Quantum energy landscape and circuit optimization

Joonho Kim and Yaron Oz

Phys. Rev. A **106**, 052424 — Published 22 November 2022

DOI: 10.1103/PhysRevA.106.052424

Quantum Energy Landscape and Circuit Optimization

Joonho Kim[1] and Yaron Oz[1,2]

[1] *School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA.*
[2] *Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Tel-Aviv 69978, Israel.*

We study the effects of entanglement and control parameters on the symmetry of the energy landscape and optimization performance of the variational quantum circuit. Through a systematic analysis of the Hessian spectrum, we characterize the local geometry of the energy landscape at a random point and along an optimization trajectory. We argue that decreasing the entangling capability and increasing the number of circuit parameters have the same qualitative effect on the Hessian eigenspectrum. Both the low-entangling capability and the abundance of control parameters increase the curvature of non-flat directions, contributing to the efficient search of area-law entangled ground states as to the optimization accuracy and the convergence speed.

I. INTRODUCTION

The variational quantum algorithm (VQA) is arguably the most promising framework to achieve the near-term quantum advantage [1, 2]. The structure of typical VQA computation consists of three parts: First, the quantum processor constructs the wavefunction $|\psi(\boldsymbol{\theta})\rangle$ by acting a sequence of unitary gate operations, which often depend on randomly chosen control parameters $\boldsymbol{\theta}$, on the initial product state $|0\rangle^{\otimes n}$. Second, the quantum processor measures the variational wavefunction, where outputs of the repeated measurements, e.g., $\langle\psi(\boldsymbol{\theta})|Z_i|\psi(\boldsymbol{\theta})\rangle$ for $1 \leq i \leq n$, are passed to the classical processor for quantum state tomography. Third, the classical processor estimates the energy function $\mathcal{L}(\boldsymbol{\theta}) \equiv \langle\psi(\boldsymbol{\theta})|\mathcal{H}|\psi(\boldsymbol{\theta})\rangle$, where $\mathcal{H}$ is the Hamiltonian that encodes a given problem, and searches an optimal parameter $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ that minimizes it. Such optimization is typically done by the local gradient search that requires the iterative evaluation of the energy function and the updated parameter. See [3, 4] for the recent reviews on the VQA algorithms.

At the heart of these VQA approaches lies the variational circuit that generates quantum wavefunctions depending on a set of control parameters stored and manipulated in classical devices. Common choices of unitary gates are usually limited to one-qubit rotation gates and two-qubit entangling gates acting only upon adjacent qubit pairs for the feasibility of hardware implementation [5]. There are numerous ways to design the variational circuits even within this limited class. In most applications, we rely on the heuristic approach to find an effective circuit whose expected performance is not a priori known. The goal of this paper is to bring design principles for an efficient circuit ansatz concerning its entangling capability and number of control parameters, by measuring how these factors influence the quantum energy landscape defined by $\mathcal{L}(\boldsymbol{\theta})$ and the performance of parameter optimization.

The overwhelming majority of the Hilbert space is occupied by highly entangled generic quantum states that exhibit the volume-law scaling of entanglement entropies, i.e., proportional to the number of subsystem qubits. As a result, the broader range of states the circuit ansatz $|\psi(\boldsymbol{\theta})\rangle$ can express, the higher the mean entanglement entropy over randomly sampled states $\{|\psi(\boldsymbol{\theta}_s)\rangle\}_s$ becomes. In this regard, the average entanglement entropy $\mathcal{R}^{(k)}$ of the circuit generated states can represent the expressibility [6] of the variational ansatz.

There is, however, a negative correlation between the average entanglement entropy and the optimization success of the circuit parameters $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$ [7–9]. The area law scaling of the entanglement entropy is a commonly expected correlation pattern of the ground states of local gapped Hamiltonians [10]. In contrast, most of the circuit parameter space is associated with highly-entangled typical quantum states, especially if the mean entanglement entropy of the variational circuit is close to the maximum. That makes the local parameter search of a highly expressible ansatz less likely to succeed in finding a trajectory towards the low-entangled ground states.

A closely related geometric statement is known as the barren plateau theorem: Assuming the 2-design characteristic of the random circuit ensemble, the gradient of the energy function $\mathcal{L}(\boldsymbol{\theta})$ with respect to the circuit variables $\boldsymbol{\theta}$ is zero on average, with the variance exponentially suppressed for the growing system size n [11]. This has been shown in [12] to be equivalent to the exponential decay of $\mathrm{Var}_{\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\alpha}) - \mathcal{L}(\boldsymbol{\theta})]$ with respect to n, indicating how generically flat the quantum energy landscape of the highly expressible variational circuit is. In this work, we will further investigate how flatness of the quantum landscape is correlated with the entangling capability of the circuit by examining the local geometry near generic random points as well as certain special points, which we describe below, extracted from the parameter optimization trajectory.

Since the quantum states generated by the circuit are controlled by continuous variables $\boldsymbol{\theta}$ stored and manipulated in the classical computer, the number of classical parameters can also be a crucial factor that affects the energy landscape and optimization performance. One extreme case was studied in [13]: The local gradient search for the over-parameterized circuit can approximate the Hamiltonian ground state very precisely, in both cases where the ground state entanglement entropy follows the volume-law scaling and the area-law scaling, despite the

high expressibility of the circuit ansatz.[1] Note that while over-parametrization does not lift the flat directions in the parameter space, it makes the convergence faster by developing a few steep directions. See [15] for a landscape analysis of the over-parameterized QAOA ansatz with the MaxCut Hamiltonian. More systematically, given a fixed amount of the average entanglement entropies, one can vary the number of control variables by adding single-qubit Pauli rotation gates. We will quantify how it affects the flatness of the energy landscape and the optimization performance.

Our study will be conducted by investigating the Hessian matrix of the energy function, $H_{ab}(\boldsymbol{\theta}) = \nabla_a \nabla_b \mathcal{L}(\boldsymbol{\theta})$, where a, b denote the circuit parameter indices. The Hessian will be evaluated at random initial points [14], final convergence points [13], and multiple intermediate points chosen from the optimization trajectory [16]. The Hessian eigenvalue spectrum reveals the information about the shape of the quantum energy landscape and how the local parameter search works. We will characterize an important relationship between its eigenvalues and the entangling capability of the circuit: For high-entangling circuits, the Hessian spectrum shows an overall high concentration near 0, while low-entangling circuits develop a few large outliers among a massive bulk of zero eigenvalues. Such spectral pattern is also observed in classical deep neural networks with over-parameterization [17–19]. We will indeed observe the close similarity between the spectral evolution caused by adding more circuit parameters and by reducing the circuit entangling capability.

The rest of the paper is organized as follows: Section II introduces the basic form of the variational circuit used in this paper and analyzes correlation functions of the circuit density matrices. Section III systematically studies how stochastic dropout of two-qubit entangling gates affects the VQA performance of the variational circuit and the quantum energy landscape through the numerical evaluation of the Hessian matrix. It includes theorems on the top Hessian eigenvalue as well as the optimization rate. The impact of the control parameters on the VQA performance and the shape of the quantum energy landscape is studied by adding single-qubit rotation gates in Section IV. Finally, Section V summarizes and provides suggestions for future research.

II. CIRCUIT ANSATZ

A. Circuit Architecture

The parametrized quantum circuit used in this paper for various numerical experiments is a hardware-efficient ansatz [5], made of 1-qubit Pauli-Y rotations and 2-qubit CZ entanglers acting on qubit pairs,

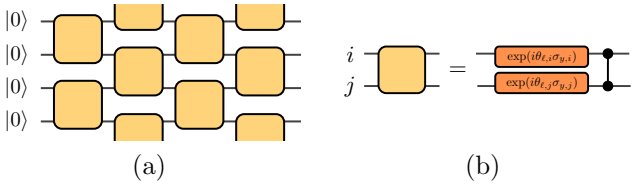


Figure 1. The variational quantum circuit used in this paper. (a) circuit architecture, (b) structure of the two-qubit gate.

$$\begin{aligned} R_{y,i}(\varphi) &= \left[e^{i\sigma_y \varphi}\right]_i, \\ CZ_{i,j} &= \mathrm{diag}(1,1,1,-1)_{i,j}. \end{aligned} \tag{1}$$

The basic building block of our unitary circuit in its primary form is the 2-qubit unitary operator of Figure 1b,

$$U_{i,j}(\varphi_a, \varphi_b) = CZ_{i,j} \cdot \left(R_{y,i}(\varphi_a) \otimes R_{y,j}(\varphi_b)\right), \tag{2}$$

acting on the 4-dimensional hyperplane spanned for the (i,j) qubit pair, embedded in the n-qubit Hilbert space.

The operators (2) acting on consecutive $(i, i+1)$ qubits compose together the following layer unitary operators:

$$U_\ell = \begin{cases} \bigotimes_{m=1}^{\lfloor n/2 \rfloor} U_{2m-1,2m}(\varphi_{\ell,2m-1}, \varphi_{\ell,2m}) & \text{odd } \ell \\ \bigotimes_{m=1}^{\lfloor n/2 \rfloor} U_{2m,2m+1}(\varphi_{\ell,2m}, \varphi_{\ell,2m+1}) & \text{even } \ell \end{cases} \tag{3}$$

where the periodic boundary condition $i \simeq i+n$ is imposed on the n-qubit lattice. It is convenient to use the collective notation $U_\ell(\boldsymbol{\varphi}_\ell)$ where $\boldsymbol{\varphi}_\ell$ denotes all $\{\varphi_{\ell,i}\}_{i=1}^n$. The variational circuit states (see Figure 1a for an illustration of the circuit architecture) are then generated by sequentially acting L instances of the layer unitary operators on the initial product state $|0\rangle^{\otimes n}$, i.e.,

$$|\psi(\boldsymbol{\theta})\rangle = U_L(\boldsymbol{\varphi}_L) \cdots U_1(\boldsymbol{\varphi}_1)|0\rangle^{\otimes n} = U(\boldsymbol{\theta})|0\rangle^{\otimes n}, \tag{4}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\varphi}_\ell\}_{\ell=1}^L$. We will index the nL components of the circuit parameter $\boldsymbol{\theta}$ by $1 \le a, b, \cdots \le nL$.

B. Circuit Density Matrix

The variational circuit generates a pure quantum state (4) whose corresponding density matrix is given by $\rho_{\alpha\beta}(\boldsymbol{\theta}) = U_{\alpha 1}(\boldsymbol{\theta}) U^*_{\beta 1}(\boldsymbol{\theta})$ for all $1 \le \alpha, \beta \le 2^n$. In particular, the circuit unitary matrix $U_{\alpha\beta}(\boldsymbol{\theta})$ of Figure 1 is real-valued and orthogonal, being parameterized by nL circular variables $\{\theta_a\}_{a=1}^{nL}$ that have the period of π [20]. The associated parameter space is therefore the compact torus T^{nL}. We find that the density matrix $\rho_{\alpha\beta}(\boldsymbol{\theta})$ can be written as follows in its Fourier expansion form:

$$\rho_{\alpha\beta} = \frac{\delta_{\alpha\beta}}{2^n} + \sum_q c^q_{\alpha\beta} \prod_{a=1}^{nL} \left(\sin(2\theta_a)\right)^{q_{2a-1}} \left(\cos(2\theta_a)\right)^{q_{2a}} \tag{5}$$

[1] As discussed in many literatures [8–11, 14], it is very challenging, if not impossible, to train highly expressible/entangling circuits *without overparameterization*.

where the sum is taken over the set of $(2nL)$-dimensional discrete vectors, $q \in \{0,1\}^{2nL}$, except the zero $\{0^{2nL}\}$.

The expectation value of the density matrix $\rho_{\alpha\beta}(\boldsymbol{\theta})$ with respect to the uniform measure on $\boldsymbol{\theta} \in T^{nL}$ reads:

$$\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})] = \delta_{\alpha\beta}/2^n \ , \tag{6}$$

where we used the orthogonality of the sine and cosine functions. Consequently, the expectation values of the energy function $\mathcal{L}(\boldsymbol{\theta})$ and its derivatives are given by:

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{\theta})] = \mathrm{Tr}(\mathcal{H})/2^n \ , \tag{7}$$
$$\mathbb{E}_{\boldsymbol{\theta}}[\nabla_a \mathcal{L}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_a \nabla_b \mathcal{L}(\boldsymbol{\theta})] = \cdots = 0 \ . \tag{8}$$

Computation of the second-order correlation functions of the energy function $\mathcal{L}(\boldsymbol{\theta})$ and its derivatives requires the knowledge of two-point functions, $\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\rho\sigma}(\boldsymbol{\theta})]$, of the density matrices.

Theorem 1. *The two-point correlation function of the density matrix $\rho_{\alpha\beta}(\boldsymbol{\theta})$ takes the following structural form:*

$$\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\rho\sigma}(\boldsymbol{\theta})] = A_{\alpha\beta}(\delta_{\alpha\rho}\delta_{\beta\sigma} + \delta_{\alpha\sigma}\delta_{\beta\rho}) + \\ + A_{\alpha\rho}\delta_{\alpha\beta}\delta_{\rho\sigma} \ . \tag{9}$$

There is no summation over repeated indices in (9).

Proof. For real symmetric matrices $\rho_{\alpha\beta}$, the general form of the two-point functions reads:

$$\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\rho\sigma}(\boldsymbol{\theta})] = A_{\alpha\beta}(\delta_{\alpha\rho}\delta_{\beta\sigma} + \delta_{\alpha\sigma}\delta_{\beta\rho}) + B_{\alpha\rho}\delta_{\alpha\beta}\delta_{\rho\sigma}, \tag{10}$$

where the values of the matrices A and B depend on the distribution of $\boldsymbol{\theta}$. It should satisfy the following relations:

$$\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\alpha}(\boldsymbol{\theta})\rho_{\beta\beta}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\beta\alpha}(\boldsymbol{\theta})] \\ = \mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\alpha\beta}(\boldsymbol{\theta})], \tag{11}$$

$$\textstyle\sum_{\beta=1}^{2^n} \mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\beta\sigma}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\sigma}(\boldsymbol{\theta})] \tag{12}$$

where the second equality (12) reflects the purity of the density matrix. By substituting (10) into (11), we obtain the relation $A_{\alpha\beta} = B_{\alpha\beta}$ for $\alpha \neq \beta$. There is a redundancy in keeping both $A_{\alpha\alpha}$ and $B_{\alpha\alpha}$, because only the combination $2A_{\alpha\alpha} + B_{\alpha\alpha}$ appears independently for $\alpha = \beta$. Without loss of generality, we can rewrite (10) as (9) for which (12) becomes equivalent to

$$(3A_{\alpha\alpha} + \textstyle\sum_{\beta\neq\alpha} A_{\alpha\beta})\delta_{\alpha\sigma} = 2^{-n}\,\delta_{\alpha\sigma} \ . \tag{13}$$

□

One can analytically derive the coefficients $A_{\alpha\beta}$ in two limiting cases. When two-qubit entanglers are completely omitted from the circuit ansatz (4), the coefficients are

$$A_{\alpha\beta} = \begin{cases} 3^{c_0(\alpha_2\vee\beta_2)+c_1(\alpha_2\wedge\beta_2)}/2^{3n} & \text{for } \alpha \neq \beta \\ 3^{n-1}/2^{3n} & \text{for } \alpha = \beta \ , \end{cases} \tag{14}$$

where the subscripts 2 in α_2 and β_2 indicate that α and β are in their binary representation. The function $c_{0/1}(s_2)$ gives the number of 0/1's in the input binary string s_2. We also have checked that (14) satisfies (13) by manipulating symbolic expressions in `Mathematica`.

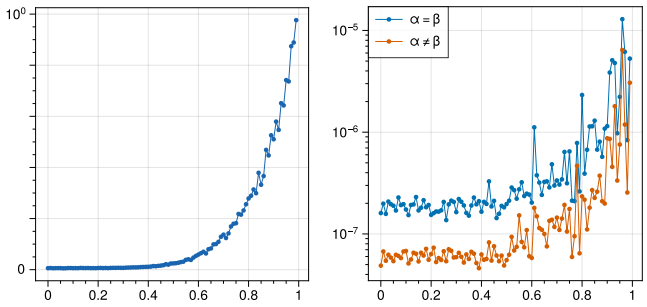


Figure 2. (a) $\mathrm{Var}_{\boldsymbol{\theta}}[\nabla_a \mathcal{L}]$, (b) $\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\alpha\beta}(\boldsymbol{\theta})]$ with 56 circuit layers, estimated from 1500 samples over different probability $p \in \{0.0, 0.1, \cdots, 0.9\}$ that removes the two-qubit entanglers.

Another case that allows the exact analysis is when the circuit distribution is the Haar orthogonal ensemble, for which one can replace the integration over T^{nL} with the Haar integral over the orthogonal group $O(2^n)$. The one-point and two-point functions are obtained by the Weingarten calculus [21] as:

$$\mathbb{E}_{\rho\in O(2^n)}[\rho_{\alpha\beta}] = 2^{-n}\delta_{\alpha\beta} \ , \tag{15}$$

$$\mathbb{E}_{\rho\in O(2^n)}[\rho_{\alpha\beta}\rho_{\rho\sigma}] = \frac{\delta_{\alpha\beta}\delta_{\rho\sigma} + \delta_{\alpha\rho}\delta_{\beta\sigma} + \delta_{\alpha\sigma}\delta_{\beta\rho}}{2^n(2^n+2)} \ . \tag{16}$$

Note that (16) implies the existence of the barren plateau problem [11] for the orthogonal 2-design ensemble. When the variational circuit (4) behaves as an approximate orthogonal 2-design, the variance of random energy gradients decays exponentially with the system size n as:

$$\mathrm{Var}_{\rho\in O(2^n)}[\nabla_a \mathcal{L}] \sim \frac{\mathrm{Tr}(\mathcal{H}^2)}{4^n} \sim \mathcal{O}(2^{-n}) \ , \tag{17}$$

assuming the scaling behavior $\mathrm{Tr}(\mathcal{H}^2) \sim \mathcal{O}(2^n)$ of various 1d spin-chain Hamiltonian systems.

In Section III, we will explore systematic reduction of the average entanglement entropy in the circuit states (4) by randomly and repeatedly removing the CZ entanglers with the probability p. Given sufficient circuit depth, the case with $p = 0$ corresponds to (16) that follows the Haar orthogonal ensemble, while the case with $p = 1$ leads to the n-qubit product state for which we find (14). We can interpolate these two extreme cases through numerical estimation of $\mathrm{Var}_{\boldsymbol{\theta}}[\nabla_a \mathcal{L}]$ and $\mathbb{E}_{\boldsymbol{\theta}}[\rho_{\alpha\beta}(\boldsymbol{\theta})\rho_{\alpha\beta}(\boldsymbol{\theta})]$ for $0 \leq p < 1$. Specifically in the $L = 56$ case, the results are summarized in Figure 2.

III. ENTANGLEMENT, ENERGY LANDSCAPE AND OPTIMIZATION

This section will explore how the shape of the quantum energy landscape varies with different levels of entangling

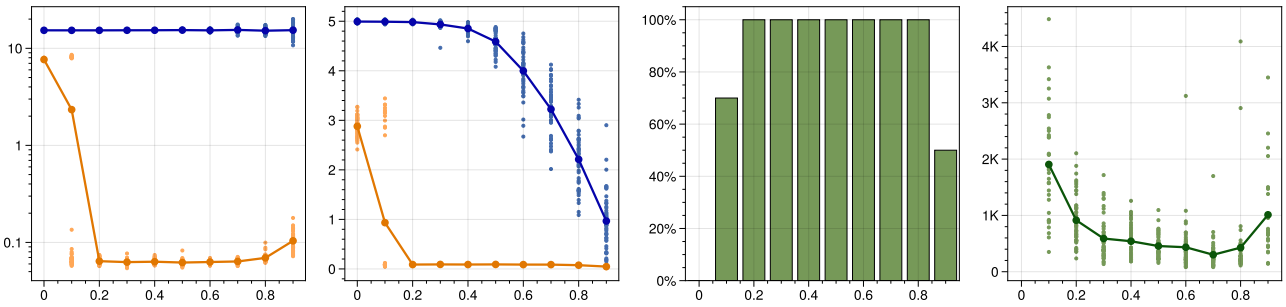


Figure 3. Collections of 50 VQA instances to approximate the $n = 12$ Ising ground state with the circuits with $L = 56$ layers, for each probability p of omitting the CZ gates. The VQA optimizations are successful for $p \in [0.2, 0.8]$. Each subplot displays: (a) energy difference ΔE, (b) Renyi-2 entropy $\mathcal{R}^{(2)}$, (3) % that reaches $\Delta E < 0.1$, (4) number of updates τ to reach $\Delta E < 0.1$.

capability for a fixed circuit architecture. We will consider the average geometry at random generic parameters [11, 14] as well as the local geometry around optimization trajectories [16]. We will fix the number of the circuit control parameters and systematically vary the entangling capability of the circuit. That can be done by dropping out the CZ entanglers contained in the two-qubit operator (2) with probability p. Dialing the drop-out probability p allows us to reach a desired level of the circuit entanglement, by retaining on average

$$m = \tfrac{1}{2} nL(1-p) \tag{18}$$

entanglers in the circuit ansatz of depth L in Figure 1. The $p = 1$ limit gives a non-entangling circuit that maps an initial product state to another product state, while the circuit at $p = 0$ maximally entangles the qubits for a sufficient number L of layers.

A. Circuit State Entanglement

We define the entangling capability of the variational circuit as the average entanglement entropy over the circuit state ensemble, $\{|\psi(\boldsymbol{\theta})\rangle : \boldsymbol{\theta} \in [0, 2\pi)^{\otimes nL}\}$, estimated through the sample average over M circuit states:

$$\frac{1}{M} \sum_{q=1}^{M} \mathcal{R}^{(k)}(|\psi(\boldsymbol{\theta}_q)\rangle) \quad \text{where} \quad \boldsymbol{\theta}_q \sim \mathcal{U}(0, 2\pi)^{\otimes nL} \ . \tag{19}$$

As typical quantum states that comprise an exceedingly large portion of the Hilbert space are highly-entangled, it represents how expressible the variational ansatz is, i.e., how various quantum states $|\Phi\rangle$ can be closely approximated by the circuit ansatz within tolerance ε,

$$\| |\Phi\rangle - |\psi(\boldsymbol{\theta}^*)\rangle \| < \varepsilon, \tag{20}$$

at a certain parameter $\boldsymbol{\theta}^* \in [0, 2\pi)^{\otimes nL}$.

It was shown in [9] that the average entanglement entropy of the dense circuit at $p = 0$ grows *linearly* for an increasing circuit depth L, then saturates to a maximum possible value $n_A - c_k$ beyond a critical depth $L_s < L$. n_A denotes the size of subsystem A, c_k is a non-negative constant that varies with the circuit architecture and the order k of the Renyi entropy,

$$\mathcal{R}^{(k)} \equiv \frac{1}{1-k} \log \mathrm{Tr}\left(\rho_A^k\right) . \tag{21}$$

Since the saturation depth L_s itself scales linearly with the system size n [9], the number of two-qubit CZ gates introduced to reach the entanglement saturation scales as $m_s \sim nL_s = O(n^2)$ for the circuit ansatz in Figure 1. For numerical simulations, we will choose the depth L^* such that the mean number of the CZ gates in the stochastic circuit, i.e., the integral part of $m = \frac{1}{2} nL^*(1-p)$, can be

$$\begin{cases} m \gtrsim m_s & \text{for } p \to 0 \\ m < m_s & \text{for } p \to 1 \ . \end{cases} \tag{22}$$

Specifically, $L^* = 56$ will be sufficient for our purposes.

B. Optimization Accuracy and Speed

To reveal the connection between the entangling capability and VQA performance of the variational circuit, we consider solving the ground state of the most prototypical system, i.e., the 1d transverse-field Ising model:

$$\mathcal{H} = J \sum_{\langle i,j \rangle} Z_i Z_j + g \sum_i X_i \quad \text{with} \quad J = 1,\ g = 1, \tag{23}$$

and measure the deviation of the circuit energy from the exact ground-level energy,

$$\Delta E \equiv \langle \psi(\theta) | \mathcal{H} | \psi(\theta) \rangle - E_g \ . \tag{24}$$

Figure 3 is the collection of 50 independent optimization results for the $L = 56$ circuits with $p \in \{0, 0.1, \cdots, 0.9\}$.

The curves in Figures 3a and 3b are respectively the energy gap ΔE and Renyi-2 entanglement entropy $\mathcal{R}^{(2)}$ evaluated under an equal partitioning of $n = 12$ qubits.

The blue/orange colors (up/down) indicate whether the displayed values are before/after applying the gradient descent (29) to circuit parameters $\tau = 5000$ times. When the average entanglement entropy of the pre-optimization states saturates to the maximum possible value, as the cases for $p \leq 0.1$, Figure 3a exhibits the formulation of orange dot clusters (on the curve below) around $\Delta E \sim 9$. It means the failure of many circuit instances in reducing ΔE via the local gradient search (29). The gradient descent fails to make a trajectory towards the Ising ground state, while stopping at a suboptimal extremum in the quantum energy landscape. It happens because for the circuit with maximum entangling capability, finding the desired parameter $\boldsymbol{\theta}^*$ such that $|\psi(\boldsymbol{\theta}^*)\rangle \simeq |\Psi\rangle$ is roughly as hard as searching the Ising ground state $|\Psi\rangle$ over the entire Hilbert space. In contrast, the circuit with less entangling capability limits the local search to a subregion of the Hilbert space consisting of low-entangled states, which may still include the ground state, thereby facilitating its discovery [9, 22].

Figures 3c and 3d display two complementary metrics about the circuit performance with respect to the VQA optimization, representing how difficult the local gradient descent is to find the circuit parameter $\boldsymbol{\theta}^*$. When the entangling capability of the circuit is too low or too high, the VQA optimization may fail to approximate the ground state, and ΔE does not fall within a tolerance range. Figure 3c shows the sample success rate of VQA trials lying within an acceptable error margin $\Delta E < 0.1$. It not only confirms the failure of maximally entangling circuits in finding the ground state, but also displays the dropping VQA performance of the low-entangling circuits with $p > 0.8$. Also consistent is the minimum number of parameter updates for each successful circuit instance to satisfy $\Delta E < 0.1$, as summarized in Figure 3d. Its mean and variance are minimized at $p = 0.7$, while being larger as p approaches the boundary value of 0.1 or 0.9. These results highlight that the VQA optimization works most efficiently with variational circuits in an intermediate range of the entangling capability [9].

A specific value of the optimal p may vary for a different choice of the depth L or target state $|\Psi\rangle$ that follows the area-law entanglement. Nevertheless, we believe the basic shape of the curves will remain the same, and the circuit with medium expressibility will be most outperforming to approximate low-entangled target states.

C. Hessian Eigenspectrum and Landscape Geometry

We recall from (16) that high entangling capability incurs the barren plateau phenomenon [11, 14, 23] that any partial derivative of the energy function $\mathcal{L}(\boldsymbol{\theta})$ becomes statistically zero (8) on average over $\boldsymbol{\theta} \in T^{nL}$ with an exponentially decaying variance (17) as to the system size n. Since an initial gradient at arbitrarily chosen $\boldsymbol{\theta}$ vanishes with exponentially large probability, the gradient-based optimization typically cannot even start moving towards local minima in the large n limit.

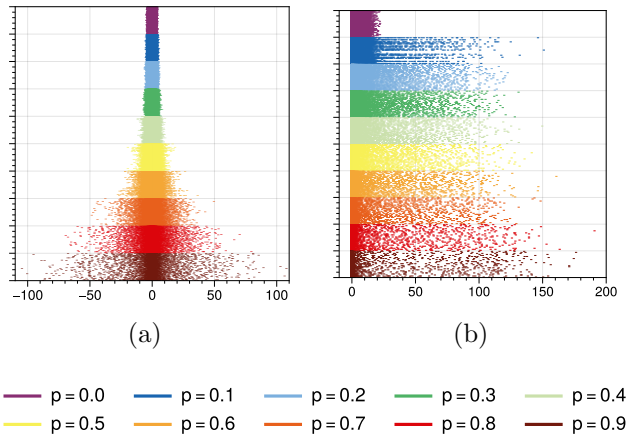


Figure 4. A collection of 500 sample Hessian eigenspectra at $L = 56$ with different probabilities p of omitting CZ-gates, after τ parameter updates. (a) $\tau = 0$, (b) $\tau = 5 \times 10^3$.

However, the systematic control of p can make the circuit ansatz move from the high entangling limit and dramatically improve the VQA performance, as evident in Figure 3. Thus we want to further characterize the geometric implication of entangling capability, i.e., how exactly it eases the barren plateau problem and makes a large impact on the VQA performance. We will compare side-by-side the Hessian eigenspectra of the circuits with a fixed number of parameters and different values of p.

The repeated application of the parameter-shift rule allows us to express a higher-order derivative of the energy function as a combination of the first-order gradients. As a result, [14] found a probabilistic inequality for the second-order derivatives, which in turn leads to the following probabilistic bound on the Hessian eigenvalues.

Theorem 2. *The largest absolute eigenvalue of the Hessian, $H_{ab} = \nabla_a \nabla_b \mathcal{L}(\boldsymbol{\theta})$, is probabilistically bounded as*

$$\Pr(|h_{\max}| \geq c) \leq \frac{2n^2 L^2 \operatorname{Var}_{\boldsymbol{\theta}}(\nabla_a \mathcal{L}(\boldsymbol{\theta}))}{c^2} \ . \tag{25}$$

Proof. Let us denote the Hessian eigenvalues by h_a where $1 \leq a \leq nL$. Since $\operatorname{Tr}(H^2) = \sum_{a,b} H_{ab}^2 = \sum_a h_a^2$,

$$\begin{aligned} |h_{\max}| &\leq \left(\sum_a h_a^2\right)^{1/2} = \left(\sum_{a,b} H_{ab}^2\right)^{1/2} \\ &\leq nL \max_{a,b}(|H_{ab}|) \ , \end{aligned} \tag{26}$$

where $\max_{a,b}(|H_{ab}|)$ is the largest absolute value of the Hessian matrix elements. Using the inequality that holds for every $1 \leq a, b \leq nL$ [14],

$$\Pr(|H_{ab}| \geq c) \leq \frac{2 \operatorname{Var}_{\boldsymbol{\theta}}(\nabla_a \mathcal{L}(\boldsymbol{\theta}))}{c^2} \ , \tag{27}$$

we arrive at the probabilistic bound (25) from (26). □

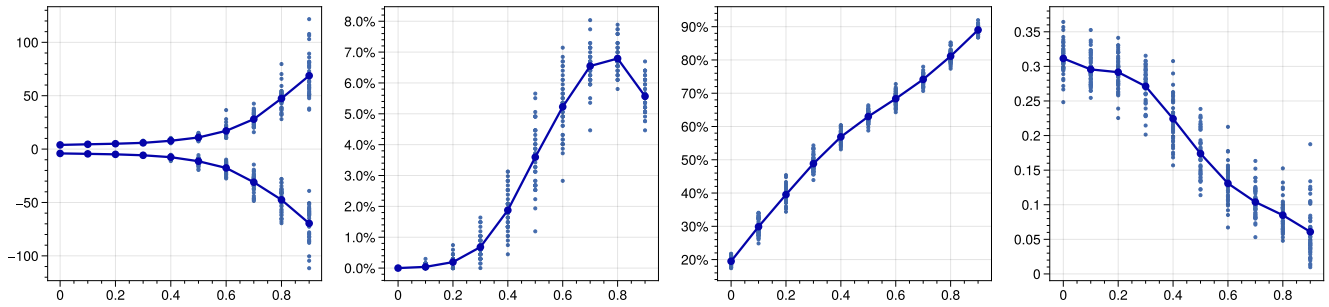


Figure 5. Characteristic plots for the Hessian eigenspectrum with $n = 12$ qubits and $L = 56$ layers, based on 50 instances for each probability p to omit the CZ-gates, at randomly initialized circuit parameters. (a) top/bottom eigenvalues, (b) % of large eigenvalues satisfying $|\lambda| > 5$, (c) % of small eigenvalues satisfying $|\lambda| < 0.2$, (d) gradient overlap with $\mathcal{P}_{\text{small}}$.

Figure 4a visualizes a collection of Hessian eigenspectra evaluated at 50 independent random circuit parameters for each p. Its horizontal axis denotes the eigenvalues, and the vertical axis extends across 500 sample Hessians distinguished by colors according to their p values. Their top/bottom eigenvalues are also depicted in Figure 5a as a function of p. The consequence of the inequality (25) is apparent in both Figures 4a and 5a: While the largest absolute eigenvalues rapidly grow as $p \to 1$, all absolute eigenvalues at $p = 0$ are bounded above as $|h_i| < 5$. Such robust concentration of the Hessian spectrum towards 0 by moving p closer to 0 shows that enhanced entangling capability causes a geometric crossover that smooths all the steepest directions in the quantum energy landscape.

Notably, the tighter concentration of Hessian eigenvalues as $p \to 0$ does not indicate higher degeneracy at zero eigenvalues. Let us count the number of stringent flat directions whose corresponding Hessian eigenvalues satisfy $|h_a| < 0.2$ and denote by $\mathcal{P}_{\text{small}}$ the flat subspace spanned by them. Figure 5c summarizes, for each different p, the average percentage between the dimensionality of $\mathcal{P}_{\text{small}}$ and that of the entire parameter space. Initially at $p = 0$, the flat subspace $\mathcal{P}_{\text{small}}$ makes up only 20% of the full dimensionality of the circuit parameter $\boldsymbol{\theta}$. This percentage steadily increases as p moves towards 1, i.e., reducing the entangling capability of the circuit. It shows that the energy landscape of low-entangling circuits is surprisingly similar to that of over-parameterized systems where most of the variables are used up to parameterize the flat directions. We will observe this geometric resemblance also in the local geometry of intermediate points on the VQA parameter trajectory, serving as the ground for the optimization efficiency of low-entangling circuits. See [24] for how efficiently the gradient descent performs the energy optimization in over-parameterized systems.

It is also informative to examine how aligned an initial gradient vector $\nabla\mathcal{L}(\boldsymbol{\theta})$ is with $\mathcal{P}_{\text{small}}$. For each different p, we estimate the overlap by projecting the normalized gradient onto the subspace $\mathcal{P}_{\text{small}}$ and computing the norm, written as

$$\frac{1}{\|\nabla\mathcal{L}(\boldsymbol{\theta})\|}\sqrt{\sum_{v\in\mathcal{P}_{\text{small}}}\left(v\cdot\nabla\mathcal{L}(\boldsymbol{\theta})\right)^2}, \tag{28}$$

where the sum is over the orthonormal basis v of $\mathcal{P}_{\text{small}}$. Figure 5d exhibits that the lower the entangling capability of the circuit, the smaller the overlap between $\nabla\mathcal{L}(\boldsymbol{\theta})$ and $\mathcal{P}_{\text{small}}$ despite higher dimensionality of $\mathcal{P}_{\text{small}}$. It is another characteristic of the low-entangling circuits, which is also found in over-parameterized classical deep learning systems [19, 25], contributing to the fast convergence of the gradient descent minimization of the energy.

D. Optimization Trajectory

Initial circuit states generated at random parameters, $\boldsymbol{\theta}_0 \sim \mathcal{U}(0, 2\pi)^{\otimes nL}$, can typically be well described by the average characteristics of the quantum energy landscape studied so far. We now turn to investigate the properties of those intermediate circuit states $|\psi(\boldsymbol{\theta}_\tau)\rangle$ obtained after τ steps of the gradient descent update.

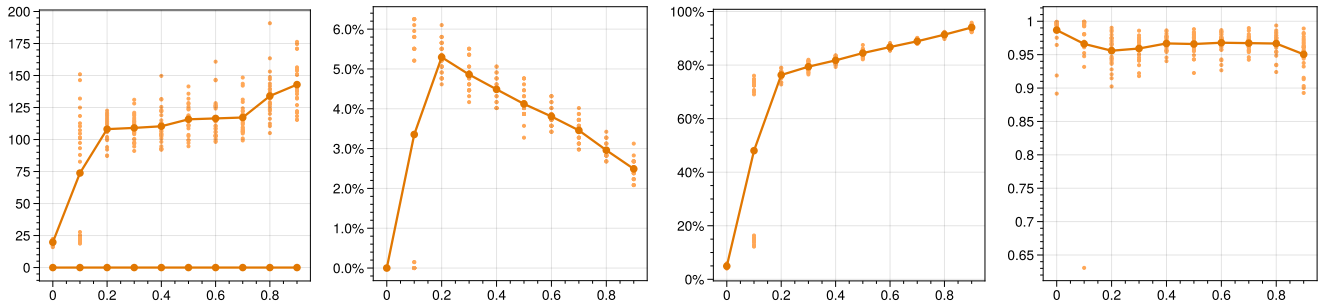
Unlike the initial circuit energy $\mathcal{L}(\boldsymbol{\theta}_0)$ that cannot differ much from the ensemble average (7), the energy $\mathcal{L}(\boldsymbol{\theta}_\tau)$ at an intermediate time τ should significantly deviate almost by definition (29) of the steepest descent method. It indicates how distinctive the intermediate states $|\psi(\boldsymbol{\theta}_\tau)\rangle$ are from initial states, thus requiring independent exploration of their geometric properties.

1. *Optimization Rate*

The gradient descent aims to solve the task of minimizing the energy function by reaching $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ through the iterative and discrete parameter updates,

$$\begin{aligned}\mathbf{v}_{\tau+1} &= \beta\mathbf{v}_\tau + \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\ ,\\ \boldsymbol{\theta}_{\tau+1} &= \boldsymbol{\theta}_\tau + \mathbf{v}_{\tau+1}\ ,\end{aligned} \tag{29}$$

where we set the learning rate η and momentum coefficient β to be $(\eta, \beta) = (0.9, 0.01)$ throughout all numerical



(a) (b) (c) (d)

Figure 6. Characteristic plots for the Hessian eigenspectrum with $n = 12$ qubits and $L = 56$ layers, based on 50 VQA instances for each probability p to omit the CZ-gates, after 5000 steps of the parameter update. (a) top/bottom eigenvalues, (b) % of large eigenvalues satisfying $|\lambda| > 25$, (c) % of small eigenvalues satisfying $|\lambda| < 0.2$, (d) gradient overlap with $\mathcal{P}_{\text{small}}$.

experiments in the paper. Taking the continuum limit, (29) turns into the gradient flow equation, written as [26]

$$(1-\beta)\frac{d\theta_a}{d\tau} = -\nabla_a \mathcal{L}(\theta) \ . \tag{30}$$

Any operator in the system shows no explicit dependence on τ. Therefore, the chain rule implies that

$$(1-\beta)\frac{d}{d\tau} = (1-\beta)\sum_{a=1}^{nL}\frac{d\theta_a}{d\tau}\frac{\partial}{\partial\theta_a} = -(\nabla\mathcal{L}(\boldsymbol{\theta})\cdot\nabla) \ . \tag{31}$$

As a consequence, the average rate of an operator $\mathcal{O}$ along the optimization trajectory is

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{d\mathcal{O}}{d\tau}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\nabla\mathcal{L}(\boldsymbol{\theta})\cdot\nabla\mathcal{O}}{\beta-1}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\mathcal{O}\nabla^2\mathcal{L}(\boldsymbol{\theta})}{1-\beta}\right]. \tag{32}$$

Note that the last equality is obtained after the integration by parts, where the averaging integral over the compact space $\boldsymbol{\theta} \sim T^{nL}$ cannot produce a boundary term.

Theorem 3. *The following statements hold along the optimization trajectory $\{\boldsymbol{\theta}_\tau\}_\tau$ parameterized by discrete integer steps τ:*

(i) *The optimization rate satisfies:*

$$\Pr\left(\left|\frac{d\boldsymbol{\theta}}{d\tau}\right| \geq c\right) \leq \frac{\mathrm{Var}_{\boldsymbol{\theta}}(\nabla\mathcal{L}(\boldsymbol{\theta}))}{c^2(1-\beta)^2} \tag{33}$$

(ii) *The average rate of the energy function is:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{d\mathcal{L}(\boldsymbol{\theta})}{d\tau}\right] = -\frac{1}{1-\beta}\sum_{a=1}^{nL}\mathrm{Var}_{\boldsymbol{\theta}}(\nabla_a\mathcal{L}(\boldsymbol{\theta})) \ . \tag{34}$$

(iii) *The average rate of the energy gradient vanishes:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{d\nabla\mathcal{L}(\boldsymbol{\theta})}{d\tau}\right) = 0 \ . \tag{35}$$

(iv) *The average rates of the Hessian and the higher-order energy derivatives $T_{a_1\cdots a_k}(\boldsymbol{\theta}) = \nabla_{a_1}\cdots\nabla_{a_k}\mathcal{L}(\boldsymbol{\theta})$ are:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{dH_{ab}(\boldsymbol{\theta})}{d\tau}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\sum_{c=1}^{nL}H_{ac}(\boldsymbol{\theta})H_{cb}(\boldsymbol{\theta})}{1-\beta}\right] \tag{36}$$

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{dT_{a_1\cdots a_k}(\boldsymbol{\theta})}{d\tau}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\sum_{c=1}^{nL}T_{a_1\cdots a_{k-1}c}(\boldsymbol{\theta})H_{ca_k}(\boldsymbol{\theta})}{1-\beta}\right] \tag{37}$$

Proof. (i) The inequality follows from the gradient flow equation (30) inserted into the Chebyshev inequality:

$$\Pr(|\nabla_a\mathcal{L}(\boldsymbol{\theta})| \geq c) \leq \frac{\mathrm{Var}_{\boldsymbol{\theta}}(\nabla_a\mathcal{L}(\theta))}{c^2} \ . \tag{38}$$

(ii) Along the optimization curve,

$$\frac{d\mathcal{L}(\boldsymbol{\theta})}{d\tau} = \frac{\nabla\mathcal{L}(\boldsymbol{\theta})\cdot\nabla\mathcal{L}(\boldsymbol{\theta})}{\beta-1} \ . \tag{39}$$

By taking the expectation value $\mathbb{E}_{\boldsymbol{\theta}}$ on both sides of (39), we arrive at (34) thanks to (8) that $\mathbb{E}_{\boldsymbol{\theta}}[\nabla_a\mathcal{L}(\boldsymbol{\theta})] = 0$.
(iii) Along the optimization curve,

$$\frac{d\nabla_a\mathcal{L}(\boldsymbol{\theta})}{d\tau} = \frac{1}{2}\nabla_a\left(\nabla\mathcal{L}(\boldsymbol{\theta})\cdot\nabla\mathcal{L}(\boldsymbol{\theta})\right) \ . \tag{40}$$

The RHS of (40) vanishes upon taking the expectation value $\mathbb{E}_{\boldsymbol{\theta}}$, as it is a total derivative on the compact torus.
(iv) Along the optimization curve,

$$\begin{aligned}\frac{dH_{ab}}{d\tau} &= \frac{\nabla H_{ab}\cdot\nabla\mathcal{L}}{\beta-1} \\ \frac{dT_{a_1\cdots a_k}}{d\tau} &= \frac{\nabla T_{a_1\cdots a_k}\cdot\nabla\mathcal{L}}{\beta-1} \ ,\end{aligned} \tag{41}$$

where we apply the chain rule to (39) and use the commutativity of derivatives. After taking the expectation value $\mathbb{E}_{\boldsymbol{\theta}}$ on (41) and doing the integration by parts, we find (36) and (37), of which the integral over the torus $\boldsymbol{\theta} \sim T^{nL}$ does not make a boundary contribution. □

Alongside the exponential decay (17) of $\mathrm{Var}_{\boldsymbol{\theta}}[\nabla_a \mathcal{L}(\boldsymbol{\theta})]$ in highly expressible circuits [11] and the consequent suppression (27) of the Hessian elements [14], these theorems (34), (36), and (37) highlight the issue of trainability at initial steps τ when circuit-generated states $\{|\psi(\boldsymbol{\theta}_\tau)\rangle\}_\tau$ are similar to the average Haar-random states.

2. Geometry at Endpoints

We now characterize the local geometry near the optimization endpoint $\boldsymbol{\theta}_{5000}$ by collecting 260 sample Hessian eigenspectra evaluated after 5000 steps of the parameter update (29). They are visualized in Figure 4b at a glance, whose horizontal and vertical directions extend across the spectral values and circuit instances, being distinguished by 10 different colors according to the p values. Besides, their top/bottom eigenvalues are also plotted in Figure 6a as a function of p. Having inspected these spectral data, we make the following observations:

First, all negative eigenvalues in the Hessian spectrum are, if not zero, negligible in their absolute values. This illustrates that the local geometry around $\boldsymbol{\theta}_{5000}$ no longer contains a concave direction because the gradient descent has already converged to a local extremum.

Second, the Hessian spectrum at $\boldsymbol{\theta}_{5000}$ distributes more widely as the circuit entangling capability decreases, i.e., $p \to 1$. All the top eigenvalues at $p = 0$ are upper bound by 25, while the top eigenvalues for $p \geq 0.1$ are frequently greater than 100. Such widespread/concentration of the Hessian spectrum correlated to the entangling capability was also observable at the initial points $\boldsymbol{\theta}_0$.

Third, the top Hessian eigenvalues across all $0 \leq p < 1$ can be roughly classified into two clusters, i.e., $h_{\mathrm{top}} \sim 20$ and $h_{\mathrm{top}} \gtrsim 100$, connected to the success/failure of the VQA optimization in approximating the ground state. It is a notable distinction from the initial Hessian spectrum where the top eigenvalues gradually increase from $h_{\mathrm{top}} \sim 5$ to $h_{\mathrm{top}} \gtrsim 70$ as p approaches to 1.

One can also infer the geometric structure of the trajectory endpoint $\boldsymbol{\theta}_{5000}$ by examining the percentages of small and large eigenvalues in the Hessian eigenspectrum. We regard a Hessian eigenvalue h_a as small if $|h_a| < 0.2$ and large if $|h_a| > 25$, such that no eigenvalue at $p = 0$ can be classified as large. The corresponding fractions of large/small eigenvalues are depicted in Figures 6b and 6c. We again find two clusters therein according to the success/failure of VQA samples.

Some circuit instances with high entangling capability, i.e., $p \leq 0.1$, converge to non-optimal extrema, whose local Hessian spectrum contains no large and only a few small values. Hence, the shape of non-optimal extrema is mildly convex and nearly isolated, as there are no steep directions and only a handful of flat directions.

In contrast, the local geometry at the endpoints, $\boldsymbol{\theta}_{5000}$, close to the ground-level energy $E(\boldsymbol{\theta}_{5000}) \simeq E_g$ exhibits the dominance of the flat directions, increasing from 70% to 95% of the total dimension of circuit parameters as p grows. Then the subspace spanned by the large or intermediate eigenvectors makes only up a smaller portion of the parameter space dimensions. Accordingly, the optimal minima should resemble a steep-sided valley with exceedingly high-dimensional flat directions.

We notice the similarity in the geometric structure near the optimization endpoints between the quantum energy landscape with low-entangling capability, i.e., $p \to 1$, and classical over-parameterized systems such as deep neural networks [17]. It is tempting to speculate that the variational circuit with lower entanglement capability can effectively enter the over-parameterized regime with fewer parameters, which explains why the local gradient search can quickly reach the optimal parameter $\boldsymbol{\theta}^*$ that corresponds to the Hamiltonian ground state $|\psi(\boldsymbol{\theta}^*)\rangle \simeq |\Psi\rangle$ [13, 24]. Somewhat tangentially, the effect of adding more circuit parameters while keeping the same amount of the entangling capability will be considered in Section IV.

3. Evolution of the Geometry

We extend the investigation on the Hessian spectrum to various intermediate points along the optimization trajectory $\{\boldsymbol{\theta}_\tau\}_{\tau=1}^{5000}$ at different steps τ. Figure 7 illustrates the evolution of the local geometry shown through a sample collection of 26 Hessian eigenspectra for each $0 \leq p < 1$ at $\tau \in \{a \times 10^b : a \times 10^b \leq 5000 \text{ and } 0 \leq a, b \leq 9\}$. Besides reconfirming how the entangling capability affects the local energy landscape around initial/final points, we also make the following new observations from that:

First, Figures 7a and 7b exhibits the abrupt rise of top eigenvalues followed by the slow adjustment and also the regular convergence of bottom eigenvalues to 0, especially for those successful VQA instances with $\mathcal{L}(\boldsymbol{\theta}_{5000}) \simeq E_g$. It corresponds to the rapid movement of $\boldsymbol{\theta}_\tau$ rolling down into an attractor basin and then fine-tuning itself to minimize the energy $\mathcal{L}(\boldsymbol{\theta}_\tau)$ inside the basin.

Second, Figures 7a shows that the surge of top eigenvalues occurs sooner with the lower entangling capability. For example, the average number of steps τ until the rise of h_{top} reduces from a few hundreds down to a few tens as p grows from 0.2 to 0.7. We remark that the rapid convergence of the gradient descent is another resemblance between the low-entangling circuits and over-parametrized systems [24].

Third, Figures 7c and 7d show the gradual crossover of the landscape geometry into either a steep canyon or a convex bowl along the optimization trajectory, depending on whether the energy at the convergence point $\mathcal{L}(\boldsymbol{\theta}_{5000})$ is sufficiently close to or far from the ground energy E_g.

Fourth, Figures 7e and 7f display a sharp transition in the alignment of the gradient vector $\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ through the norm of the unit gradient projected onto $\mathcal{P}_{\mathrm{small/large}}$ spanned by the small/large Hessian eigenvectors v

$$\frac{1}{\|\nabla\mathcal{L}(\theta)\|}\sqrt{\sum_{v\in\mathcal{P}_{\mathrm{small/large}}}\left(v\cdot\nabla\mathcal{L}(\theta)\right)^2}. \tag{42}$$

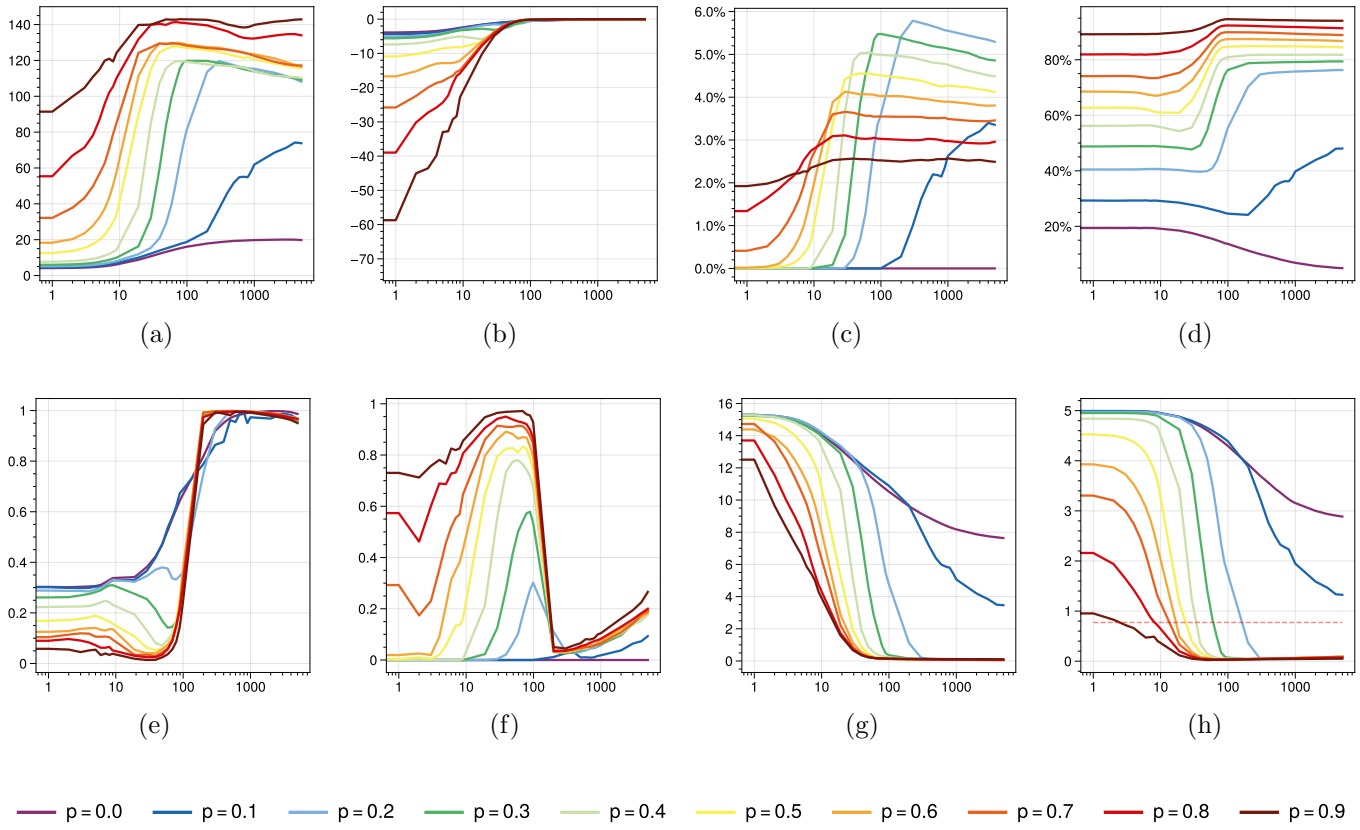


Figure 7. The evolution of the Hessian eigenvalues, the overlap of the energy gradient with the small/large curvature subspaces, the energy gap from the ground state, and the Renyi-2 entropy. (a) top eigenvalue $\lambda_{\max}$, (b) bottom eigenvalue $\lambda_{\min}$, (c) % of large eigenvalues ($|\lambda| > 25$), (d) % of small eigenvalues ($|\lambda| < 0.2$), (e) gradient overlap with $\mathcal{P}_{\text{small}}$, (f) gradient overlap with $\mathcal{P}_{\text{large}}$, (g) energy difference ΔE, (h) Renyi-2 entropy $\mathcal{R}^{(2)}$. Their estimation is based on a collection of 26 Hessian samples for every $0 \leq p < 1$ and $\tau \in \{a \times 10^b : a \times 10^b \leq 5000 \text{ and } 0 \leq a, b \leq 9\}$.

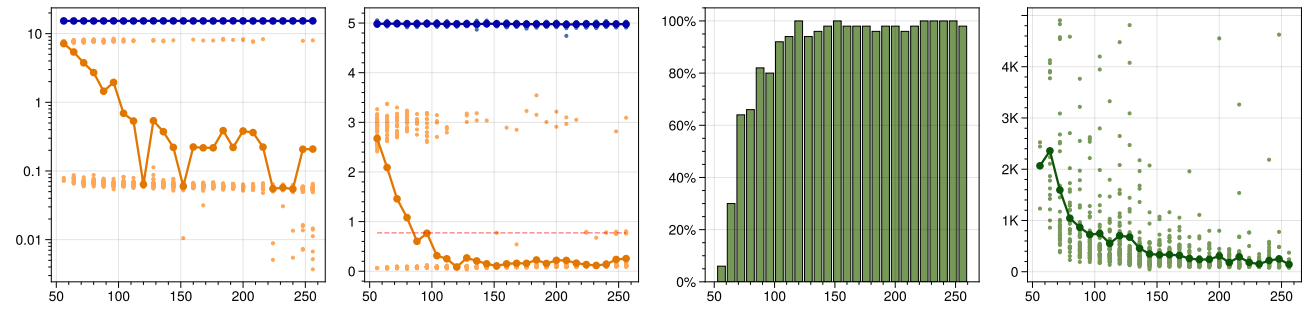


Figure 8. Collections of 50 VQA instances to approximate the $n = 12$ Ising ground state for each circuit depth L that counts both 56 entangling and $(L - 56)$ rotation layers. Adding control parameters with a fixed amount of the entanglement capability improves the optimization performance. Each subplot displays: (a) energy difference ΔE, (b) Renyi-2 entropy $\mathcal{R}^{(2)}$, (3) % that reaches $\Delta E < 0.1$, (4) number of parameter updates τ to reach $\Delta E < 0.1$.

The parameter update only makes $\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ more aligned with $\mathcal{P}_{\text{large}}$ during the initial stage. Afterwards, the gradient overlap with $\mathcal{P}_{\text{small/large}}$ surges/drops suddenly around the transition point $\tau_t \simeq 100$, indicating the initial phase of the gradient descent is over, and the parameter $\boldsymbol{\theta}_{\tau \geq \tau_t}$ has been already confined to an attractor basin. We note that such existence of two phases means the gradient descent is not always aligned with the subspace $\mathcal{P}_{\text{large}}$, not as for the deep neural networks [25].

IV. CONTROL PARAMETERS

This section studies the effect of the number of control parameters on the quantum energy landscape and VQA performance. To run controlled experiments with a fixed degree of the entangling capability, we will always start with the $L^* = 56$ circuits in Figure 1 and introduce extra parameters by randomly sandwiching between the L^* layers n copies of the Pauli-y rotation gate acting on every qubit $(L - L^*)$ times. The total number of circuit parameters is therefore $nL = nL^* + n(L - L^*)$. Notice that the additional parameters are redundant as Pauli-y rotations can commute, i.e.,

$$\begin{aligned} R_{y,i}(\varphi_1)R_{y,i}(\varphi_2) &= R_{y,i}(\varphi_2)R_{y,i}(\varphi_1) \\ &= R_{y,i}(\varphi_1 + \varphi_2) \end{aligned} \tag{43}$$

To see if the enlarged parameter space significantly impacts the VQA performance, we start with 50 random circuit instances and minimize the mean energy $\mathcal{L}(\boldsymbol{\theta})$ of the Ising Hamiltonian (23) by the gradient descent method. All the VQA optimization results for $n = 12$ qubits are summarized in Figure 8 as a function of the total number of layers L. We check that more control parameters, despite their redundancy (43), can still facilitate the local gradient search of the optimal parameters: A cluster of orange dots (on the curve below) near $\Delta E \sim 9$, which represents unsuccessful attempts in reaching the ground-level energy E_g, becomes less populated as L increases. Getting deeper not only increases the rate of optimization success, defined by $\Delta E < 0.1$, but also reduces the parameter update steps τ required for it. Moreover, with exceedingly many parameters, such as $L \gtrsim 200$, the variational circuit can often achieve a high precision approximation that even satisfies $\Delta E \lesssim 10^{-2}$ and $\Delta\mathcal{R}^{(2)} \sim 0$ [13].

Having found the redundant parameters can positively impact the VQA optimization performance, we look into the geometric changes of the quantum energy landscape driven by increasing the parameter space dimension. Let us compute sample collections of 26 Hessian eigenspectra for each $L \in \{56, 64, \cdots, 112, 120, 144, 168\}$ before/after 5000 steps of the gradient descent update. Figure 9 show all the Hessian spectra at a glance, whose horizontal and vertical axes extend over the values and circuit instances, colored differently according to their L values. Some supplementary characteristic curves on the sample Hessian spectra are also presented in Figures 10 and 11, where the upper bounds 5 and 25 of absolute eigenvalues at $L = 56$ are respectively referred to distinguish large eigenvalues before and after the optimization.

Our key observation is that the overall patterns of Figures 4–6 and Figures 9–11 are analogous, demonstrating the qualitative similarity between reducing entanglement capability and adding control parameters: In general, the energy gradient of highly entangling circuits exhibits an exponentially decaying variance. The landscape's flatness manifests in the Hessian eigenspectrum that shows

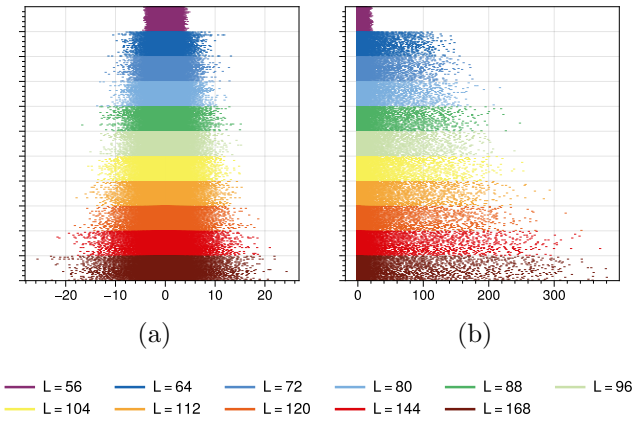


Figure 9. A visualization of 286 sample Hessian eigenspectra, after τ gradient descent updates, for each circuit depth L that counts 56 entangling and $(L - 56)$ rotation layers. (a) $\tau = 0$, (b) $\tau = 5 \times 10^3$. It looks qualitatively similar to Figure 4.

a strong concentration near 0. However, as we increase the parameter space dimension, the gap between the top and bottom Hessian eigenvalues broadens, and the percentage of small or large eigenvalues enlarges. Such evolution in the geometric structure, in turn, improves the optimization performance of the variational circuit. The resemblance that we observe between the energy landscapes of over-parameterized circuits and low-entangling circuits is therefore consistent and somewhat inevitable; One expects the optimization performance to enhance as well by limiting the search scope to a small subset of low entangled states rather than the entire Hilbert space [9]. The low-entangling circuits behave under the gradient descent as if they are over-parameterized since they represent only a limited subset in the Hilbert space.

That implies the following design principle for variational circuits: to avoid reaching both the high entangling capability and over-parameterization if the VQA task involves low-entangled target states. One should rely on highly expressible over-parameterized circuits otherwise.

V. DISCUSSION

Throughout this work, we addressed the design principles for quantum variational circuits by proving several theorems and conducting systematic experiments. We demonstrated how the circuit entanglement and the parameter space dimension affect the local geometry of the energy landscape and thus the VQA optimization performance. Our central object of study was the Hessian of the energy function in the parameter space, which shows the curvature eigenspectrum, evaluated at a random initial point and along the optimization trajectory.

Several analyses on the landscape geometry illustrated that the efficiency of the low-entangling circuits under the VQA optimization is related to their resemblance with an over-parameterized system, despite having a rel-

atively small number of variables. Fewer parameters than 2^n may be sufficient to parameterize a subset of Hilbert space that the circuit with limited expressibility can represent. This leaves us with the following question: How many circuit parameters are optimal for a given amount of entangling capability under the gradient descent.[2]

A study of over-parametrization was carried out in [30], where it was defined as having more circuit parameters than the critical number needed to explore the relevant directions in the state space. The emphasis of [30] was on deriving a bound on the critical dimension, which holds without a particular assumption on the entangling capability of the variational circuit. While our analysis is consistent with that, considering the opposite effect between the enhanced entangling capability and the increased parameter space dimension, we expect finding an optimal point in the trade-off relationship will be a critical task. It includes examining the critical dimension like [30] for non-maximally entangling circuits that can efficiently approximate low-entangled VQA target states.

Another important question unexplored in this paper is the effect of the Hamiltonian on the energy landscape and optimization accuracy. We recall that the low-entangling circuits are not successful in simulating volume-law entangled ground states [9], whose entanglement scaling is actually determined by the Hamiltonian. Generally, we would like to characterize how certain defining properties of the Hamiltonian, e.g., locality or degree of spin interactions, can steepen/flatten the energy landscape and thus influence the VQA performance.

Although all numerical computations in the main text were carried out specifically with the Ising Hamiltonian, we expect the above theorems and empirical observations to hold for generic Hamiltonian systems. The specialty of individual Hamiltonians should appear only in the numerical values of constants. While finding conclusive evidence of Hamiltonian independency will require substantial additional computation beyond the scope of this work, we know there are several supporting evidence that agrees well with the general picture given in this paper: the VQA optimization curve that shows energy vs. gradient descent iteration [7, 9], the demonstration of the attractor basin that shows energy vs. distance in parameter space along the optimization trajectory [13], for both non-local spin-chain model and Sachdev-Ye-Kitaev model of interacting fermions.

Finally, we would like to understand how various types of noise can change the quantum energy landscape [20, 31, 32], which may lead to suitable error mitigation schemes for noisy VQA optimization.

ACKNOWLEDGEMENTS

We thank Kishor Bharti, Thi Ha Kyaw, Dario Rosa for valuable discussions. The work of J.K. is supported by the NSF grant PHY-1911298 and the Sivian fund. The work of Y.O. is supported in part by Israel Science Foundation Center of Excellence, the IBM Einstein Fellowship and John and Maureen Hendricks Charitable Foundation at the Institute for Advanced Study in Princeton. Our Python code for the numerical experiments is written in TensorFlow Quantum [33]. The experimental data are managed by using Comet.ML [34].

[1] 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** no. 1, (Jul, 2014) . http://dx.doi.org/10.1038/ncomms5213.

[2] E. Farhi, J. Goldstone, and S. Gutmann, "A Quantum Approximate Optimization Algorithm," arXiv:1411.4028 [quant-ph].

[3] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, "Hybrid quantum-classical algorithms and quantum error mitigation," *Journal of the Physical Society of Japan* **90** no. 3, (Mar, 2021) 032001. http://dx.doi.org/10.7566/JPSJ.90.032001.

[4] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "Variational quantum algorithms," arXiv:2012.09265 [quant-ph].

[5] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets," *Nature* **549** no. 7671, (Sep, 2017) 242–246. http://dx.doi.org/10.1038/nature23879.

[6] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," *Advanced Quantum Technologies* **2** no. 12, (Oct, 2019) 1900070.

[2] An analysis of [27] that identifies the parameter redundancy can be useful. See also [6, 28, 29] for expressibility-related discussions.

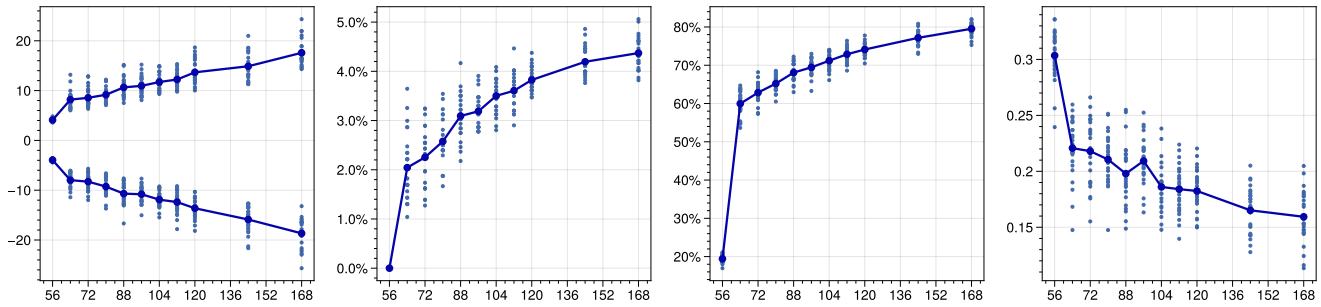


Figure 10. Characteristic plots for the Hessian eigenspectrum with 56 entangling and $(L-56)$ rotation layers, based on 26 instances for each $L \in \{56, 64, \cdots, 112, 120, 144, 168\}$, at randomly initialized circuit parameters. Adding control parameters with a fixed amount of the entanglement capability has qualitatively the same effect as fixing the number of control parameters and reducing the entanglement capability. (a) top/bottom eigenvalues, (b) % of large eigenvalues satisfying $|\lambda| > 5$, (c) % of small eigenvalues satisfying $|\lambda| < 0.2$, (d) gradient overlap with $\mathcal{P}_{\text{small}}$. See Figure 5 that looks qualitatively analogous.

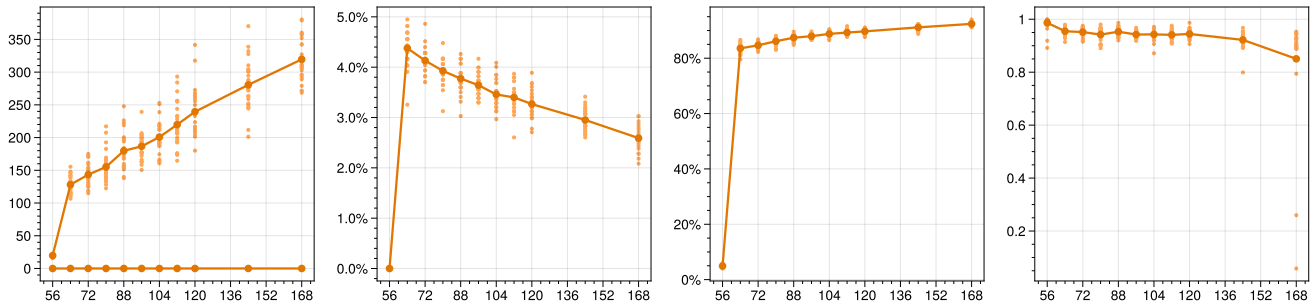


Figure 11. Characteristic plots for the Hessian eigenspectrum with 56 entangling and $(L-56)$ rotation layers, based on 26 VQA instances for each $L \in \{56, 64, \cdots, 112, 120, 144, 168\}$, after 5000 steps of the parameter update. (a) top/bottom eigenvalues, (b) % of large eigenvalues satisfying $|\lambda| > 25$, (c) % of small eigenvalues satisfying $|\lambda| < 0.2$, (d) gradient overlap with $\mathcal{P}_{\text{small}}$. They are qualitatively similar to Figure 6.

https://doi.org/10.1002%2Fqute.201900070.

[7] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, "Exploring entanglement and optimization within the hamiltonian variational ansatz," *PRX Quantum* **1** no. 2, (Dec, 2020) . http://dx.doi.org/10.1103/PRXQuantum.1.020319.

[8] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, "Connecting ansatz expressibility to gradient magnitudes and barren plateaus," arXiv:2101.02138 [quant-ph].

[9] J. Kim and Y. Oz, "Entanglement Diagnostics for Efficient Quantum Computation," arXiv:2102.12534 [quant-ph].

[10] M. B. Hastings, "An area law for one-dimensional quantum systems," *Journal of Statistical Mechanics: Theory and Experiment* **2007** no. 08, (Aug, 2007) P08024–P08024. http://dx.doi.org/10.1088/1742-5468/2007/08/P08024.

[11] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nature Communications* **9** no. 1, (Nov, 2018) . http://dx.doi.org/10.1038/s41467-018-07090-4.

[12] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, "Equivalence of quantum barren plateaus to cost concentration and narrow gorges," arXiv:2104.05868 [quant-ph].

[13] J. Kim, J. Kim, and D. Rosa, "Universal effectiveness of high-depth circuits in variational eigenproblems," *Physical Review Research* **3** no. 2, (Jun, 2021) . http://dx.doi.org/10.1103/PhysRevResearch.3.023203.

[14] M. Cerezo and P. J. Coles, "Impact of barren plateaus on the hessian and higher order derivatives," arXiv:2008.07454 [quant-ph].

[15] J. Lee, A. B. Magann, H. A. Rabitz, and C. Arenz, "Progress toward favorable landscapes in quantum combinatorial optimization," *Physical Review A* **104** no. 3, (Sep, 2021) . http://dx.doi.org/10.1103/PhysRevA.104.032401.

[16] P. Huembeli and A. Dauphin, "Characterizing the loss landscape of variational quantum circuits," *Quantum Science and Technology* **6** no. 2, (Feb, 2021) 025011. http://dx.doi.org/10.1088/2058-9565/abdbc9.

[17] L. Sagun, L. Bottou, and Y. LeCun, "Eigenvalues of the hessian in deep learning: Singularity and beyond," arXiv:1611.07476 [cs.LG].

[18] B. Ghorbani, S. Krishnan, and Y. Xiao, "An investigation into neural net optimization via hessian eigenvalue density," arXiv:1901.10159 [cs.LG].

[19] S. Fort and S. Ganguli, "Emergent properties of the local geometry of neural loss landscapes," arXiv:1910.05929 [cs.LG].

[20] E. Fontana, M. Cerezo, A. Arrasmith, I. Rungger, and P. J. Coles, "Optimizing parametrized quantum circuits via noise-induced breaking of symmetries," arXiv:2011.08763 [quant-ph].

[21] B. Collins and P. Śniady, "Integration with respect to the haar measure on unitary, orthogonal and symplectic group," *Communications in Mathematical Physics* **264** no. 3, (Mar, 2006) 773–795. http://dx.doi.org/10.1007/s00220-006-1554-3.

[22] J. Eisert, M. Cramer, and M. B. Plenio, "Colloquium: Area laws for the entanglement entropy," *Reviews of Modern Physics* **82** no. 1, (Feb, 2010) 277–306. http://dx.doi.org/10.1103/RevModPhys.82.277.

[23] C. O. Marrero, M. Kieferová, and N. Wiebe, "Entanglement induced barren plateaus," arXiv:2010.15968 [quant-ph].

[24] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," arXiv:2003.00307 [cs.LG].

[25] G. Gur-Ari, D. A. Roberts, and E. Dyer, "Gradient descent happens in a tiny subspace," arXiv:1812.04754 [cs.LG].

[26] D. Kunin, J. Sagastuy-Breña, S. Ganguli, D. L. K. Yamins, and H. Tanaka, "Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics," *CoRR* **abs/2012.04728** (2020) , arXiv:2012.04728. https://arxiv.org/abs/2012.04728.

[27] L. Funcke, T. Hartung, K. Jansen, S. Kühn, and P. Stornati, "Dimensional expressivity analysis of parametric quantum circuits," *Quantum* **5** (Mar, 2021) 422. http://dx.doi.org/10.22331/q-2021-03-29-422.

[28] T. Hubregtsen, J. Pichlmeier, P. Stecher, and K. Bertels, "Evaluation of parameterized quantum circuits: on the relation between classification accuracy, expressibility and entangling capability," arXiv:2003.09887 [quant-ph].

[29] S. E. Rasmussen, N. J. S. Loft, T. Bækkegaard, M. Kues, and N. T. Zinner, "Reducing the amount of single-qubit rotations in vqe and related algorithms," *Advanced Quantum Technologies* **3** no. 12, (Oct, 2020) 2000063. http://dx.doi.org/10.1002/qute.202000063.

[30] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, "Theory of overparametrization in quantum neural networks,". https://arxiv.org/abs/2109.11676.

[31] G. S. Barron and C. J. Wood, "Measurement error mitigation for variational quantum algorithms," arXiv:2010.08520 [quant-ph].

[32] J. Zeng, Z. Wu, C. Cao, C. Zhang, S. Hou, P. Xu, and B. Zeng, "Simulating noisy variational quantum eigensolver with local noise models," arXiv:2010.14821 [quant-ph].

[33] M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. H. Yoo, S. V. Isakov, P. Massey, M. Y. Niu, R. Halavati, E. Peters, M. Leib, A. Skolik, M. Streif, D. V. Dollen, J. R. McClean, S. Boixo, D. Bacon, A. K. Ho, H. Neven, and M. Mohseni, "TensorFlow Quantum: A Software Framework for Quantum Machine Learning," arXiv:2003.02989 [quant-ph].

[34] "Comet.ML." https://www.comet.ml/.