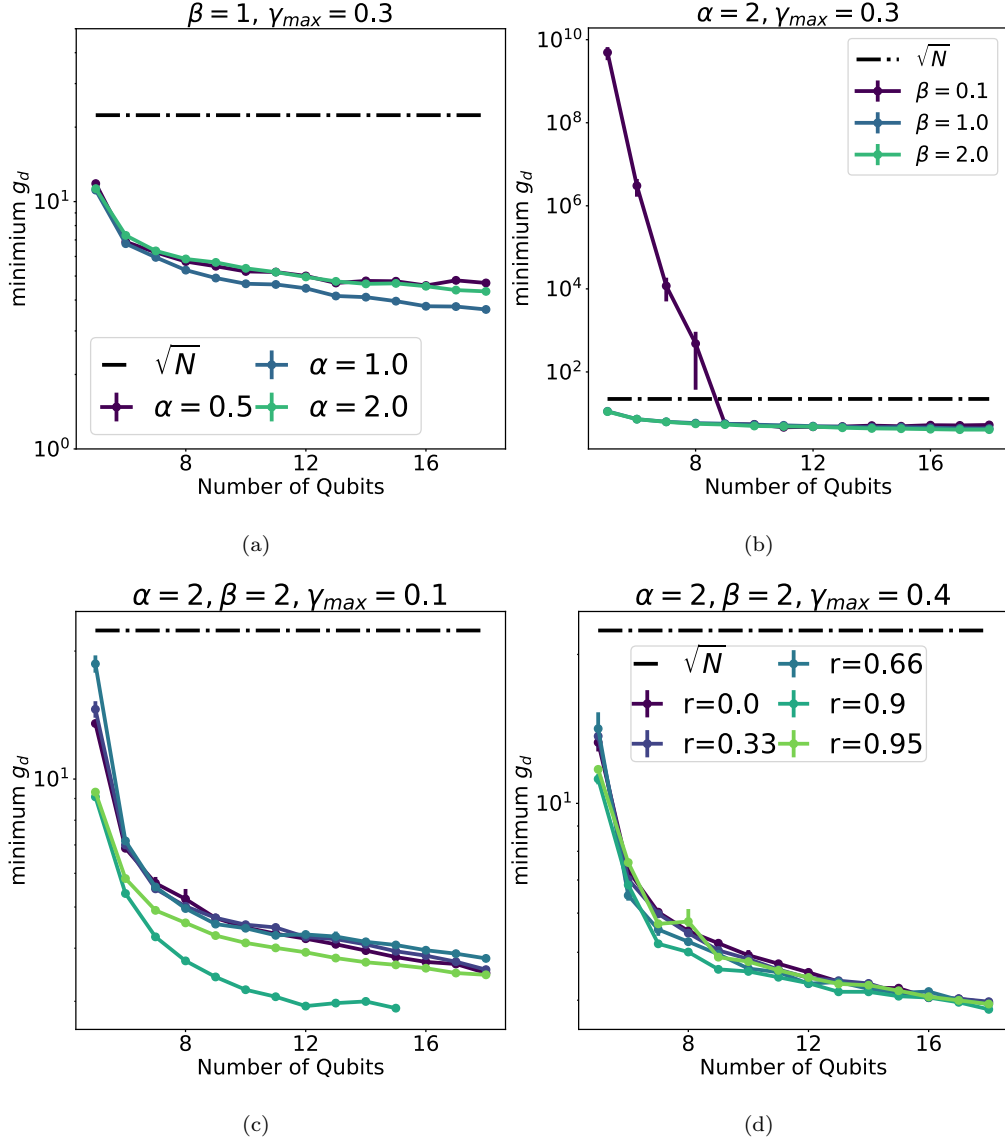First, the condition that the number of training points

FIG. 4. Plots of the minimum geometric difference vs. number of qubits using the IQP style feature map. We show results for different data distributions and values of α. In (a) and (b), we show results for uncorrelated data distributions while in (c) and (d) we show results for correlated data distributions. (a) Minimum geometric difference, g_d, vs. number of qubits for $\beta = 1.0$, $\gamma_{max} = 0.3$ and varying α. (b) Minimum geometric difference, g_d, vs. number of qubits for $\alpha = 2.0$, $\gamma_{max} = 0.3$ and varying β. Minimum geometric difference, g_d, vs. number of qubits for $\alpha = 2.0$, $\beta = 2.0$ and (c) $\gamma_{max} = 0.1$ ((d) $\gamma_{max} = 0.4$) for varying correlation coefficient r.

(or, equivalently, $\gamma_{\max}$) stays constant with the number of qubits may be overly restrictive. Specifically, efficient generalization is possible as long as $\gamma_{\max}$ does not decay faster than polynomially with the number of qubits. Therefore, there may be a possibility of quantum advantage with kernels where $\gamma_{\max}$ decays polynomially and geometric difference stays large. However, this regime is difficult to probe due to only a small number of qubits accessible in simulation, which is why it is outside of the scope of the current study.

Second, we make a specific choice of technique to improve the generalization (namely, bandwidth tuning pro-

posed in [21, 25]). While no other technique has been demonstrated to enable generalization of quantum kernel methods on a broad range of problems, it is possible that such technique will be proposed in the future.

Third, we only consider a few of the best-performing quantum feature maps proposed in the literature. While we expect similar results to hold for any sufficiently expressive quantum feature maps, the decay of geometric difference need not hold in general for a carefully designed feature map. In particular, we do not expect to see a decay in geometric difference for the feature map based on Shor's algorithm that was proposed in [11] for a learning
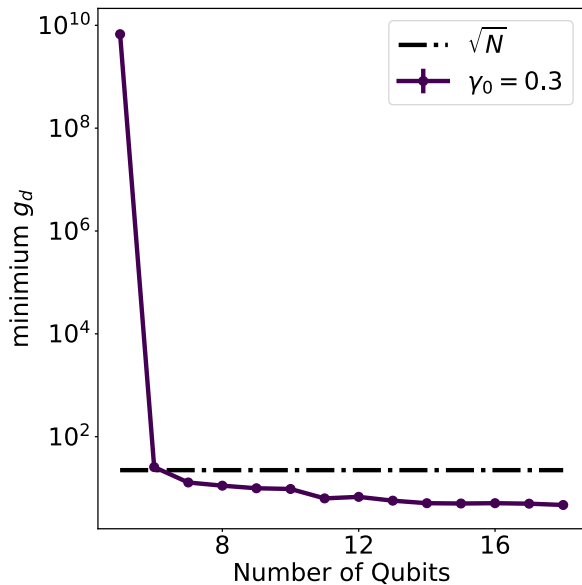
(fidelity) kernels are unlikely to provide advantage

FIG. 5. Plot of the minimum g_d vs. number of qubits for a the Heisenberg feature map (Eq. 8) with $n_l = 4$ and an uncorrelated data distribution with $\beta = 1.0$.

version of the discrete logarithm problem. Whether or not such quantum feature maps exist for a broader range of machine learning problems remains an open question.

Finally, our results for both synthetic and real data rely on the data coming from a well-behaved distribution, such as the distribution of pixel values in images. This condition holds for the diverse real-world datasets considered, and we expect it to hold for more general common datasets. For classical datasets with specific exotic properties that align well with the quantum model, quantum advantage is known to be possible [11, 12, 45]. Additionally, quantum advantage is possible on classical data corresponding to parameters of a quantum model [46]. However, the quantum advantage in this case relies explicitly on the ability of quantum computers to perform quantum simulation efficiently. Therefore we do consider this to be another example of classical data with specific structure enabling quantum advantage.

VI. DISCUSSION

In this work we provide numerical evidence that quantum (fidelity) kernels are unlikely to provide advantage over classical machine learning models. Unlike [22], we consider explicitly quantum kernels that perform well on classical data. Nonetheless our conclusion is similar to that of [22]. Without further algorithmic advances, quantum advantage on classical data is unlikely.

The central challenge is the need to encode inductive bias into the quantum model to prevent decay of the largest eigenvalue and "flattening" of the spectrum. While considering parameterized quantum circuits (also called "Quantum Neural Networks" or QNNs) may provide some benefit by using the training method to introduce inductive bias. For example, there exist cases where QNNs require exponentially fewer datapoints that the analogous kernel method [47]. However, QNNs suffer from trainability issues such as barren plateaus [48] and proliferation of low-quality local minima [49]. Moreover, in general it is not clear if it's easier to introduce structure into QNNs than into quantum feature maps. It may be possible to leverage group structure of the data [35, 36] to overcome this challenge, though applying it to a broad range of classical problems appears challenging. Without such bias, general-purpose techniques like bandwidth tuning [21, 25] or projected kernels [10, 22] must be used, which do not appear to provide a path to quantum advantage on classical data.

ACKNOWLEDGEMENTS

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, 2011).

[2] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, Nature Computational Science 2022 2:9 **2**, 567 (2022).

[3] T. Xiao, J. Huang, H. Li, J. Fan, and G. Zeng, npj Quantum Information **8** (2022), 10.1038/s41534-022-00649-6.

[4] T. Xiao, J. Fan, and G. Zeng, npj Quantum Information **8** (2022), 10.1038/s41534-021-00513-z.

[5] H. Xu, J. Li, L. Liu, Y. Wang, H. Yuan, and X. Wang, npj Quantum Information **5** (2019), 10.1038/s41534-019-0198-z.

[6] M. Pistoia, S. F. Ahmad, A. Ajagekar, A. Buts, S. Chakrabarti, D. Herman, S. Hu, A. Jena, P. Minssen, P. Niroula, A. Rattew, Y. Sun, and R. Yalovetzky, in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)* (IEEE, 2021).

[7] D. Herman, C. Googin, X. Liu, A. Galda, I. Safro, Y. Sun, M. Pistoia, and Y. Alexeev, arXiv:2201.02773 (2022).

[8] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature **549**, 195 (2017).

[9] H. Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, and J. R. McClean, Science **376**, 1182 (2022), arXiv:2112.00778.

[10] J. M. Kübler, S. Buchholz, and B. Schölkopf, (2021), arXiv:2106.03747v2.

[11] Y. Liu, S. Arunachalam, and K. Temme, Nature Physics 2021 17:9 **17**, 1013 (2021), arXiv:2010.02174.

[12] A. Schmidhuber and S. Lloyd, arXiv:2209.14286 (2022).

[13] M. Schuld and N. Killoran, Phys. Rev. Lett. **122**, 040504 (2019), arXiv:1803.07128.

[14] D. K. Park, C. Blank, and F. Petruccione, Physics Letters A **384**, 126422 (2020), arXiv:2004.03489.

[15] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Nature 2019 567:7747 **567**, 209 (2019), arXiv:1804.11326.

[16] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, *On the Expressive Power of Deep Neural Networks*, Tech. Rep. (2017) arXiv:1606.05336v6.

[17] M. Schuld, R. Sweke, and J. J. Meyer, (2021), arXiv:2008.08605v2.

[18] C. Blank, D. K. Park, J. K. K. Rhee, and F. Petruccione, npj Quantum Information 2020 6:1 **6**, 1 (2020), arXiv:1909.02611.

[19] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, (2020), arXiv:2011.00027v1.

[20] L. Banchi, J. Pereira, and S. Pirandola, PRX QUANTUM **2**, 40321 (2021).

[21] A. Canatar, E. Peters, C. Pehlevan, S. M. Wild, and R. Shaydulin, arXiv:2206.06686v2.

[22] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. Mcclean, Nature (2021), 10.1038/s41467-021-22539-9.

[23] A. E. Paine, V. E. Elfving, and O. Kyriienko, (2022), arXiv:2203.08884v1.

[24] E. Peters and M. Schuld, (2022), arXiv:2209.05523v1.

[25] R. Shaydulin and S. M. Wild, arXiv:2111.05451v3.

[26] N. Shirai, K. Kubo, K. Mitarai, and K. Fujii, arXiv:2111.02951v1.

[27] H. Y. Huang, R. Kueng, and J. Preskill, Nature Physics 2020 16:10 **16**, 1050 (2020), arXiv:2002.08953.

[28] A. Canatar, B. Bordelon, and C. Pehlevan, Nature Communications **12**, 1 (2021).

[29] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Nature Communications 2021 12:1 **12**, 1 (2021), arXiv:2001.00550.

[30] M. Cerezo and P. J. Coles, Quantum Science and Technology **6**, 035006 (2021), arXiv:2008.07454.

[31] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Nature Communications **13** (2022), 10.1038/s41467-022-32550-3.

[32] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Physical Review Letters **117** (2015), 10.1103/PhysRevLett.117.080501, arXiv:1504.07999v2.

[33] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Quantum Science and Technology **7**, 045015 (2022), arXiv:2104.05868.

[34] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, 10.22331/q-2021-10-05-558, arXiv:2011.12245v2.

[35] M. Larocca, F. Sauvage, F. M. Sbahi, G. Verdon, P. J. Coles, and M. Cerezo, PRX Quantum **3** (2022), 10.1103/prxquantum.3.030341.

[36] F. Sauvage, M. Larocca, P. J. Coles, and M. Cerezo, "Building spatial symmetries into parameterized quantum circuits for faster training," (2022), arXiv:2207.14413.

[37] E. Farhi and H. Neven, arXiv:1802.06002 (2018), 10.48550/arXiv.1802.06002.

[38] M. Schuld, arXiv:2101.11020 (2021), 10.48550/arXiv.2101.11020.

[39] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, 2002).

[40] H. S. Seung, H. Sompolinsky, and N. Tishby, Physical Review A **45**, 6056 (1992).

[41] R. Dietrich, M. Opper, and H. Sompolinsky, Physical Review Letters **82**, 2975 (1999).

[42] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).

[43] M. Advani, S. Lahiri, and S. Ganguli, Journal of Statistical Mechanics: Theory and Experiment **2013**, P03014 (2013).

[44] V. N. Vapnik, *The Nature of Statistical Learning* (Springer, 1996).

[45] D. W. Berry, Y. Su, C. Gyurik, R. King, J. Basso, A. D. T. Barba, A. Rajput, N. Wiebe, V. Dunjko, and R. Babbush, (2022), 10.48550/ARXIV.2209.13581.

[46] Y. Wu, B. Wu, J. Wang, and X. Yuan, "Provable advantage in quantum phase learning via quantum kernel alphatron," (2021).

[47] S. Jerbi, L. J. Fiderer, H. P. Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Nature Communications **14** (2023), 10.1038/s41467-023-36159-y.

[48] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Nature Communications 2018 9:1 **9**, 1 (2018), arXiv:1803.11173.

[49] X. You and X. Wu, in *International Conference on Machine Learning* (PMLR, 2021) pp. 12144–12155.

DISCLAIMER