Barren plateaus in quantum neural network training landscapes

Jarrod R. McClean,[1,*] Sergio Boixo,[1,†] Vadim N. Smelyanskiy,[1,‡] Ryan Babbush,[1] and Hartmut Neven[1]

[1] *Google Inc., 340 Main Street, Venice, CA 90291, USA*

(Dated: March 30, 2018)

Many experimental proposals for noisy intermediate scale quantum devices involve training a parameterized quantum circuit with a classical optimization loop. Such hybrid quantum-classical algorithms are popular for applications in quantum simulation, optimization, and machine learning. Due to its simplicity and hardware efficiency, random circuits are often proposed as initial guesses for exploring the space of quantum states. We show that the exponential dimension of Hilbert space and the gradient estimation complexity make this choice unsuitable for hybrid quantum-classical algorithms run on more than a few qubits. Specifically, we show that for a wide class of reasonable parameterized quantum circuits, the probability that the gradient along any reasonable direction is non-zero to some fixed precision is exponentially small as a function of the number of qubits. We argue that this is related to the 2-design characteristic of random circuits, and that solutions to this problem must be studied.

Rapid developments in quantum hardware have motivated advances in algorithms to run in the so-called noisy intermediate scale quantum (NISQ) regime [1]. Many of the most promising application-oriented approaches are hybrid quantum-classical algorithms that rely on optimization of a parameterized quantum circuit [2–8]. The resilience of these approaches to certain types of errors and high flexibility with respect to coherence time and gate requirements make them especially attractive for NISQ implementations [3, 9–11].

The first implementation of such algorithms was developed in the context of quantum simulation with the variational quantum eigensolver [2, 3]. This algorithm has been successfully demonstrated on a number of experimental setups with extensions to excited states and other forms of incoherent error mitigation [2, 9, 12–16]. Since then, the quantum approximate optimization algorithm was developed in a similar context to address hard optimization problems [5, 17–19]. This algorithm has also been demonstrated on quantum devices [20]. These approaches have even been extended to both quantum machine learning and error correction [6, 7, 20–23].

While the precise formulation of these methods and their domains of applicability differ considerably, they typically tend to rely upon the optimization of some parameterized unitary circuit with respect to an objective function that is typically a simple sum of Pauli operators or fidelity with respect to some state. This framework is reminiscent of the methodology of classical neural networks [23, 24]. As with any non-linear optimization, the choice of both the parameterization and the initial state is important. In quantum simulation, there is often a choice inspired by physical domain knowledge [3, 17, 25–29]. However, in all domains of applicability, there have been implementations that utilize parametrized random circuits of varying depth [7, 13, 21, 23, 30]. Within quantum simulation that approach has been referred to as a "hardware efficient ansatz" [13].

When little structure is known about the problem or constraints of the existing quantum hardware may prevent utilizing that structure, choosing a random implementable circuit seems to provide an unbiased choice. One might also expect, based on recent experimental designs for "quantum supremacy", that random quantum circuits are a powerful tool for such a task [31]. Also, despite concerns about gradient-based methods in classical deep neural networks [32–34], they are successful [24], even if using random initialization [33, 35]. However, in the quantum case one must remember that the estimation of even a single gradient component will scale as $O(1/\epsilon^{\alpha})$ for some small power α [36] as opposed to classical implementations where the same is achieved in $O(\log(1/\epsilon))$ time.

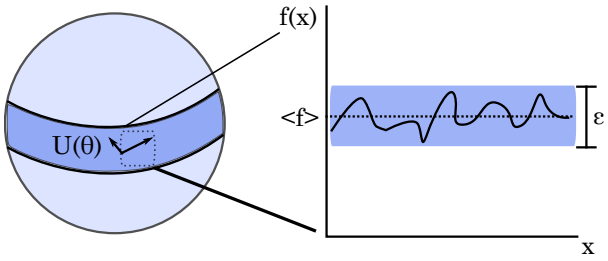

FIG. 1. A cartoon of the general geometric results from this work. The sphere depicts the phenomenon of concentration of measure in quantum space: the fraction of states that fall outside a fixed angular distance from zero along any coordinate decreases exponentially in the number of qubits [37]. This implies a flat plateau where observables concentrate on their average over Hilbert space and the gradient is exponentially small. The fact that only an exponentially small fraction of states fall outside of this band means that searches resembling random walks will have an exponentially small probability of exiting this "barren plateau".

We will present results related to random quantum circuits in the context of the exponential dimension of Hilbert space and gradient based hybrid quantum-classical algorithms. A cartoon depiction of this is given

in Figure 1. We show that for a large class of random circuits, the average value of the gradient of the objective function is zero, and the probability that any given instance of such a random circuit deviates from this average value by a small constant ϵ is exponentially small in the number of qubits. This can be understood in the geometric context of concentration of measure [38–40] for high dimensional spaces. When the measure of the space concentrates in this way, the value of any reasonably smooth function will tend towards its average with exponential probability, a fact made formal by Levy's lemma [37]. In our context, this means that the gradient is zero over vast reaches of quantum space.

The region where the gradient is zero does not correspond to local minima of interest, but rather an exponentially large plateau of states that have exponentially small deviations in the objective value from the average the totally mixed state. We argue that the depth of circuits which achieve these undesirable properties are modest, requiring only $O(n^{1/d})$ depth circuits on a d dimensional array, and numerically evaluate the constant factors one expects to encounter for small instances of this kind. We close with an outlook on how this result should shape strategies in ansatz design for scaling to larger experiments.

Gradient concentration in random circuits

We will discuss random parameterized quantum circuits (RPQCs)

$$U(\vec{\theta}) = U(\theta_1, ..., \theta_L) = \prod_{l=1}^{L} U_l(\theta_l) W_l \tag{1}$$

where $U_l(\theta_l) = \exp(-i\theta_l V_l)$, V_l is a hermitian operator, and W_l is a generic unitary operator that does not depend on any angle θ_l. Circuits of this form are a natural choice due to a straightforward evaluation of the gradient with respect to most objective functions and have been introduced in a number of contexts already [26, 41]. Consider an objective function $E(\theta)$ expressed as the expectation value over some hermitian operator H,

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

When the RPQCs are parameterized in this way, the gradient of the objective function takes a simple form:

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

where we introduce the notations $U_- \equiv \prod_{l=0}^{k-1} U_l(\theta_l) W_l$, $U_+ \equiv \prod_{l=k}^{L} U_l(\theta_l) W_l$, and henceforth drop the subscript k from $V_k \to V$ for ease of exposition. Finally, we will define our RPQCs $U(\vec{\theta})$ to have the property that for any

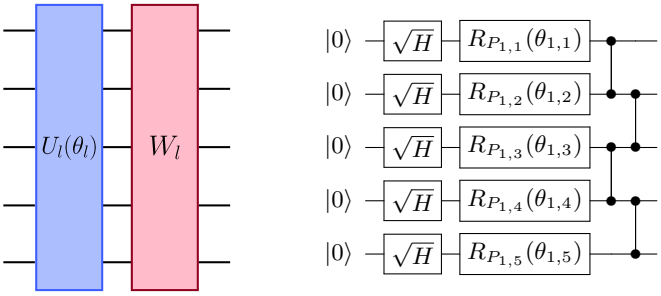

FIG. 2. Left: The generic subunit of circuits we study in this work, with a parameterized component $U_l(\theta_l)$ and non-parameterized unit W_l for each layer l. Right: Example schematic of the 1D random circuits used in our numerical experiments. The circuit begins with a square root of Hadamard applied to all qubits followed by a specified number of layers of randomly chosen Pauli rotations applied to each qubit and then a 1D ladder of controlled Z gates. The initial square root of Hadamard gates are not repeated in each layer. The indices i and j in $\theta_{i,j}$ index the layer and qubit respectively. For each layer and qubit $P_{i,j} \in \{X, Y, Z\}$ and $\theta_{i,j} \in [0, 2\pi)$ are sampled independently.

gradient direction $\partial_k E$ defined above, the circuit implementing $U(\vec{\theta})$ is sufficiently random such that either U_-, U_+, or both match the Haar distribution up to the second moment, and the circuits U_- and U_+ are independent.

Our results make use of properties of the Haar measure on the unitary group $d\mu_{\text{Haar}}(U) \equiv d\mu(U)$, which is the unique left- and right-invariant measure such that

$$\int_{U(N)} d\mu(U) f(U) = \int d\mu(U) f(VU) = \int d\mu(U) f(UV) \tag{4}$$

for any $f(U)$ and $V \in U(N)$, where the integration domain will be implied to be $U(N)$ when not explicitly listed. While this property is valuable for proofs, quantum circuits that exactly achieve this invariance generically require exponential resources. This motivates the concept of unitary t-designs [42–44], which satisfy the above properties for restricted classes of $f(U)$, often requiring only modest polynomial resources. Suppose $\{p_i, V_i\}$ is an ensemble of unitary operators, with unitary V_i being sampled with probability p_i. The ensemble $\{p_i, V_i\}$ is a k-design if

$$\sum_i p_i V_i^{\otimes t} \rho (V_i^\dagger)^{\otimes t} = \int d\mu(U) U^{\otimes t} \rho (U^\dagger)^{\otimes t}. \tag{5}$$

This definition is equivalent to the property that if $f(U)$ is a polynomial of at most degree t in the matrix elements of U and at most degree t in the matrix elements of U^*, then averaging over the t-design $\{p_i, V_i\}$ will yield the same result as averaging over the unitary group with the respect to the Haar measure.

The average value of the gradient is a concept that requires additional specification because, for a given point, the gradient can only be defined in terms of the circuit

that led to that point. We will use a practical definition that leads to the value we are interested in, namely

$$\langle \partial_k E \rangle = \int dU p(U) \partial_k \langle 0| U(\vec{\theta})^\dagger H U(\vec{\theta}) |0\rangle \tag{6}$$

where $p(U)$ is the probability distribution function of U. A review on the properties of products of independent random matrices can be found in Ref. [45]. The assumptions of independence and at least one of U_- or U_+ forming a 1-design in our RPQCs implies that $\langle \partial_k E \rangle = 0$, as shown in the appendix.

Levy's lemma informs our intuition about the the expected variance of this quantity through simple geometric arguments. In particular, Haar random unitaries on n qubits will output states uniformly in the $D = 2^n - 1$ dimensional hypersphere. The derivative with respect to the parameters θ is Lipschitz continuous with some parameter η that depends on the operator H. Levy's lemma then implies that the variance of measurements will decrease exponentially in the number of qubits. This intuition may be made more precise through explicit calculation of the variance, which is done in more detail in the appendix. The result is that

$$\mathrm{Var}\left[\partial_k E\right] = \begin{cases} -\frac{\mathrm{Tr}(\rho^2)}{2^{2n}} \mathrm{Tr}\left\langle \left[V, u^\dagger H u\right]^2 \right\rangle_{U_+} \\ -\frac{\mathrm{Tr}(H^2)}{2^{2n}} \mathrm{Tr}\left\langle \left[V, u \rho u^\dagger\right]^2 \right\rangle_{U_-} \\ 2\ \mathrm{Tr}\left(H^2\right) \mathrm{Tr}\left(\rho^2\right) \left(\frac{\mathrm{Tr}(V^2)}{2^{3n}} - \frac{\mathrm{Tr}(V)^2}{2^{4n}} \right) \end{cases} \tag{7}$$

where the notation $\langle f(u) \rangle_{U_x}$ indicates the average with u drawn from $p(U_x)$, and the first case corresponds to U_- being a 2 design and not U_+, the second to U_+ being a 2-design but not U_-, and the third to both U_+ and U_- being 2-designs. We emphasize the fact that this variance depends at most on polynomials of degree 2 in U and polynomials of degree 2 in U^*. Whereas a unitary 2-design will exhibit the correct variance [43, 46], a unitary 1-design will exhibit the correct average value, but not necessarily the variance. As a result, if a circuit is of sufficient depth such that for any $\partial_k E$, either U_- or U_+ forms a 2-design, then with high probability one will produce an ansatz state on a barren plateau of the quantum landscape, with no interesting search directions in sight.

Numerical simulations

The previous section shows that for reasonable classes of RPQCs at a sufficient number of qubits and depth, one will end up on a barren plateau. Here we check this result for even modest depth 1D random circuits with numerical simulations. This helps to clarify the rate of concentration for realistic circuits and shows the transition as the circuit grows in length from a single layer to a circuit demonstrating statistics analogous to a 2-design.

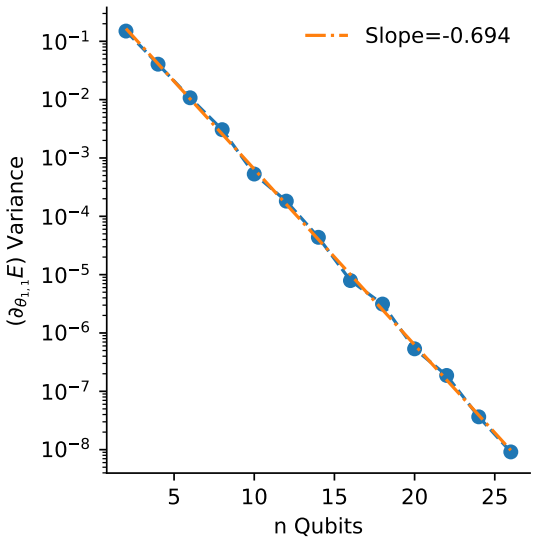

FIG. 3. The sample variance of the gradient of a two-local Pauli term plotted as a function of the number of qubits on a semi-log plot. As predicted, an exponential decay is observed as a function of the number of qubits for both the expected value and its spread.

The circuits and objective functions used in our numerical experiments begin with a layer of $R_Y(\pi/4) = \exp(-i\pi/8\ Y) = \sqrt{H}$ gates to prevent X, Y, or Z from being an especially preferential direction with respect to gradients. Then, the circuit proceeds by a number of layers. Each layer consists of a parallel application of single qubit rotations to all qubits, given by $R_P(\theta)$ where $P \in \{X, Y, Z\}$ is chosen with uniform probability and $\theta \in [0, 2\pi)$ is also chosen uniformly. This layer is followed by a layer of 1D nearest neighbor controlled phase gates, as in Figure 2. Thus, the number of angles is the number of qubits times the number of layers.

The objective operator H is chosen to be a single Pauli ZZ operator acting on the first and second qubit, $H = Z_1 Z_2$. The gradient is evaluated with respect to the first parameter, $\theta_{1,1}$. This simple choice helps to extract the exponential scaling. As complex objectives can be written as sums of these operators, the results for large objectives can be inferred from these numbers. Moreover, it's clear that for any polynomial sum of these operators, the exponential decay of the signal in the gradient will not be circumvented.

From Figure 3 we see that for a single 2-local Pauli term, both the expected value of the gradient and its spread decay exponentially as a function of the number of qubits even when the number of layers is a modest linear function. We also observe in Figure 4 that as the number of layers increases, there is a transition to a 2-design where the variance converges. This leads to a distinct plateau as the circuit length increases, where the

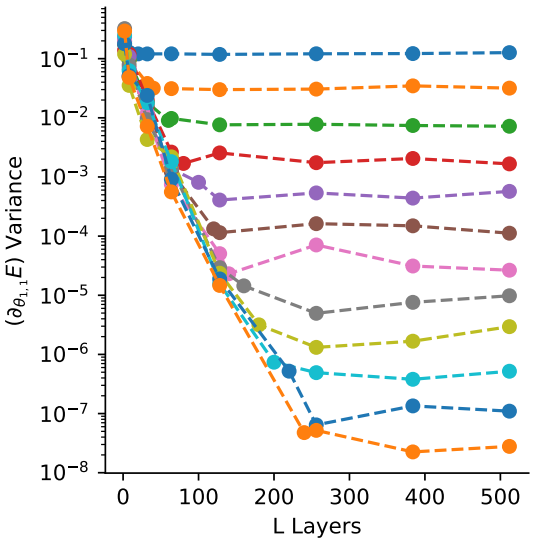

FIG. 4. Here we show the sample variance of the gradient of a two-local Pauli term plotted as a function of the number of layers in the 1D quantum circuit. The different lines correspond to all even numbers of qubits between 2 and 24, with 2 qubits being the top line, and the rest being ordered by qubit number. This shows the convergence of the second moment as a function of the number of layers to a fixed value determined by the number of qubits.

height of the plateau is determined by the number of qubits. These results substantiate our conclusion that gradients in modest-sized random circuits tend to vanish without additional mitigating steps.

Contrast with gradients in classical deep networks

Finally, we contrast our results with the vanishing (and exploding) gradient problem of classical deep neural networks [32–34, 47]. At least two key differences are present in the quantum case: (i) the different scaling of the vanishing gradient and (ii) the complexity of computing expected values.

The gradient in a classical deep neural network can vanish exponentially in the number of layers [32, 33], while in the a quantum circuit is exponentially small in the number of qubits, as shown above. Therefore, the later will generally be exponentially smaller than the former. In the classical case, the gradient for a weight in a neuron depends on the sum of all the paths connecting that neuron to the output, and when the weights are initialized with random values the paths have random signs which cancels the signal [32]. The number of paths is exponential in the number of layers. In the quantum case, the number of paths is exponential in the number of gates, and also have random signs [31]. The gradient saturates to an exponential in the number of qubits because the output state is normalized.

The estimation of the gradient for each training batch for a classical neural network is limited by machine precision and scales with $O(\log(1/\epsilon))$. Even if the gradient is small, as long as it is consistent enough between batches, the method may eventually succeed. On a quantum device, the cost of estimating the gradient scales as $O(1/\epsilon^{\alpha})$ [36]. For a number of measurements much lower than this limit with ϵ the size of the gradient, a gradient based optimization will result in a random walk. By concentration of measure, a random walk will have exponentially small probability of exiting the barren plateau. As a result, gradient descent without some additional strategy cannot circumvent this challenge on a quantum device in polynomial time.

Conclusions

We have seen both analytically and numerically that for a wide class of random quantum circuits, the expected values of observables concentrate to their averages over Hilbert space and gradients concentrate to zero. This represents an interesting statement about the geometry of quantum circuits and landscapes related to hybrid-quantum classical algorithms. More practically, it means that randomly initialized circuits of sufficient depth will find relatively little utility in hybrid quantum-classical algorithms.

Historically, vanishing gradients may have played a role in the early winter of deep neural networks [32–34, 47]. However, multiple techniques have been proposed to mitigate this problem [24, 35, 48, 49], and the amount of training data and computational power available has grown substantially. One approach to avoid these landscapes in the quantum setting is to use structured initial guesses, such as those adopted in quantum simulation. Another possibility is to use pre-training segment by segment, which was an early success in the classical setting [48, 50]. These or other alternatives must be studied if these ansätze are to be succesful beyond a few qubits.

* jmcclean@google.com
† boixo@google.com
‡ smelyan@google.com

[1] J. Preskill, "Quantum Computing in the NISQ era and beyond," arXiv:1801.00862 (2018).
[2] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, "A variational eigenvalue solver on a photonic quantum processor," Nature Communications **5**, 1 (2014).
[3] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-

Guzik, "The theory of variational hybrid quantum-classical algorithms," New Journal of Physics **18**, 023023 (2016).
[4] M.-H. Yung, J. Casanova, A. Mezzacapo, J. McClean, L. Lamata, A. Aspuru-Guzik, and E. Solano, "From transistor to trapped-ion computers for quantum chemistry," Scientific Reports **4**, 9 (2014).
[5] E. Farhi, J. Goldstone, and S. Gutmann, "A Quantum Approximate Optimization Algorithm," arXiv:1411.4028 (2014).
[6] P. D. Johnson, J. Romero, J. Olson, Y. Cao, and A. Aspuru-Guzik, "QVECTOR: an algorithm for device-tailored quantum error correction," arXiv:1711.02249 (2017).
[7] Y. Cao, G. Giacomo Guerreschi, and A. Aspuru-Guzik, "Quantum Neuron: an elementary building block for machine learning on quantum computers," arXiv:1711.11240 (2017).
[8] C. Hempel, C. Maier, J. Romero, J. R. McClean, T. Monz, H. Shen, P. Jurcevic, B. Lanyon, P. Love, R. Babbush, A. Aspuru-Guzik, R. Blatt, and C. Roos, "Quantum chemistry calculations on a trapped-ion quantum simulator," arXiv:1803.10238 (2018).
[9] P. J. J. O'Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tranter, N. Ding, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Jeffrey, A. Megrant, J. Y. Mutus, C. Neill, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, P. V. Coveney, P. J. Love, H. Neven, A. Aspuru-Guzik, and J. M. Martinis, "Scalable Quantum Simulation of Molecular Energies," Physical Review X **6**, 31007 (2016).
[10] J. R. McClean, M. E. Schwartz, J. Carter, and W. A. de Jong, "Hybrid Quantum-Classical Hierarchy for Mitigation of Decoherence and Determination of Excited States," Physical Review A **95**, 42308 (2017).
[11] D. Wecker, M. B. Hastings, and M. Troyer, "Progress towards practical quantum variational algorithms," Physical Review A **92**, 42303 (2015).
[12] Y. Shen, X. Zhang, S. Zhang, J.-N. Zhang, M.-H. Yung, and K. Kim, "Quantum Implementation of Unitary Coupled Cluster for Simulating Molecular Electronic Structure," Physical Review A **95**, 020501(R) (2017).
[13] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, J. M. Chow, and J. M. Gambetta, "Hardware-efficient Quantum Optimizer for Small Molecules and Quantum Magnets," Nature **549**, 242 (2017).
[14] J. I. Colless, V. V. Ramasesh, D. Dahlen, M. S. Blok, M. E. Kimchi-Schwartz, J. R. McClean, J. Carter, W. A. de Jong, and I. Siddiqi, "Computation of molecular spectra on a quantum processor with an error-resilient algorithm," Physical Review X **8**, 011021 (2018).
[15] R. Santagati, J. Wang, A. A. Gentile, S. Paesani, N. Wiebe, J. R. McClean, S. Morley-Short, P. J. Shadbolt, D. Bonneau, J. W. Silverstone, D. P. Tew, X. Zhou, J. L. O'Brien, and M. G. Thompson, "Witnessing eigenstates for quantum simulation of hamiltonian spectra," Science Advances **4** (2018).
[16] E. F. Dumitrescu, A. J. McCaskey, G. Hagen, G. R. Jansen, T. D. Morris, T. Papenbrock, R. C. Pooser, D. J. Dean, and P. Lougovski, "Cloud Quantum Computing of an Atomic Nucleus," arXiv:1801.03897 (2018).
[17] D. Wecker, M. B. Hastings, and M. Troyer, "Training a quantum optimizer," Physical Review A **94**, 022309 (2016).
[18] Z. Wang, S. Hadfield, Z. Jiang, and E. G. Rieffel, "Quantum approximate optimization algorithm for maxcut: A fermionic view," Physical Review A **97**, 022304 (2018).
[19] N. Moll, P. Barkoutsos, L. S. Bishop, J. M. Chow, A. Cross, D. J. Egger, S. Filipp, A. Fuhrer, J. M. Gambetta, M. Ganzhorn, *et al.*, "Quantum optimization using variational algorithms on near-term quantum devices," arXiv:1710.01022 (2017).
[20] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. Schuyler Fried, S. Hong, P. Karalekas, C. B. Osborn, A. Papageorge, E. C. Peterson, G. Prawiroatmodjo, N. Rubin, C. A. Ryan, D. Scarabelli, M. Scheer, E. A. Sete, P. Sivarajah, R. S. Smith, A. Staley, N. Tezak, W. J. Zeng, A. Hudson, B. R. Johnson, M. Reagor, M. P. da Silva, and C. Rigetti, "Unsupervised Machine Learning on a Hybrid Quantum Computer," arXiv:1712.05771 .
[21] J. Romero, J. P. Olson, and A. Aspuru-Guzik, "Quantum autoencoders for efficient compression of quantum data," Quantum Science and Technology **2**, 045001 (2017).
[22] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," Nature **549**, 195 (2017).
[23] E. Farhi and H. Neven, "Classification with quantum neural networks on near term processors," arXiv:1802.06002 (2018).
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436 (2015).
[25] J. R. McClean, R. Babbush, P. J. Love, and A. Aspuru-Guzik, "Exploiting locality in quantum computation for quantum chemistry," The Journal of Physical Chemistry Letters **5**, 4368 (2014).
[26] J. Romero, R. Babbush, J. R. McClean, C. Hempel, P. Love, and A. Aspuru-Guzik, "Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz," arXiv:1701.02691 (2017).
[27] R. Babbush, N. Wiebe, J. McClean, J. McClain, H. Neven, and G. K.-L. Chan, "Low-depth quantum simulation of materials," Physical Review X **8**, 011044 (2018).
[28] N. C. Rubin, R. Babbush, and J. R. McClean, "Application of fermionic marginal constraints to hybrid quantum algorithms," arXiv:1801.03524 (2018).
[29] I. D. Kivlichan, J. McClean, N. Wiebe, C. Gidney, A. Aspuru-Guzik, G. K.-L. Chan, and R. Babbush, "Quantum simulation of electronic structure with linear depth and connectivity," Physical Review Letters **120**, 110501 (2018).
[30] E. Farhi, J. Goldstone, S. Gutmann, and H. Neven, "Quantum Algorithms for Fixed Qubit Architectures," arXiv:1703.06199 (2017).
[31] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, "Characterizing Quantum Supremacy in Near-Term Devices," arXiv:1608.00263 (2016).
[32] D. M. Bradley, *Learning in modular systems* (Carnegie Mellon University, 2010).
[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS* (2010) pp. 249–256.
[34] S. Shalev-Shwartz, O. Shamir, and S. Shammah, "Failures of Gradient-Based Deep Learning,"

arXiv:1703.07950 (2017).
[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML* (2015) pp. 448–456.
[36] E. Knill, G. Ortiz, and R. D. Somma, "Optimal quantum measurements of expectation values of observables," Physical Review A **75**, 012328 (2007).
[37] M. Ledoux, *The concentration of measure phenomenon*, 89 (American Mathematical Soc., 2005).
[38] S. Popescu, A. J. Short, and A. Winter, "Entanglement and the foundations of statistical mechanics," Nature Physics **2**, 754 (2006).
[39] M. J. Bremner, C. Mora, and A. Winter, "Are random pure states useful for quantum computation?" Physical Review Letters **102**, 190502 (2009).
[40] D. Gross, S. T. Flammia, and J. Eisert, "Most quantum states are too entangled to be useful as computational resources," Physical Review Letters **102**, 190501 (2009).
[41] G. G. Guerreschi and M. Smelyanskiy, "Practical optimization for hybrid quantum-classical algorithms," arXiv:1701.01450 (2017).
[42] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, "Symmetric informationally complete quantum measurements," Journal of Mathematical Physics **45**, 2171 (2004).
[43] C. Dankert, R. Cleve, J. Emerson, and E. Livine, "Exact and approximate unitary 2-designs and their application to fidelity estimation," Physical Review A **80**, 012304 (2009).
[44] A. W. Harrow and R. A. Low, "Random quantum circuits are approximate 2-designs," Communications in Mathematical Physics **291**, 257 (2009).
[45] J. R. Ipsen, "Products of Independent Gaussian Random Matrices," arXiv:1510.06128 (2015).
[46] D. A. Roberts and B. Yoshida, "Chaos and complexity by design," Journal of High Energy Physics **2017**, 121 (2017).
[47] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," (2001).
[48] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural computation **18**, 1527 (2006).
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.
[50] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems* (2007) pp. 153–160.
[51] Z. Puchała and J. A. Miszczak, "Symbolic integration with respect to the haar measure on the unitary groups," Bulletin of the Polish Academy of Sciences Technical Sciences **65**, 21 (2017).

Appendix I

Here we explicitly show the expectation value of the gradient is 0 and that under our assumptions the variance decays exponentially in the number of qubits. By our definition of RPQCs, we have that for any specified direction $\partial_k E$, both U_- and U_+ are independently distributed and either U_- or U_+ match the Haar distribution up to at least the second moment (the are a 2-design). The assumption of independence is equivalent to

$$p(U) = \int dU_+ p(U_+) \int dU_- p(U_-) \times \delta(U_+ U_- - U). \tag{8}$$

which allows us to rewrite the expression as

$$\langle \partial_k E \rangle = i \int dU_- p(U_-) \mathrm{Tr}\left\{\rho_- \times \int dU_+ p(U_+) \left[V, U_+^\dagger H U_+\right]\right\} \tag{9}$$

We will utilize explicit integration over the unitary group with respect to the Haar measure, which up to the first moment can be expressed as [51]

$$\int d\mu(U)\ U_{ij} U^\dagger_{km} = \int d\mu(U) U_{ij} U^*_{mk} = \frac{\delta_{im}\delta_{jk}}{N}. \tag{10}$$

where N is the dimension of the space, typically 2^n for n qubits. Using this expression, one may readily verify that

$$M = \int d\mu(U) U O U^\dagger = \frac{\mathrm{Tr} O}{N} I \tag{11}$$

which we use in the following. Now, making use of the assumption that either U_+ or U_- matches the Haar measure up to the first moment (it is a 1-design), we first examine the case where U_- is at least a 1-design and find that

$$\begin{aligned}\langle \partial_k E \rangle &= i \int d\mu\left(U_-\right) \mathrm{Tr}\left\{\rho_- \times \left[V, \int dU_+ p(U_+) U_+^\dagger H U_+\right]\right\} \\ &= \frac{i}{N} \mathrm{Tr}\left\{\left[V, \int dU_+ p(U_+) U_+^\dagger H U_+\right]\right\} \\ &= 0\end{aligned} \tag{12}$$

where we have defined $\rho_- = U_- |0\rangle \langle 0| U_-^\dagger$ and used the fact that the trace of a commutator of trace class operators is zero. In the second case, where we assume U_+ is at least a 1-design,

$$\begin{aligned}\langle \partial_k E \rangle &= i \int dU_- p(U_-) \mathrm{Tr}\left\{\rho_- \int d\mu\left(U_+\right) \left[V, U_+^\dagger H U_+\right]\right\} \\ &= i \frac{\mathrm{Tr} H}{N} \int dU_- p(U_-) \mathrm{Tr}\left\{\rho_- \left[V, I\right]\right\} \\ &= 0.\end{aligned} \tag{13}$$

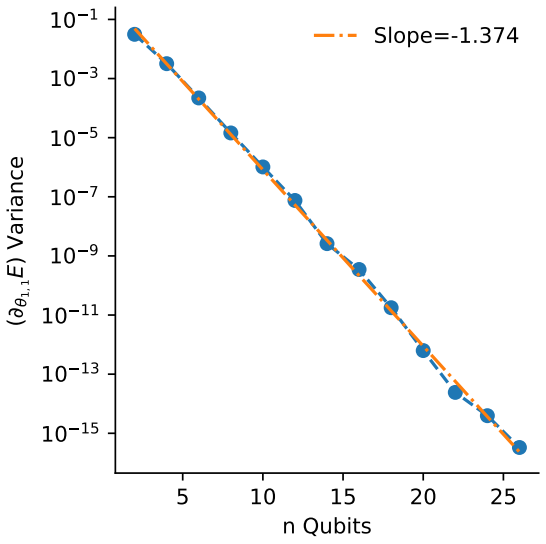

FIG. 5. The sample variance of the gradient of $H = |00...0\rangle\langle 00...0|$ plotted as a function of the number of qubits on a semi-log plot. As predicted, an exponential decay is observed as a function of the number of qubits for both the expected value and its spread.

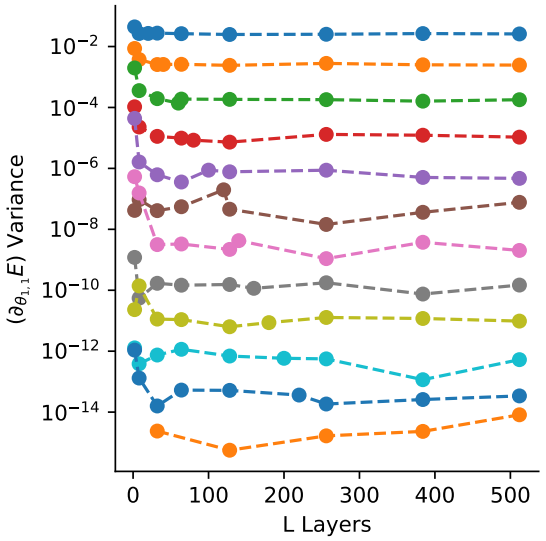

FIG. 6. Here we show the sample variance the gradient of $H = |00...0\rangle\langle 00...0|$ plotted as a function of the number of layers in the 1D quantum circuit. The different lines correspond to all even numbers of qubits between 2 and 24, with 2 qubits being the top line, and the rest being ordered by qubit number. This shows the convergence of the second moment as a function of the number of layers to a fixed value determined by the number of qubits.

An advantage of the explicit polynomial formulas are that they allow an analytic calculation of the variance as well, which allows precise specification of the coefficient in Levy's lemma. In cases where the integrals depend on up to two powers of elements of U and U^*, one may make use of the elementwise formula [51]

$$\int d\mu(U) U_{i_1 j_1} U_{i_2 j_2} U^*_{i'_1 j'_1} U^*_{i'_2 j'_2} = \frac{\delta_{i_1 i'_1}\delta_{i_2 i'_2}\delta_{j_1 j'_1}\delta_{j_2 j'_2} + \delta_{i_1 i'_2}\delta_{i_2 i'_1}\delta_{j_1 j'_2}\delta_{j_2 j'_1}}{N^2-1} - \frac{\delta_{i_1 i'_1}\delta_{i_2 i'_2}\delta_{j_1 j'_2}\delta_{j_2 j'_1} + \delta_{i_1 i'_2}\delta_{i_2 i'_1}\delta_{j_1 j'_1}\delta_{j_2 j'_2}}{N(N^2-1)} \tag{14}$$

The variance of the gradient is defined by

$$\mathrm{Var}[\partial_k E] = \langle (\partial_k E)^2 \rangle \tag{15}$$

as we have seen above that $\langle \partial_k E \rangle = 0$. Through use of the above formula for integration up to the second moment of the Haar distribution, one may evaluate this expression in 3 separate cases. In the case where U_- is a 2-design but not U_+,

$$\begin{aligned}\mathrm{Var}[\partial_k E] &= \frac{2\mathrm{Tr}(\rho^2)}{N^2}\mathrm{Tr}\langle H_u^2 V^2 - (H_u V)^2\rangle_{U_+} \\ &= -\frac{\mathrm{Tr}(\rho^2)}{2^{2n}}\mathrm{Tr}\langle [V, H_u]^2\rangle_{U_+}\end{aligned} \tag{16}$$

where $H_u = u^\dagger H u$ and we have defined the notation $\langle f(u)\rangle_{U_x}$ to mean the average over u sampled from $p(U_x)$. In the case where U_+ is a 2-design but not U_-,

$$\begin{aligned}\mathrm{Var}[\partial_k E] &= \frac{2\mathrm{Tr}(H^2)}{N^2}\mathrm{Tr}\langle \rho_u^2 V^2 - (\rho_u V)^2\rangle_{U_-} \\ &= -\frac{\mathrm{Tr}(H^2)}{2^{2n}}\mathrm{Tr}\langle [V, \rho_u]^2\rangle_{U_-}\end{aligned} \tag{17}$$

where $\rho_u = u\rho u^\dagger$. Finally in the case where both U_+ and U_- are 2-designs

$$\mathrm{Var}[\partial_k E] = 2\ \mathrm{Tr}\left(H^2\right)\mathrm{Tr}\left(\rho^2\right)\left(\frac{\mathrm{Tr}\left(V^2\right)}{2^{3n}} - \frac{\mathrm{Tr}\left(V\right)^2}{2^{4n}}\right). \tag{18}$$

In all cases, the exponential decay of the gradient as a function of the number of qubits is evident.

Appendix II

Here we provide data for an additional numerical example that is particularly relevant to circuit and state learning tasks. Explicitly, we take as the objective function the projection onto a state of interest, which due to rotational invariance we can set to be the all 0 computational state. Alternatively, we can write $H = |00...0\rangle\langle 00...0|$. The results of the simulation of the gradient variance as a function of the number of qubits and layers are shown in Figure 5 and Figure 6.