Variational Embeddings in Quantum Machine Learning for Classification Problems

Narges Alavi Samani,[1] Mudassir Moosa,[2] Syed Raza,[3] and Aroosa Ijaz[4, 5]

[1] *Department of Computer Science, Université Paris Diderot, Paris, France*
[2] *Department of Physics, Cornell University, Ithaca, NY, USA*
[3] *Department of Physics, University of Virginia, Charlottesville, VA, USA*
[4] *Vector Institute for Artificial Intelligence, Toronto, ON, Canada*
[5] *University of Waterloo, Ontario, Canada*

(Dated: January 23, 2021)

Placeholder for abstract. To learn and compare variational embeddings that can optimally classify data with two classes

CONTENTS

I. INTRODUCTION

Summarize the goal of the project and the initial results. Discuss progress, open problems and future directions

A brief overview of what we studied during the program

Problem of classification of data points using quantum computation have recently been studied in detail [1? ?]. In earlier works [? ?], the data was first embedded in a Hilbert space by a fixed quantum circuit and then the embedded state was passed through a trainable variational circuit that performs a measurement to classify the data. The recent work [1], on the other hand, trained the embedding circuit instead. The goal of training, in this case, is to find an embedding circuit that maximally separate the embedded data in a Hilbert space. Once the embedded data is sufficiently separated, the data can be classified by measuring the fidelity between the embedded states.

We now briefly review the proposal of [1]. Suppose we are given a data set of in which each point is either from class A or class B. We embed the data point x using a variational embedding circuit to a state $|x;\theta\rangle = U(x;\theta)\,|0\rangle$, where θ collectively denotes all the variational parameters. A uniform ensembles of embedded data points from class A and from class B can be represented by density matrices $\rho_A(\theta) = \frac{1}{M_A}\sum_{a\in A}|a;\theta\rangle\langle a;\theta|$ and $\rho_B(\theta) = \frac{1}{M_B}\sum_{b\in B}|b;\theta\rangle\langle b;\theta|$ respectively. It was proposed that the values of parameters θ for which the embedded states are maximally separated should be found by optimizing the Hilbert-Schmidt cost function, $C_{HS} \equiv 1 - \frac{1}{2}\mathrm{tr}(\rho_A - \rho_B)^2$. Once these optimal parameters are found, we can classify the data point x to class A (class B) if the fidelity $\langle x|\,\rho_A - \rho_B\,|x\rangle$

II. RISK FUNCTION FOR VARIATIONAL EMBEDDING CIRCUITS

It was proposed in [1] that one should train the circuit that embeds the data into a Hilbert space. The goal of the training is to find a set of variational parameters for which the data from different classes are maximally separated in a Hilbert space. It was further proposed that if the fidelity is used to classify the data points, then these parameters should be found by minimizing the Hilbert-Schmidt cost function between ensembles of embedded states of different classes.

Our goal in this section is to test this proposal by applying it on an exactly solvable toy problem. This problem is simple enough that we can analytically determine for what value of the variational parameter the embedded data is separable in the Hilbert space. We will then show that the optimization of the Hilbert-Schmidt cost function does not converge to that value of variational parameter, but the optimization of the *empirical risk* function does.

A. Toy problem

In this toy problem, we consider an engineered set of 2-dimensional points (x_1, x_2) where $-L \leq x_{1,2} \leq L$ and we restrict to $L < \pi/2$ (see Fig. (1)). We assign these points to different classes depending of whether $x_1 x_2 > 0$ (blue dots in Fig. (1)) or $x_1 x_2 < 0$ (red dots in Fig. (1)). The data point (x_1, x_2) is then embedded on a single qubit state $|x_1, x_2; \theta\rangle$, where

$$|x_1, x_2; \theta\rangle \;=\; RX(x_2) RY(\theta) RX(x_1)\;|0\rangle\;, \qquad (2.1)$$

and θ is the only variational parameter in the problem. This circuit is simple enough that we can study analytically where each data point (x_1, x_2) is getting mapped on a Bloch sphere as a function of θ. To do this, we first define $\rho(x_1, x_2; \theta) \equiv |x_1, x_2; \theta\rangle \langle x_1, x_2; \theta|$ which we write as

$$\rho(x_1, x_2; \theta) \;=\; \frac{1}{2}\Big(\mathbf{1} \,+\, \vec{n}(x_1, x_2; \theta)\cdot\vec{\sigma}\Big)\,. \qquad (2.2)$$

The Pauli vector $\vec{n}(x_1, x_2; \theta)$ is given by

$$\vec{n}(x_1, x_2; \theta) \;=\; \langle x_1, x_2; \theta|\;\vec{\sigma}\;|x_1, x_2; \theta\rangle\;, \qquad (2.3)$$

and can be evaluated component wise to get

$$\begin{aligned} n_z(x_1, x_2; \theta) \;&=\; \cos(x_2)\cos(x_1)\cos(\theta) \;-\; \sin(x_2)\sin(x_1)\,, \\ n_y(x_1, x_2; \theta) \;&=\; -\sin(x_2)\cos(x_1)\cos(\theta) \;-\; \cos(x_2)\sin(x_1)\,, \\ n_x(x_1, x_2; \theta) \;&=\; \cos(x_1)\sin(\theta)\,. \qquad (2.4) \end{aligned}$$

Specializing to $\theta = \pi/2$, we find that $n_z = -\sin(x_1)\sin(x_2)$, $n_y = -\cos(x_2)\sin(x_1)$, and $n_x = \cos(x_1)$. This implies that all the points with $x_1 x_2 > 0$ (i.e. blue points in Fig. (1)) maps on a Bloch sphere below the 'equator' (i.e. $z = 0$ plane) where all the points with $x_1 x_2 < 0$ (red points) map above the equator. Therefore, the embedded data in this case is separable by $z = 0$ plane.

Let us also consider the case of $\theta = 0$, the significance of which will be apparent shortly. In this case, $n_z = \cos(x_1 + x_2)$, $n_y = -\sin(x_1 + x_2)$, and $n_x = 0$. This implies that all the data points with the same value of $x_1 + x_2$ are mapped to the same point of a Bloch sphere. It can be easily from Fig. (1) that some of the blue and red dots lie on a contour of constant $x_1 + x_2$. We, therefore, conclude that the embedding with $\theta = 0$ does not separate all of the data points. It in fact makes some of the data point less distinguishable by mapping them to the same state.

This suggests that the embedding circuit with $\theta = \pi/2$ is much better than the embedding circuit with $\theta = 0$. To see if this is consistent with the proposal of [1], we numerically calculate the Hilbert Schmidt cost function between ensembles of embedded states as a function of θ. We found that the minima of the Hilbert-Schmidt cost function is at $\theta = 0$; see Fig. (1).

This raises an interesting question: Is there a cost function that is minimum at $\theta = \pi/2$? We propose a possible answer to this question in the following subsection.

B. Empirical risk

Suppose we are given M data points $\{x_i\}$ with their corresponding labels $\{y_i = \pm 1\}$ for $i = 1, 2, ..., M$ and our task is to find the optimal *indicator function* $f(x)$ (i.e. functions that take only two values from $\{-1, 1\}$) to predict the label of a point x. According to Vapnik [2], we should find $f(x)$ which minimizes the empirical risk $R[f] \equiv \frac{1}{M}\sum_i L\big(y_i\,;\,f(x_i)\big)$, where $L\big(y_i\,;\,f(x_i)\big) \equiv \frac{1}{2}\big(1 - y_i f(x_i)\big)$ is a *loss* function and is defined such that $L = 0$ $(L = 1)$ when $f(x_i) = y_i$ $(f(x_i) \neq y_i)$. With this, the empirical risk or the probability of error becomes

$$R[f] \;=\; \frac{1}{2} \,-\, \frac{1}{2M}\sum_i y_i\, f(x_i)\,. \qquad (2.5)$$

If we take $f(x)$ to be the fidelity function $\langle x|\,\rho_A - \rho_B\,|x\rangle$ where ρ_A (ρ_B) denotes the ensemble of embedded states of of data points with label $y = 1$ $(y = -1)$, then the risk function in Eq. (2.5) reduces to the Hilbert-Schmidt cost function [1]. However, note that $f(x) = \langle x|\,\rho_A - \rho_B\,|x\rangle$ is not an indicator or classification function as it takes continuous values.

Note that to predict the label of a data point x, the actual value of the fidelity is not important. Rather, only the sign of the fidelity function is important. Therefore, the classification function that we use is $f(x) = \text{sign}\big(\,\langle x|\,\rho_A - \rho_B\,|x\rangle\,\big)$. The empirical risk for this classification function is then given by

$$R[f] = \frac{1}{2} - \frac{1}{2M}\sum_i y_i\;\text{sign}\big(\,\langle x_i|\,\rho_A - \rho_B\,|x_i\rangle\,\big)\,, \qquad (2.6)$$

and this is no longer proportional to the Hilbert-Schmidt cost function.

We numerically compute this risk function as a function of θ for the toy problem discussed above. The result is shown in Fig. (1). We found that this risk function has a minima around $\theta = \pi/2$ which is precisely the value of θ for which the embedded data is separable by the equator. This provides a possible answer to the question we raised at the end of the last subsection.

Given the success of the risk function in Eq. (2.6) in picking out the embedding which separates the embedded data (i.e. $\theta = \pi/2$) and given the general theory of statistical learning [2], we speculate that the risk function may be a better cost function than the Hilbert-Schmidt cost function to train the embedding circuit. However, the analysis of the toy problem is not enough to reach such a conclusion. A detailed comparison between these cost functions for more complex data sets and for embedding circuits more complicated than the one in Eq. (2.1) is needed. We plan on doing a detailed analysis on these cost functions as part of a future project.

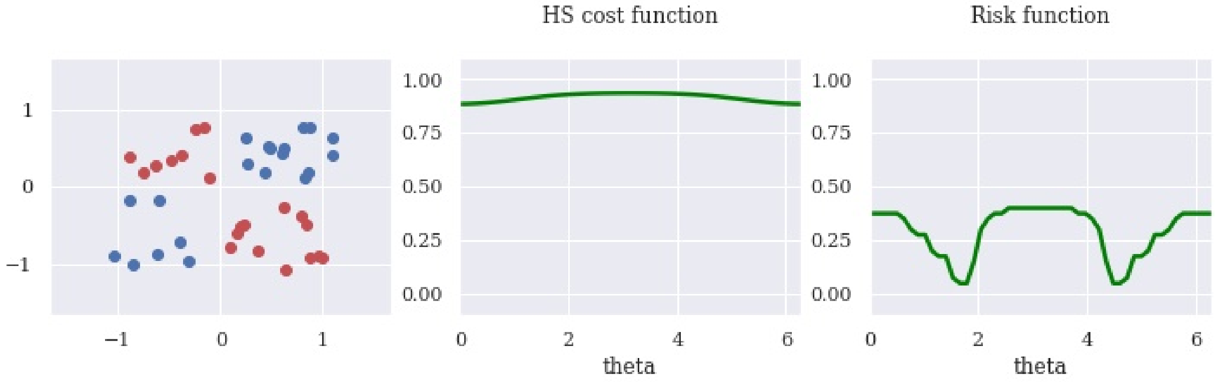

FIG. 1. *(left)* Randomly generated 2-dimensional data set for the toy problem in Sec. (II A). The Hilbert-Schmidt (HS) cost function *(center)* and the risk function *(right)* as a function of $0 \leq \theta \leq 2\pi$.

III. OVERLAP VS HILBERT-SCHMIDT COST FUNCTION

The optimization of the Hilbert-Schmidt cost function is a computationally expensive task. This is because the computation of the gradient of the Hilbert-Schmidt cost between two density matrices involves the computation of the gradients of their purities and the gradient of their overlap. In this subsection, we argue that for a single wire circuit, optimizing the overlap between two density matrices automatically optimizes the Hilbert-Schmidt cost between them. Hence, we propose that for a single wire circuit, it may be more efficient to optimize the overlap instead of the Hilbert-Schmidt cost function. This is a useful result since optimizing the overlap takes around one-third of the time it takes to optimize the Hilbert-Schmidt cost function.

Our argument is based on the following theorem:

Theorem 1 *Consider a 2-dimensional Hilbert space and suppose two density matrices ρ_A and ρ_B are such that their overlap is small: $tr(\rho_A\rho_B) = \epsilon$ where $\epsilon \ll 1$. Then these density matrices are almost pure, i.e. $tr(\rho_A^2) \sim tr(\rho_B^2) = 1 - O(\epsilon)$. Moreover the HS cost between these matrices satisfies $\epsilon \leq C_{HS}(\rho_A, \rho_B) \leq 2\,\epsilon$.*

We relegate the proof of this theorem to Appendix (A). Intuitively, this can be understood as follows. Note that every state of a single qubit can be identified with a point in a Bloch sphere. Points on a Bloch sphere correspond to pure states whereas points inside a Bloch sphere correspond to mixed states. Moreover, if two states are orthogonal, their corresponding points are 'antipodal'. Therefore, if we are given two states which are almost orthogonal (i.e. their overlap is arbitrarily small), then their corresponding points must lie arbitrarily close to the surface of a Bloch sphere. This guarantees that these states are almost pure.

We now continue with our argument. Suppose a variational embedding circuit that we use is such that we can achieve arbitrarily small Hilbert-Schmidt cost value for some choices of parameters. Then for those choices of parameters, the overlap is also arbitrarily small. Hence, we can find the optimal set of parameters by minimizing the overlap by gradient flow method.

We numerically test our proposal by comparing the result of optimizing the overlap with that of optimizing the Hilbert-Schmidt cost function. We used the data set from [1] and found that the results of optimizing the overlap are similar to those of optimizing the Hilbert-Schmidt cost function; see Fig. (2). Moreover, 300 steps of optimizing the overlap took 2.5 minutes whereas the same number of steps of optimizing the HS cost took almost 9.0 minutes.

It is worthwhile to note that the theorem 1 is special for a single qubit and such a statement is not true for Hilbert-spaces of more than 2 dimensions. easiest way to see this is through a counter example. Consider a three dimensional Hilbert space and let $\{|0\rangle, |1\rangle, |2\rangle\}$ be an orthonormal basis. Now take $\rho_A = |0\rangle\langle 0|$ and $\rho_B = \frac{1}{2}|1\rangle\langle 1| + \frac{1}{2}|2\rangle\langle 2|$. Even though there is no overlap between these states, the state ρ_B is not pure.

IV. RANDOM VARIATIONAL EMBEDDINGS

In this section, we consider random variational embedding circuits and compare their efficiency and performance with that of the QAOA circuit studied in [1]. We studied both 1-qubit random circuits and 2-qubits random circuits and we discuss these separately below.

A. 1-qubit random circuits

The 1-qubit QAOA circuit considered in [1] consisted of L layers where each layer was of the form $U^{\text{QAOA}}_{(\ell)} = RX(x)RY(\theta_\ell)$ for $\ell = 1, 2, ..., L$. Here, x is the data point whereas $\{\theta_1, \theta_2, ..., \theta_L\}$ are variational parameters. Following these L layers, there is final layer of $RX(x)$ to ensure that the gradient of the cost function w.r.t θ_L is not zero.

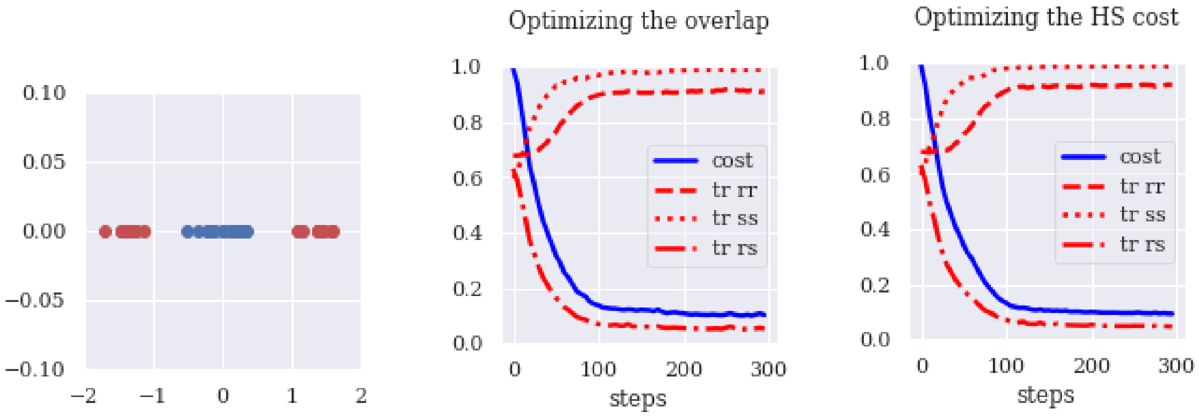

FIG. 2. Data set *(left)* from [1]. The result when we optimize the overlap *(center)* and when we optimize the Hilbert-Schmidt (HS) cost *(right)*.

We examined two approaches of making the above circuit random. In the first approach, we have a circuit with L identical layers as above but the quantum gates in these layers are chosen randomly from the set of $\{RX, RY, RZ\}$ with equal probability. In the second approach, we relax the condition that the layers are identical and hence, quantum gates in each layer are chosen uniformly and independently from $\{RX, RY, RZ\}$. We found that there is a high probability of the existence of 'flat directions' in the cost function for both of these types of random circuits, and hence, they are not as efficient as the QAOA circuit.

In the first approach, there is a $2/3$ probability that the circuit will consist of alternating layers of non-commuting rotation gates. In this case, the circuit is similar to the QAOA circuit. However, there is a $1/3$ probability that all of the gates in the circuit are the same and hence, commute with each other. In this case, the whole circuit can be replaced by a single rotation operator $R\big((L+1)x + \theta_1 + ... + \theta_L\big)$ where R is either RX, RY, or RZ. As a result, the overlap $\langle x'|x\rangle$ between two embedded state is independent of variational parameters: $\langle x'|x\rangle = \langle 0|\, R\big((L+1)(x-x')\big)\,|0\rangle$. Since the Hilbert-Schmidt cost function consists of a weighted sum of overlaps between emdedded states [1], we deduce that the Hilbert-Schmidt cost function is independent of the variational parameters. This implies that this approach of random variational embedding will not work $1/3^{\text{th}}$ of the times because the cost function is 'flat' in every direction. The rest of the times, it will be as good as the QAOA circuit.

The second approach is more interesting as the probability that the cost function is flat in every directions is very small. However, there are still high probabilities of the existence of some flat directions. One such situation is where three adjacent gates in the circuit are the same. For example, suppose that a part of the circuit looks like $\cdots R(\theta_\ell)\, R(x)\, R(\theta_{\ell+1})\, \cdots$, where again R is either RX, RY, or RZ. In this case, the cost function will only be a function of $\theta_+ = \theta_\ell + \theta_{\ell+1}$. Therefore, $\theta_- = \theta_\ell - \theta_{\ell+1}$ is a flat direction.

B. 2-qubits random circuits

The 2-qubit embedding circuit that we considered started with a layer of $RY(\pi/4)$ acting on each of the qubits and ended with a layer of of $RX(x)$ acting on each of the qubits. In between these fixed layers, we have L additional layers of the form $U_{(\ell)}(x;\theta_{2\ell-1},\theta_\ell) = (R_{\ell,1}(x)\otimes R_{\ell,2}(x))\; CZ\; (R_{\ell,3}(\theta_{2\ell-1})\otimes R_{\ell,4}(\theta_{2\ell}))\; CZ\,,$ where $R_{\ell,i}$ are independently and randomly chosen from $\{RX, RY, RZ\}$. Hence, there are $2L$ variational parameters. The model of this circuit is inspired by the circuit from [3].

We restricted our attention to a 1-dimensional data set shown in Fig. (3). Since our goal is to compare the results of our circuit with that of the QAOA circuit studied in [1], we followed [1] and choose the variational parameters to be small before the optimization. We tried different number of layers and different seeds used to randomly choose the quantum gates. In each case, we empirically found that the optimized value of the Hilbert-Schmidt cost function was significantly higher than 0.1, which was the optimized value achieved in [1] with a 4 layer QAOA circuit. We first fix the number of layers of our circuit to 2 so that it has the same number of parameters (four) as a 4 layer QAOA circuit and try different values of seed (i.e. 11, 137, and 92). The result of this analysis is presented in Fig. (3). We then fixed the random seed to 42 and tried different number of layers ($L = 2, 3, 4$). The result of this analysis is presented in Fig. (4).

V. VARIATIONAL EMBEDDINGS AS FOURIER SERIES (RAZA)

Summarize the Fourier expressibility paper. Then the hypothesis on the relation between decision boundaries

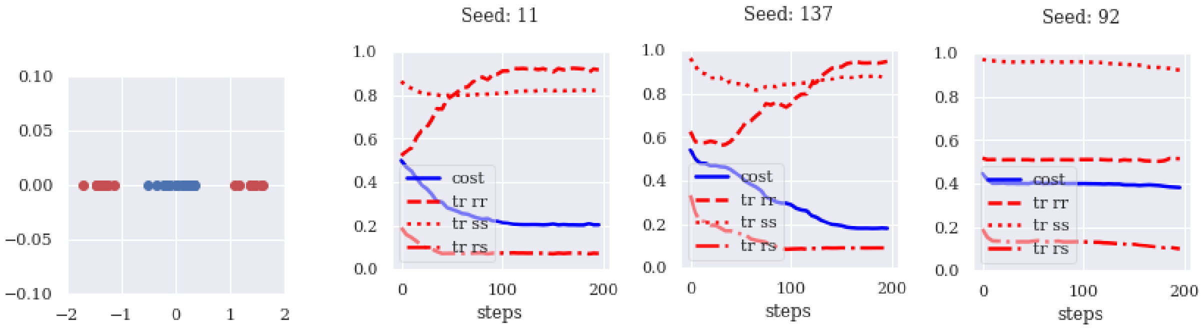

FIG. 3. Data set *(left)* from [1]. The optimization of the Hilbert-Schmidt cost function for a random embedding circuit with 2 layers. Different seeds (11, 137, and 92) were used to randomly chose the rotation gates.

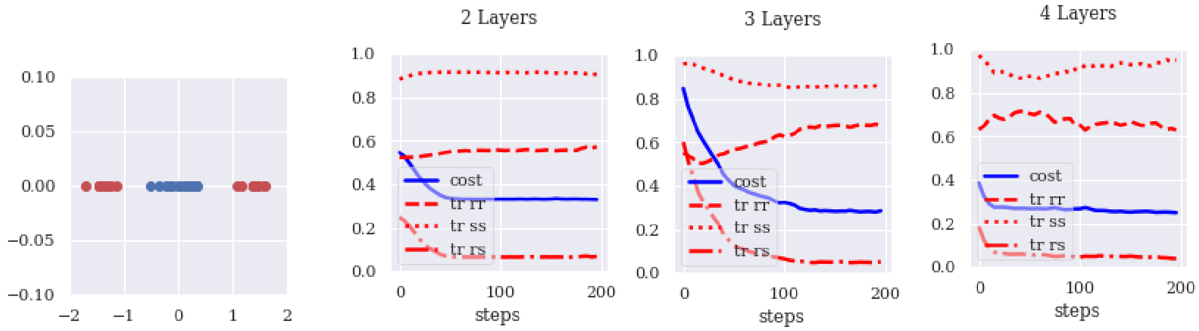

FIG. 4. Data set *(left)* from [1]. The optimization of the Hilbert-Schmidt cost function for a random embedding circuit with 2, 3, and 4 layers. A fix random seed of 42 was used to randomly chose the rotation gates.

expressibility and embedding circuits complexity.

Raza to add IQP results to Narges' chapter

In this section, we will review some existing results on variational embeddings [1] and quantum models as Fourier series [4]. We will then pose an hypothesis; although we have some initial insights but the work is still in its early stages and results will be presented in a future work.

VI. COMPARISON OF VARIATIONAL EMBEDDING CIRCUITS

One of the challenges in implementing variational quantum embedding for classification tasks is to choose an effective circuit that maps classical inputs into "well-separated" quantum states in Hilbert space. To find the best variational circuit embedding for a classification task, we need to find links between the characterization of quantum circuits and their performance in classifying data. There are various descriptors to characterize Parameterized Quantum Circuits(PQC), such as expressivity, entangling capability, connectivity, circuit depth, number of parameters, and effect of barren plateaus.

A. Expressivity and Entangling Capability

To compare different PQC with respect to expressivity and entangling capability, Sim et al. [5] provide different circuit structures, varying connectivity of qubits and selection of gates. They quantify improvements in both expressivity and entangling capabilities gained by sequences of controlled-X rotation gates compared to sequences of controlled-Z rotation gates. A reason for the lower performance of circuits with controlled-Z rotation gates might be the fact that these gates commute with each other resulting in a fewer number of effective circuit parameters.

Additionally, Sim et al. [5] compare different circuits structure with respect to the arrangement of two-qubit gates. These arrangements include near-neighbor, circuit-block, and all-to-all interactions (may need figure). In near-neighbor configurations, two-qubit gates operate in a linear array of qubits. In circuit-block configuration, two-qubit gates are arranged in an array of qubits that form a closed loop. For all-to-all configuration, two-qubit gates are arranged in a fully connected graph of qubits. The results in [5] show that all-to-all configurations give rise to the highest expressivity, although the expressivity of circuit-block configurations is close to the all-to-all ones. Further, both all-to-all and circuit-block configurations have a high entangling capability. On the other hand, near-neighbor configurations have the lowest circuit depth, for the same number of two-qubit gates. Therefore, all-to-all arrangements of two-qubit gates lead to the highest expressivity and entangling capability with the cost of higher circuit depth, number of parameters, connectivity.

B. Simulation Results

In this part, we implement several variational circuit for embedding data in order to classify the data into two classes. The circuit templates are shown in Fig. Although these templates are provided for four qubit circuits, we additionally implement these circuits with lower numbers of qubits. Circuit simulations presented in this work are implemented using the PennyLane software package for hybrid optimization [6].

Circuit designs that are implemented in this study, shown in Fig., are derived or inspired by past works. For example, circuit 1 is QAOA with an extra data mapping at the last layer which is also used as an embedding circuit in [1]. For designing circuit 2, we use the template of circuit 1 and replace rotation-X and rotation-Y gates with each other. Circuits 3, 4, 7, and 8 are among the circuit templates presented in [5] of which expressivity and entangling capability are studied. Circuit 5 and 6 are similar to circuits 1 and 2 except that data is encoded only at the end of each layer.

VII. CONCLUSION AND FUTURE DIRECTIONS

ACKNOWLEDGMENT

This work was supported by the mentorship program of Quantum Open Source Foundation (QOSF).

[1] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran, "Quantum embeddings for machine learning," arXiv e-prints , arXiv:2001.03622 (2020), arXiv:2001.03622 [quant-ph].

[2] V. N. Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks **10**, 988–999 (1999).

[3] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven, "Barren plateaus in quantum neural network training landscapes," Nature Communications **9**, 4812 (2018), arXiv:1803.11173 [quant-ph].

[4] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer, "The effect of data encoding on the expressive power of variational quantum machine learning models," (2020), arXiv:2008.08605 [quant-ph].

[5] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," Advanced Quantum Technologies **2**, 1900070 (2019).

[6] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, M. Sohaib Alam, Shahnawaz Ahmed, Juan Miguel Arrazola, Carsten Blank, Alain Delgado, Soran Jahangiri, Keri McKiernan, Johannes Jakob Meyer, Zeyue Niu, Antal Száva, and Nathan Killoran, "Pennylane: Automatic differentiation of hybrid quantum-classical computations," (2020), arXiv:1811.04968 [quant-ph].

Appendix A: Proof of theorem 1

In this appendix, we present the proof for theorem 1 that we used in Sec. (III). Since ρ_A and ρ_B are 2-dimensional density matrices, we can write them as

$$\rho_A = \frac{1}{2}\big(\mathbf{1}+\mathbf{n}_A\cdot\vec{\sigma}\big) \qquad \rho_B = \frac{1}{2}\big(\mathbf{1}+\mathbf{n}_B\cdot\vec{\sigma}\big) \tag{A1}$$

where $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ is a vector Pauli operator. Note that the overlap between these two density matrices is given by

$$\mathrm{tr}(\rho_A\rho_B) = \frac{1}{2}\big(1+\mathbf{n}_A\cdot\mathbf{n}_B\big)\,. \tag{A2}$$

Therefore, if $\mathrm{tr}(\rho_A\rho_B) = \epsilon$, then $\mathbf{n}_A\cdot\mathbf{n}_B = -1+2\epsilon$, and hence $|\mathbf{n}_A\cdot\mathbf{n}_B| = 1-2\epsilon$. Now using the Cauchy-Schwarz inequality, we get

$$|\mathbf{n}_A|\,|\mathbf{n}_B| \geq 1-2\epsilon\,. \tag{A3}$$

Moreover, note that the purity of ρ_A and that of ρ_B is given by

$$\mathrm{tr}(\rho_A^2) = \frac{1}{2}\big(1+|\mathbf{n}_A|^2\big)\,, \tag{A4}$$

$$\mathrm{tr}(\rho_B^2) = \frac{1}{2}\big(1+|\mathbf{n}_B|^2\big)\,. \tag{A5}$$

Since $\mathrm{tr}(\rho_A^2) \leq 1$ and $\mathrm{tr}(\rho_B^2) \leq 1$, we deduce that $|\mathbf{n}_A| \leq 1$ and $|\mathbf{n}_B| \leq 1$. Combining these conditions with Eq. (A3), we get

$$|\mathbf{n}_A| = 1 - c_A\,\epsilon + O(\epsilon^2)\,, \tag{A6}$$

$$|\mathbf{n}_B| = 1 - c_B\,\epsilon + O(\epsilon^2), \tag{A7}$$

where $c_A \geq 0$, $c_B \geq 0$, and $c_A + c_B \leq 2$. Inserting these results in Eq. (A5), we find that $\mathrm{tr}(\rho_A^2) = 1-c_A\epsilon+O(\epsilon^2)$ and $\mathrm{tr}(\rho_B^2) = 1-c_B\epsilon+O(\epsilon^2)$.

Moreover, the HS cost between ρ_A and ρ_B becomes

$$\begin{aligned} C_{HS}(\rho_A, \rho_B) &= 1 + \mathrm{tr}(\rho_A \rho_B) - \frac{1}{2}\big(\mathrm{tr}(\rho_A^2) + \mathrm{tr}(\rho_B^2)\big)\,, \\ &= \frac{2 + (c_A + c_B)}{2}\,\epsilon + O(\epsilon^2)\,. \end{aligned} \tag{A8}$$

Since $c_A \geq 0$ and $c_B \geq 0$, we deduce that $C_{HS} \geq \epsilon$. Also since $c_A + c_B \leq 2$, we get $C_{HS} \leq 2\epsilon$.

This finishes the proof of theorem 1.