AN INITIALIZATION STRATEGY FOR ADDRESSING BARREN PLATEAUS IN PARAMETRIZED QUANTUM CIRCUITS

TECHNICAL NOTE

Edward Grant *
Rahko Limited &
Department of Computer Science
University College London
edward.grant@rahko.ai

Leonard Wossnig
Rahko Limited &
Department of Computer Science
University College London
leonard.wossnig@rahko.ai

Mateusz Ostaszewski
Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
mostaszewski@iitis.pl

Marcello Benedetti
Cambridge Quantum Computing Limited &
Department of Computer Science
University College London
m.benedetti@cs.ucl.ac.uk

November 28, 2019

ABSTRACT

Parametrized quantum circuits initialized with random initial parameter values are characterized by barren plateaus where the gradient becomes exponentially small in the number of qubits. In this technical note we theoretically motivate and empirically validate an initialization strategy which can resolve the barren plateau problem for practical applications. The technique involves randomly selecting some of the initial parameter values, then choosing the remaining values so that the circuit is a sequence of shallow blocks that each evaluates to the identity. This initialization limits the effective depth of the circuits used to calculate the first parameter update so that they cannot be stuck in a barren plateau at the start of training. In turn, this makes some of the most compact ansätze usable in practice, which was not possible before even for rather basic problems. We show empirically that variational quantum eigensolvers and quantum neural networks initialized using this strategy can be trained using a gradient based method.

***K*eywords** Quantum Neural Networks · Variational Quantum Eigensolvers · Quantum Machine Learning

1 Introduction

Parametrized quantum circuits have recently been shown to suffer from gradients that vanish exponentially towards zero as a function of the number of qubits. This is known as the 'barren plateau' problem and has been demonstrated analytically and numerically [1]. The implication of this result is that for a wide class of circuits, random initialization will cause gradient-based optimization methods to fail. Resolving this issue is critical to the scalability of algorithms such as the variational quantum eigensolver (VQE) [2, 3] and quantum neural networks (QNNs) [4, 5, 6].

In Ref. [7] the author shows that the barren plateau problem is not an issue of the specifically chosen parametrization, but rather extends the result to any direction in the tangent space of a point in the unitary group. In the Appendix we re-derive this result, and show that the gradient of the scalar $\langle 0| U(\alpha)^\dagger H U(\alpha) |0\rangle$ for any Hermitian operator H, vanishes with respect to any direction on the unitary group. Similarly the variance of the gradient decreases exponentially with the number of qubits.

*Corresponding author

Notably, this does not preclude the existence of a parametrization that would allow for efficient gradient-based optimization. Finding such a parametrization seems a non-trivial, but important, task. Indeed, as argued in Ref. [1], the barren plateau problem affects traditional ansätze such as the unitary coupled cluster, when initialized randomly, even for a small number of orbitals. The authors leave it as an open question whether the problem can be solved by employing alternative hardware-efficient ansätze.

In this technical note, we take an alternative route. Instead of proposing a new ansatz, we present a solution based on a specific way of initializing the parameters of the circuit. The strategy resolves the problem using a sequence of shallow unitary blocks that each evaluates to the identity. This limits the effective depth of the circuits used to calculate the gradient at the first iteration and allows us to efficiently train a variety of parametrized quantum circuits.

2 A quick recap of the barren plateau problem

Here we briefly recapitulate the original barren plateau problem and its generalization. Details of the derivation can be found in the Appendix. A parametrized quantum circuit can be described by a sequence of unitary operations

$$U(\boldsymbol{\theta}) = U_L(\theta_L)W_L \cdots U_1(\theta_1)W_1 = \prod_{l=L}^{1} U_l(\theta_l)W_l, \tag{1}$$

where $U_l(\theta_l) = \exp(-i\theta_l V_l)$, θ_l is a real-valued parameter, V_l is an Hermitian operator, and W_l is a fixed unitary. The objective function of a variational problem can be defined as the expectation

$$E(\boldsymbol{\theta}) = \langle 0|U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta})|0\rangle, \tag{2}$$

where H is an Hermitian operator representing the observable of interest. The partial derivatives take the form

$$\partial_{\theta_k} E(\boldsymbol{\theta}) = i \left\langle 0 \left| U_-^\dagger \left[V_k, U_+^\dagger H U_+\right] U_- \right| 0 \right\rangle, \tag{3}$$

where $U_- \equiv \prod_{l=k-1}^{1} U_l(\theta_l)W_l$, and $U_+ \equiv \prod_{l=L}^{k} U_l(\theta_l)W_l$. If either U_- or U_+ matches the Haar distribution [8] up to the second moment, e.g., 2-designs [9], the expected number of samples required to estimate Eq. (3) is exponential in the system size.

In the Appendix we re-derive the result of Ref. [7] for the more general barren plateau problem. Concretely it is shown that the gradient in a direction Z in the tangent space of the unitary group at $U(\alpha)$,

$$\partial_\alpha E(U(\alpha)) = i \langle 0| Z^\dagger H U + U^\dagger H Z |0\rangle, \tag{4}$$

where $U(0) = U$, and $\partial_\alpha U|_{\alpha=0} = Z$, vanishes in expectation, i.e.,

$$\mathbb{E}\left[\partial_\alpha E(U(\alpha))\right] = 0. \tag{5}$$

Noting that the direction always takes the form $Z = -iUM$ for some fixed Hermitian matrix M, it is further shown that the variance

$$\mathrm{Var}\left[\partial_\alpha E(U(\alpha))\right] = 2\frac{(M^2)_{00} - (M_{00})^2}{N^2 - 1}\left(\mathrm{Tr}\left(H^2\right) - \frac{\mathrm{Tr}\left(H\right)^2}{N}\right), \tag{6}$$

with $M_{00} := \langle 0| M |0\rangle$ and $(M^2)_{00} := \langle 0| M^2 |0\rangle$, becomes exponentially small in the number of qubits.

This leads us to believe that the choice of parametrization for quantum circuits is a highly non-trivial task that can determine the success of the variational algorithm. In the next Section, we describe a method which resolves the barren plateau problem for practical purposes. In Section 4 we give numerical evidence on two different use cases.

3 Initializing a circuit as a sequence of blocks of identity operators

Intuitively, the initialization strategy is as follows: we randomly select some of the initial parameter values and choose the remaining values in such a way that the result is a fixed unitary matrix, i.e., a deterministic outcome such as the identity. Additionally, we initially ensure that when taking the gradient with respect to any parameter, most of the circuit evaluates to the identity, which restricts its effective depth. This initialization strategy is optimized in order to obtain a non-zero gradient for most parameters in the first iteration. Obviously, this does not *a priori* guarantee that the algorithm stays far from the barren plateau. However, numerical results indicate that this is indeed the case and that this initialization strategy allows the circuit to be trained efficiently. This gives an immediate advantage compared

to previously known methods, which generally do not allow for training any parameter without an exponential cost incurred through the required accuracy.

Concretely, to ensure that U_- and U_+ do not approach 2-designs, we initialize the circuit via M blocks where each block is of depth L. Depth L is chosen to be sufficiently small so that the blocks are shallow and cannot approach 2-designs. In the following we will consider any fixed gate, i.e., W_l in Eq. (1), as a parametrized one in order to simplify the presentation. For any $m = 1, \ldots, M$, the corresponding block has the form

$$U_m(\boldsymbol{\theta}_m) = \prod_{l=L}^{1} U_l(\theta_{l,1}^m) \prod_{l=1}^{L} U_l(\theta_{l,2}^m). \tag{7}$$

While the initial parameter values for the $U_l(\theta_{l,1})$ can be chosen at random, the parameter values for $U_l(\theta_{l,2})$ are chosen such that $U_l(\theta_{l,2}) = U_l(\theta_{l,1})^\dagger$. Each block, and thus the whole circuit, evaluates to the identity, i.e.,

$$U(\boldsymbol{\theta}^{init}) = \prod_{m=M}^{1} U_m(\boldsymbol{\theta}_m) = \prod_{m=M}^{1} \left(\prod_{l=L}^{1} U_l(\theta_{l,1}^m) \prod_{l=1}^{L} U_l(\theta_{l,1}^m)^\dagger \right) = \prod_{m=M}^{1} I_m = I. \tag{8}$$

It is important to choose each block $U_m(\boldsymbol{\theta}_m)$ to be sufficiently deep to allow entanglement as training progresses. Yet each block should be sufficiently shallow so that $U_m(\boldsymbol{\theta}_m)$, considered in isolation, does not approach a 2-design to the extent that the gradients would become impractically small.

In Ref. [1] it was shown that the sampled variance of the gradient of a two-local Pauli term decreases exponentially as a function of the number of qubits, which also immediately follows from the barren plateau result. Furthermore, the convergence towards this fixed lower value of the variance was numerically shown to be a function of the circuit depth. This implies that for blocks $U_m(\boldsymbol{\theta}_m)$ of constant depth, the whole circuit $U(\boldsymbol{\theta}^{init})$ is not in a barren plateau, allowing us to estimate the gradient efficiently for the initial learning iteration.

The intuition behind the *identity block strategy* is the following: changing a single parameter in one block means that the other blocks still act like the identity. Therefore even if the whole circuit is deep enough to potentially be a 2-design, changing any parameter will yield a shallow circuit. Notably this holds only for the first training parameter update.

We now analyze in more depth the behaviour of the initialization at the level of each block. An interesting property is that the gradients for gates located *away* from the center of the block, e.g., in the beginning or at the end of a block, will have a larger magnitude. The reason is that the circuits required for the estimation are further from being 2-designs since they are more shallow, as shown by the following calculation

$$\partial_{\theta_{k,1}^m} U_m(\boldsymbol{\theta}_m) = \left(\prod_{l=L}^{k+1} U_l(\theta_{l,1}^m) \right) \left(\partial_{\theta_{k,1}^m} U_k(\theta_{k,1}^m) \right) \left(\prod_{l=k-1}^{1} U_l(\theta_{l,1}^m) \right) \left(\prod_{l=1}^{L} U_l(\theta_{l,2}^m) \right) \tag{9}$$

$$= \left(\prod_{l=L}^{k+1} U_l(\theta_{l,1}^m) \right) \left(\partial_{\theta_{k,1}^m} U_k(\theta_{k,1}^m) \right) U_k(\theta_{k,2}^m) \left(\prod_{l=k+1}^{L} U_l(\theta_{l,2}^m) \right) \tag{10}$$

where we used that $U_l(\theta_{l,2}^m) = U_l(\theta_{l,1}^m)^\dagger$. For a small index k we see that the circuit becomes shallow and hence the gradient is expected to be larger. A similar calculation can be done for the gradient with respect to the second set of parameters, i.e., $\partial_{\theta_{k,2}^m} U_m(\boldsymbol{\theta}_m)$. In this case, for index k close to L we see that the circuit becomes shallow.

To summarize, we expect that parameters at the boundaries of the blocks to have gradient larger than those at the center. Notice that the gradient can still be zero if H commutes with the gate, which is a consequence of Eq. (3) and Eq. (4). However, this is generally unlikely, and can be resolved by applying a small deterministic entangling circuit. More concretely, to avoid such cases we can add a shallow entangling layer B to the circuit, i.e., $U_M \cdots U_1 B$, where U_i are the blocks described above. This also resolves the barren plateau problem for the training of variational quantum eigensolvers, which we discuss in more detail in the next Section.

4 Experimental results

4.1 Initializing a parametrized quantum circuit

In this experiment we show the scaling of the variance of the gradient as a function of the number of qubits for both the random initialization and identity block strategy. For the random circuit, we use the same ansatz as in Ref. [1], Fig.

2, and the same ZZ observable. Their ansatz consists of layers of single qubit rotations $\exp(-i\theta_l V_l)$ about randomly selected axes $V_l \in \{X, Y, Z\}$, followed by nearest neighbor controlled-Z gates. We used a total of 120 layers of this kind. For the identity block initialization, we employed a single block as described by Eq. (8) with $M = 1$ and $L = 60$. This setting also accounts for a total of $2LM = 120$ layers. In both cases, initial values for the free parameters were drawn from the uniform distribution $\text{unif}(0, 2\pi)$.

In Fig. 1 (a) we compare the variance of $(\partial_{\theta_{1,1,1}} E)$, i.e., the gradient with respect to the first element of the first part of the first block, as a function of the number of qubits n, when the circuit is applied to the input state $\sqrt{H}^{\otimes n} |0\rangle$. Each point in the Figure was computed from 200 circuits. When using the random initialization, the variance decreased exponentially with the number of qubits, reproducing the plateau behaviour which was described in Ref. [1]. In contrast to this, the variance of the circuit initialized as an identity block was invariant to the system size.

In Fig. 1 (b) we again compare the variance of $(\partial_{\theta_{1,1,1}} E)$ as a function of system size when circuits are applied to random MNIST images, downsampled and normalized such that they constitute valid quantum states. This type of encoding is known as amplitude encoding and represents a realistic scenario for computer vision tasks performed on a quantum computer. Each point in the Figure was computed from 200 circuits. Similar to the previous experiment, the variance of the gradient vanished with the system size when using the random initialization. In contrast, for circuits using the identity block initialization, the variance did not vanish exponentially with the system size, showing that the plateau was avoided at initialization.

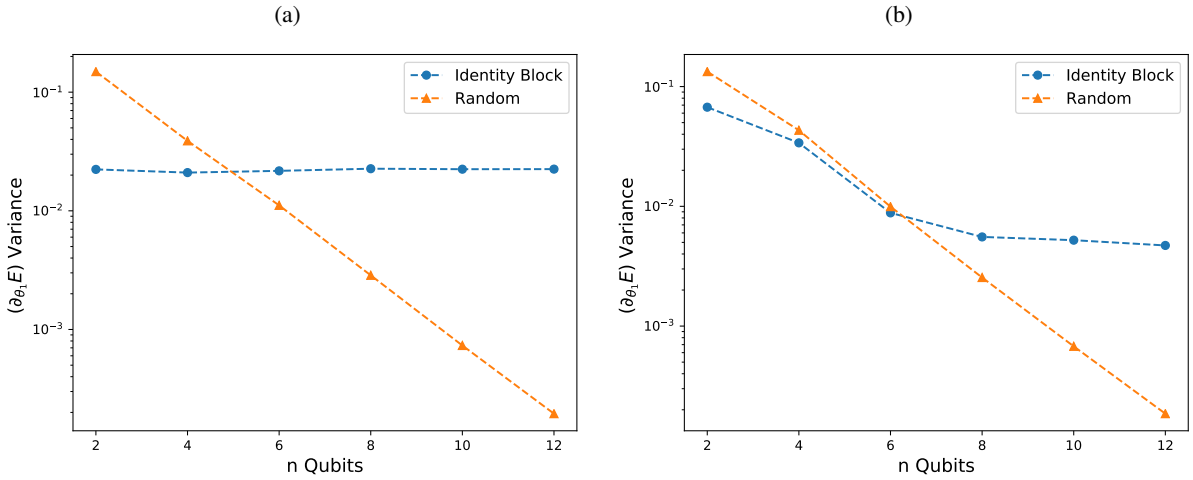

Figure 1: The variance of the gradient of the energy with respect to the first parameter as a function of number of qubits for (a) input state $\sqrt{H}^{\otimes n} |0\rangle$, and (b) input state consisting of amplitude encoded MNIST images. Using the random initialization strategy resulted in the variance of the gradient becoming exponentially small as a function of the number of qubits. The identity block initialization prevented this for both input types (a) and (b).

4.2 Training a quantum neural network classifier

We have shown both analytically and empirically that a circuit with identity block initialization does not suffer from the barren plateau problem in the first training iteration. In this experiment we examine the variance of the gradient during training time to test whether a quantum neural network (QNN) classifier for MNIST images approaches the plateau.

We used a 10-qubit circuit with $M = 2$ identity blocks, each having $L = 33$ layers, for a total of $2LM = 132$ layers. We selected $N = 700$ MNIST images at random, resized them to 32×32, and finally reshaped them into vectors of dimension $2^{10} = 1024$. We normalized each vector to unit length in order to be used as inputs to the circuit. Labels were set to $y_i = 1$ for images of 'even' digits, and $y_i = 0$ for images of 'odd' digits. For each MNIST example ψ_i, classification was performed by executing the circuit, measuring the observable ZZ, and finally ascribing a predicted probability to each class such that $P(even|\psi_i) = \frac{1}{2}(\langle\psi_i| U(\theta)^\dagger ZZ U(\theta) |\psi_i\rangle + 1)$, and $P(odd|\psi_i) = 1 - P(even|\psi_i)$. The training was performed on 200 different initial circuits, each constituting a different trial. We trained the circuit to

minimize the binary cross-entropy loss

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(P(even|\psi_i)) + (1-y_i)\log(1-P(even|\psi_i)), \tag{11}$$

Optimization was performed using the Adam optimizer with a learning rate of 0.001 [10] and a single randomly selected MNIST example used for each update.

Figure 2 (a) shows the mean accuracy and standard deviation on a binarized MNIST dataset as a function of the training iterations for a circuit initialized using identity blocks compared with a strategy where all parameters are initially set to zero. While both strategies result in a circuit that initially evaluates to the identity, initializing all parameters to zero made the training much less efficient. Figure 2 (b) shows the variance of the partial derivative for parameters associated to the first three qubits, across different trials, and as a function of training iterations for circuits initialized with identity blocks. From the figures we observe that the model does not get stuck in a barren plateau (red dashed line), and that the variance decreases only as the model converges to a minimum of the objective function.

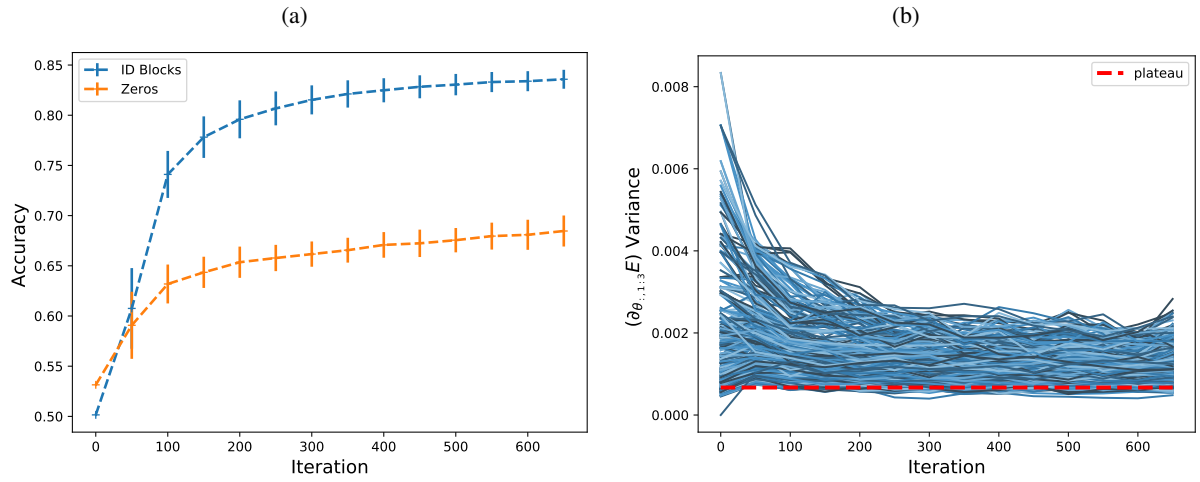

Figure 2: (a) Training accuracy and standard deviation as a function of the number of training iterations for an MNIST classifying circuit initialized using two identity blocks compared with a circuit where all parameters are initially set to zero, and (b) variance across trials of the partial derivatives for parameters associated to the first three qubits in the circuit using the identity block initialization method.
The circuit initialized using identity blocks trained successfully in all trials and never encountered the barren plateau (red dashed line).

4.3 Training a variational quantum eigensolver

In this experiment we use the identity block strategy and train a variational quantum eigensolver (VQE) to find ground state energies. We chose the 7-qubit Heisenberg model on a 1D lattice with periodic boundary conditions and in the presence of an external magnetic field. The corresponding Hamiltonian reads

$$H = J\sum_{(i,j)\in\mathcal{E}} (X_iX_j + Y_iY_j + Z_iZ_j) + h\sum_{i\in\mathcal{V}} Z_i, \tag{12}$$

where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the undirected graph of the lattice with 7 nodes, J expresses the strength of the spin-spin interactions, h corresponds to the strength of the magnetic field in the Z-direction. In the experiments we set $J = h = 1$. We chose this setting because for $J/h = 1$ the ground state is highly entangled (see also Ref. [3] for VQE simulations on the Heisenberg model).

Notice that the identity block initialization by itself cannot generate an entangled state at the very beginning. This could result in degraded performance for VQE. Hence, as the input state we chose $|\psi\rangle = B|0\rangle$, where B consists of 7 layers of random single-qubit rotations and controlled-Z gates. These gates are never updated during training, and their purpose is to provide some level of entanglement even at the beginning. Training was performed using the Adam optimizer with a learning rate of 0.001, until convergence.

Figure 3 shows the variance of the partial derivatives $(\partial_\theta E)$ across 200 trials for (a) a circuit initialized using $M = 2$ identity blocks and $L = 33$ layers per block, and (b) a randomly initialized circuit with the same number of layers in total. In (b) we observe a barren plateau for all parameters, while in (a) we observe that most of the variances are well above the plateau. As expected, within each identity block the variance increases with distance from the center.

Figure 4 (a) shows the mean and standard deviation of the expected energy across trials as a function of training iteration. A comparison is made between trials where circuits were initialized using identity blocks or where the initial parameters values were simply set to zero. Similar to the MNIST experiment, setting all parameters to zero resulted in poor training compared to the identity block strategy. The variance in the expectation when the zero parameter initialization strategy is used is explained by the random choice of the parameters in B for each trial. Panel (b) shows the variance of the gradient across different trials as a function of training iterations for circuits initialized using identity blocks. The variance of the gradient approached zero during training as the model's energy (blue line) converged to the ground state energy (green dashed line). Thus, similarly to what observed in the MNIST classification experiment, the circuit did not get stuck in a plateau.

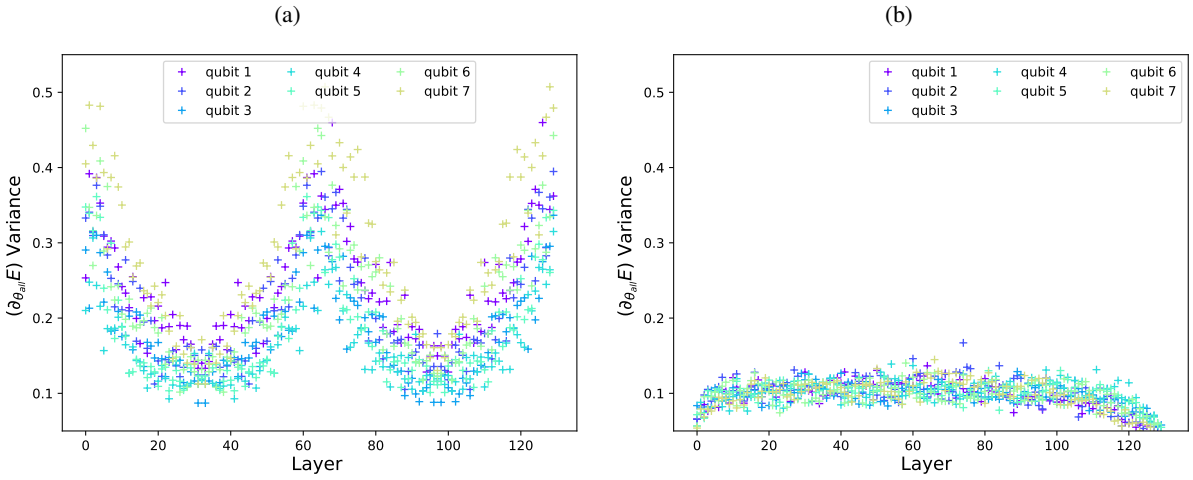

Figure 3: Variance of the gradient of a VQE circuit at the start of training across 200 trials for (a) a circuit initialized using two identity blocks, and (b) a random initialization. In (a) the variance increases with the distance from the center of each identity block and the majority of variances exceed those in (b) where a barren plateau is expected.

5 Conclusion

In this technical note we motivated and demonstrated a practical initialization strategy that addresses the problem of barren plateaus in the energy landscape of parametrized quantum circuits. In the experiments we conducted, the *identity block strategy* enabled us to perform well in two tasks: the variational quantum eigensolver (VQE), and the quantum neural network (QNN).

More work is needed to assess the impact of input states and data encoding methods. In the case of VQEs, the strategy does not initially allow the circuit to generate an entangled state. We resolved this by adding a shallow entangling layer that is fixed throughout training. In the case of QNNs, the encoded input data can already be highly entangled, thereby reducing the depth of the circuit where the plateau problem occurs. From these examples we conclude that there is a problem-dependent trade-off to be analyzed.

Finally, our approach is solely based on initialization of the parameter values. There are other potential strategies for avoiding barren plateaus such as layer-wise training, regularization, and imposing structural constraints on the ansatz. Understanding more about the relative merits of these and other approaches is a topic for future work.

6 Acknowledgements

E.G. is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [EP/P510270/1]. L.W. is supported by the Royal Society. M.O. acknowledges support from Polish National Science Center scholarship 2018/28/T/ST6/00429. M.B. is supported by EPSRC and by Cambridge Quantum Computing Limited (CQC). We

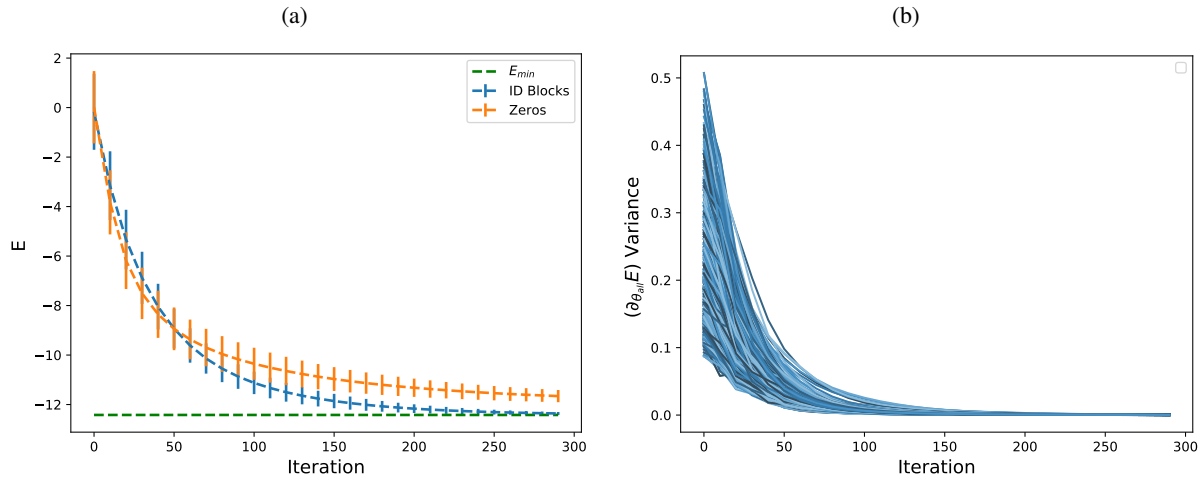

Figure 4: (a) Mean expected energy and standard deviation during training comparing a VQE initialized using identity blocks and a VQE initialized by setting all initial parameters to zero, and (b) variance of the gradient across trials as a function of training iteration for a VQE initialized using identity blocks. The circuit did not encounter a plateau during training since the variance of the gradient became small only as the model energy (blue line) converged to the ground state energy (green dashed line). In contrast to circuits initialized using identity blocks, circuits initialized by setting all parameters to zero failed to converge.

thank Raban Iten and Dominic Verdon for helpful technical discussions. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018. https://doi.org/10.1038/s41467-018-07090-4

[2] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5:4213, 2014. https://doi.org/10.1038/ncomms5213

[3] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242, 2017. https://doi.org/10.1038/nature23879

[4] Maria Schuld, Alex Bocharov, Krysta Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *arXiv preprint arXiv:1804.00633*, 2018.

[5] Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G Green, and Simone Severini. Hierarchical quantum classifiers. *npj Quantum Information*, 4(1):65, 2018. https://doi.org/10.1038/s41534-018-0116-9

[6] Hongxiang Chen, Leonard Wossnig, Simone Severini, Hartmut Neven, and Masoud Mohseni. Universal discriminative quantum neural networks. *arXiv preprint arXiv:1805.08654*, 2018.

[7] Dominic Verdon. Unitary 2-designs, variational quantum eigensolvers, and barren plateaus. https://qitheory.blogs.bristol.ac.uk/files/2019/02/barrenplateausblogpost-1xqcazi.pdf, 2019. [Online; accessed 13-March-2019].

[8] Zbigniew Puchała and Jaroslaw Adam Miszczak. Symbolic integration with respect to the haar measure on the unitary groups. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 65(1):21–27, 2017. https://doi.org/10.1515/bpasts-2017-0003

[9] Andris Ambainis and Joseph Emerson. Quantum t-designs: t-wise independence in the quantum world. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 129–140. IEEE, 2007. https://doi.org/10.1109/CCC.2007.26

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

7 Appendix

Here we provide a brief derivation of the vanishing gradient problem for the unitary group [7].

7.1 Vanishing gradient

For Hermitian $H \in \mathbb{C}^{N\times N}$ and a normalized state $|0\rangle \in \mathbb{C}^N$, we consider the function $E(U) = \langle 0| U^\dagger H U |0\rangle$ for $U \in U(N)$, where $U(N)$ denotes the unitary group of dimension N. In the following, we calculate the derivative of $E(U)$ at a unitary U in direction Z, where Z lies in the tangent space at of the unitary group at the point U. To do so, we choose a path $U(\alpha)$ such that $U(0) = U$ and $\partial_\alpha U|_{\alpha=0} = Z$. We have

$$\partial_\alpha E(U(\alpha)) = \partial_\alpha \langle 0| U(\alpha)^\dagger H U(\alpha) |0\rangle \tag{13}$$

$$= \langle 0| Z^\dagger H U + U^\dagger H Z |0\rangle . \tag{14}$$

We assume that Z in the tangent space at U has the form $Z = iUM$ for some Hermitian operator M, since it is easy to see that every Z of this form is in the tangent space at U and that every tangent vector at U can be written in this form. Then, we have

$$\partial_\alpha E(U(\alpha)) = \langle 0| (-iMU^\dagger) H U + U^\dagger H (iUM) |0\rangle \tag{15}$$

$$= i \langle 0| U^\dagger H U M - M U^\dagger H U |0\rangle \tag{16}$$

$$= i \langle 0| [U^\dagger H U, M] |0\rangle . \tag{17}$$

Now, we would like to calculate the gradient over the whole unitary group. For this, we fix the Hermitian matrix M and find

$$\mathbb{E}\left[\partial_\alpha E(U(\alpha))\right] = i \langle 0| \left[\mathbb{E}[U^\dagger H U], M\right] |0\rangle \tag{18}$$

$$= i \frac{\mathrm{Tr}(H)}{N} \langle 0| [I, M] |0\rangle \tag{19}$$

$$= 0 , \tag{20}$$

where we have used that for the Haar measure on the unitary group $\mu(U)$, $U \in U(N)$ it holds that $\int d\mu(U) U O U^\dagger = \frac{\mathrm{Tr}(O)}{N} I$, see [8].

Notably, if we initialize the matrix U to be the identity for a fixed H, which could for example be achieved by just taking half the depth of the initial parametrized circuit $U_{1/2}$ and then appending the adjoint $U_{1/2}^\dagger$. The full initial circuit becomes

$$U_{1/2}(\alpha) U_{1/2}^\dagger(\alpha) = I , \tag{21}$$

then this is always the identity, i.e., constant. Plugging the identity into the expectation of the gradient, we then obtain

$$\mathbb{E}\left[\partial_\alpha E(U(\alpha))\right] = i \langle 0| [H, M] |0\rangle , \tag{22}$$

which is only zero whenever the Hamiltonian commutes with the observable, which is generally not the case. Note that this insight also holds for any other identity initialization such as the block initialization introduced in the body of the paper.

Note that trainable gates often take the form $\exp(-i\alpha_j V_j)$. If the V_j's are chosen at random from tensor products of Pauli matrices $\{I, Z, X, Y\}^{\otimes n}$, then with high probability at least one of the derivatives is non-zero unless H is the identity, see Eq. (22). In sight of the initialization strategy, it is worth noting that initializing the circuit as $UU^\dagger$ hence does not guarantee by itself that at least one derivative is non zero.

7.2 Vanishing variance

We start with a simple identity.

Lemma 1. *It holds that*

$$\left(U^\dagger A U B U^\dagger C U\right)_{ij} = \sum_{k,l,m,n,p,q} A_{nm} B_{pq} C_{lk} U_{mp} U_{kj} U^*_{ni} U^*_{lq} . \tag{23}$$

Proof. The proof follows from entry-wise evaluation. □

We further need the following identity for the second moments in the proof.

Lemma 2 ([8]). *For $\int d\mu(U)$ being the integral over the unitary group with respect to the random Haar measure, it holds that*

$$\int d\mu(U) U_{i_1 j_1} U_{i_2 j_2} U^*_{i'_1 j'_1} U^*_{i'_2 j'_2} = \frac{\delta_{i_1 i'_1}\delta_{i_2 i'_2}\delta_{j_1 j'_1}\delta_{j_2 j'_2} + \delta_{i_1 i'_2}\delta_{i_2 i'_1}\delta_{j_1 j'_2}\delta_{j_2 j'_1}}{N^2-1} - \frac{\delta_{i_1 i'_1}\delta_{i_2 i'_2}\delta_{j_1 j'_2}\delta_{j_2 j'_1} + \delta_{i_1 i'_2}\delta_{i_2 i'_1}\delta_{j_1 j'_1}\delta_{j_2 j'_2}}{N(N^2-1)}. \tag{24}$$

First observe that we can explicitly evaluate the variance and obtain

$$\begin{aligned} \mathrm{Var}\left[\partial_\alpha E(U(\alpha))\right] &= -\mathrm{Tr}\left(\rho_0 \left(U^\dagger H U M - M U^\dagger H U\right)\right)^2 && (25)\\ &= \mathrm{Tr}\left(\rho_0 U^\dagger H U M \rho_0 M U^\dagger H U\right) - \mathrm{Tr}\left(\rho_0 U^\dagger H U M \rho_0 U^\dagger H U M\right) + \\ &\quad \mathrm{Tr}\left(\rho_0 M U^\dagger H U \rho_0 U^\dagger H U M\right) - \mathrm{Tr}\left(\rho_0 M U^\dagger H U \rho_0 M U^\dagger H U\right). && (26)\end{aligned}$$

Note that here we used the fact that the square of the trace is the trace of the square since ρ is a rank one matrix, i.e., a projector.

We can proceed now by evaluating the expectation of each term individually. As an example we calculate the first term, since the remaining terms can be evaluated in a similar fashion. We need to evaluate the following term

$$\int d\mu(U) \mathrm{Tr}\left(\rho_0 U^\dagger H U M \rho_0 M U^\dagger H U\right) = \mathrm{Tr}\left(\rho_0 \underbrace{\int d\mu(U) U^\dagger H U M \rho_0 M U^\dagger H U}_{=:\xi}\right). \tag{27}$$

Note that we can look at the integrand entry-wise before evaluating the trace. With $A = C := H$, $B := M\rho_0 M$, we have

$$\begin{aligned} \xi_{ij} &= \int d\mu([U]_{ij}) \sum_{k,l,m,n,p,q} A_{nm} B_{pq} C_{lk} U_{mp} U_{kj} U^*_{ni} U^*_{lq} && (28)\\ &= \sum_{k,l,m,n,p,q} A_{nm} B_{pq} C_{lk} \int d\mu([U]_{ij}) U_{mp} U_{kj} U^*_{ni} U^*_{lq} && (29)\\ &= \sum_{k,l,m,n,p,q} A_{nm} B_{pq} C_{lk} \left(\frac{\delta_{mn}\delta_{kl}\delta_{pi}\delta_{jq} + \delta_{ml}\delta_{kn}\delta_{pq}\delta_{ji}}{N^2-1}\right) - \\ &\quad \sum_{k,l,m,n,p,q} A_{nm} B_{pq} C_{lk} \left(\frac{\delta_{mn}\delta_{kl}\delta_{pq}\delta_{ji} + \delta_{ml}\delta_{kn}\delta_{pi}\delta_{jq}}{N(N^2-1)}\right) && (30)\\ &= \frac{\mathrm{Tr}\left(A\right)\mathrm{Tr}\left(C\right)}{N^2-1} B_{ij} + \frac{\mathrm{Tr}\left(B\right)\mathrm{Tr}\left(AC\right)}{N^2-1}\delta_{ij} - \\ &\quad \frac{\mathrm{Tr}\left(A\right)\mathrm{Tr}\left(B\right)\mathrm{Tr}\left(C\right)}{N(N^2-1)}\delta_{ij} - \frac{\mathrm{Tr}\left(AC\right)}{N(N^2-1)} B_{ij}. && (31)\end{aligned}$$

Note that plugging in A, B and C in Eq. (31), then yields

$$\frac{\mathrm{Tr}\left(H\right)^2}{N^2-1}(M\rho_0 M)_{ij} + \frac{\mathrm{Tr}\left(\rho_0 M^2\right)\mathrm{Tr}\left(H^2\right)}{N^2-1}\delta_{ij} - \frac{\mathrm{Tr}\left(H\right)^2\mathrm{Tr}\left(\rho_0 M^2\right)}{N(N^2-1)}\delta_{ij} - \frac{\mathrm{Tr}\left(H^2\right)}{N(N^2-1)}(M\rho_0 M)_{ij}. \tag{32}$$

Doing similar calculations for the other terms (using (31) for different A, B and C) and canceling and summarizing terms, yields the variance

$$\mathrm{Var}\left[\partial_\alpha E(U(\alpha))\right] = 2\frac{(M^2)_{00} - (M_{00})^2}{N^2-1}\left(\mathrm{Tr}\left(H^2\right) - \frac{\mathrm{Tr}\left(H\right)^2}{N}\right), \tag{33}$$

where H_{kl} denotes the (k,l)-entry of a matrix H and $M_{00} := \langle 0| M |0\rangle$, $(M^2)_{00} = \langle 0| M^2 |0\rangle$. This indicates that the variance indeed also decreases exponentially with the number of qubits.