Variational Embeddings in Quantum Machine Learning for Classification Problems

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

[1] *placeholder*
[2] *placeholder2*
[3] *Department of Physics, University of Virginia, Charlottesville VA 22904, USA*
[4] *placeholder3*

(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

II. RISK FUNCTIONAL FOR VARIATIONAL EMBEDDING CIRCUITS (MUDASSIR)

A. Risk

As discussed in Sec. (I), 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 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.

1. *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. (**??**)). We assign these points to different classes depending of whether $x_1 x_2 > 0$ (blue dots in Fig. (**??**)) or $x_1 x_2 < 0$ (red dots). The data point (x_1, x_2) is then embedded on a single qubit state $|x_1, x_2; \theta\rangle$, where

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

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

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

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

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

and can be evaluated component wise to get

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

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. (**??**)) 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. (**??**) 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.

B. 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 the purities (or purity?) of these matrices 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). Here, we instead continue 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. (1). 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.

C. 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. We discuss these separately below.

1. 1-*qubit random circuits*

The 1-qubit QAOA circuit considered in [1] consisted of L layers where each layer was of the form $U_{(\ell)}^{\text{QAOA}} = RX(x)RY(\theta_\ell)$ 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.

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 but independently. We found that there is a high probability 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.

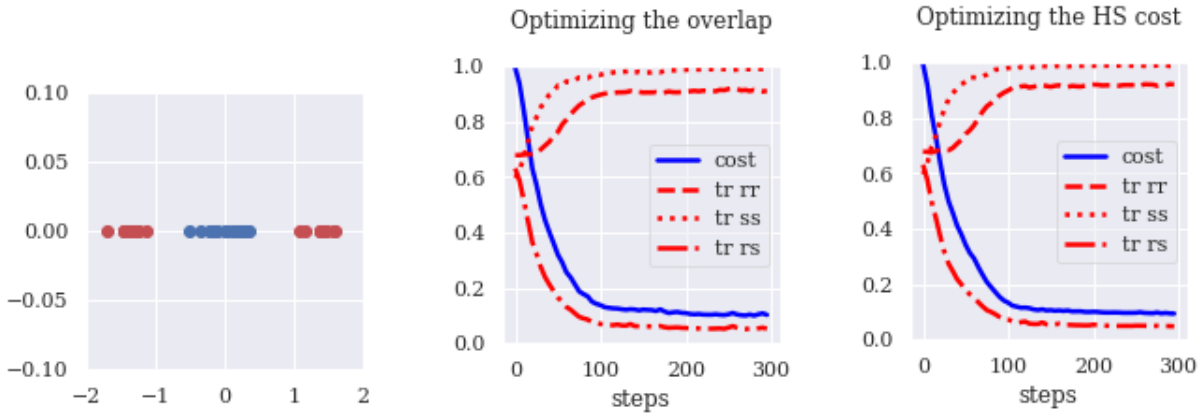

FIG. 1. Data set *(left)* from [1]. The result when we optimize the overlap *(center)* and when we optimize the Hilbert-Schmidt (HS) cost *(right)*.

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 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 $\dots R(\theta_\ell)\, R(x)\, R(\theta_{\ell+1}) \dots$, 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.

2. 2-qubits random circuits

We considered a 2-qubit embedding circuit which started with a layer of $RY(\pi/4)$ acting on both qubits and ended with a layer of of $RX(x)$ acting on each 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 $\{RX, RY, RZ\}$. The model of this circuit is inspired by the circuit from [2].

We restricted our attention to a 1-dimensional data set shown in Fig. (**??**). (Add reference) and maximally separate the embedded data in the Hilbert space by optimizing the Hilbert-Schmidt cost function. (We need a review of these ideas/concepts in the introduction.) 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 initial seeds. In each case, we empirically found that the optimized value of the Hilbert-Schmidt cost was higher that the optimized value achieved in [1] with a QAOA circuit. Add few more sentences and then add figures.

III. QUANTUM MODELS AS FOURIER SERIES (RAZA)

Testing citations [**?**]

Summarize the Fourier expressibility paper. Then the hypothesis on the relation between decision boundaries expressibility and embedding circuits complexity.

Raza to add IQP results to Narges' chapter

IV. COMPARISON OF VARIATIONAL EMBEDDING CIRCUITS (NARGES)

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. [3] 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. [3] 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 [3] 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.

V. CONCLUSION AND FUTURE DIRECTIONS

ACKNOWLEDGMENT

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

[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] 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].

[3] 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).

Appendix A: Proof of theorem 1

In this appendix, we present the proof for theorem 1 that we used in Sec. (II B). 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.