# Entanglement devised barren plateau mitigation

Taylor L. Patti,[1,*] Khadijeh Najafi,[2] Xun Gao,[1] and Susanne F. Yelin[1]
[1]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA;*
*IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, New York 10598 USA;*
*and Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, California 91125, USA*

(Received 22 December 2020; accepted 21 June 2021; published 23 July 2021)

Hybrid quantum-classical variational algorithms are one of the most propitious implementations of quantum computing on near-term devices, offering classical machine-learning support to quantum scale solution spaces. However, numerous studies have demonstrated that the rate at which this space grows in qubit number could preclude learning in deep quantum circuits, a phenomenon known as barren plateaus. In this work, we implicate random entanglement, i.e., entanglement that is formed due to state evolution with random unitaries, as a source of barren plateaus and characterize them in terms of many-body entanglement dynamics, detailing their formation as a function of system size, circuit depth, and circuit connectivity. Using this comprehension of entanglement, we propose and demonstrate a number of barren plateau ameliorating techniques, including initial partitioning of cost function and non-cost function registers, meta-learning of low-entanglement circuit initializations, selective inter-register interaction, entanglement regularization, the addition of Langevin noise, and rotation into preferred cost function eigenbases. We find that entanglement limiting, both automatic and engineered, is a hallmark of high-accuracy training and emphasize that, because learning is an iterative organization process whereas barren plateaus are a consequence of randomization, they are not necessarily unavoidable or inescapable. Our work forms both a theoretical characterization and a practical toolbox; first defining barren plateaus in terms of random entanglement and then employing this expertise to strategically combat them.

DOI: 10.1103/PhysRevResearch.3.033090

## I. INTRODUCTION

The rapid development of noisy quantum devices [1] has led to great interest in hybrid quantum-classical variational algorithms, through which classical machine-learning techniques are employed to prepare, sample, and optimize states on noisy quantum hardware [2–6]. Not only do these algorithms show potential for a variety of near-term applications [7], they are inherently robust against certain coherent errors and are free to minimize decoherence effects through the exploration of unconventional gate sequences. Of particular interest are quantum neural networks (QNNs) [8], in which quantum input states are transformed into output states by a parametrized quantum circuit (PQC). The output states then undergo a series of measurements, collectively referred to as a cost function, and the measurement results are used to optimize the circuit.

Although QNNs offer a straightforward approach, their implementation can be quite challenging. Among the greatest of these difficulties are barren plateaus [9]: regions of the cost function's parameter space where it is rather constant, varying too little for successful gradient-based optimization. While in shallow circuits these barren landscapes are cost-function dependent [10,11], the effect is cost-function independent for circuits that are sufficiently deep. Moreover, even gradient-free algorithms can be impacted [12,13]. While certain restricted subsets of PQCs are somewhat resilient to barren plateaus [14,15], the most general implementation, known as the "hardware efficient ansatz," becomes exponentially barren with increasing qubit number. Numerous techniques have been suggested for the amelioration of barren plateaus, including layer-wise and symmetry-based training [16,17], correlated and identity-esq circuit initialization [18,19], and quantum convolutional neural network protocols [20], but they have yet to form a complete toolbox that is suitable for large-scale, general purpose QNNs.

Likewise, our understanding of barren plateaus is extensive yet far from complete. For some years, it has been understood that barren plateaus are a consequence of concentration of measure [21,22], stemming from the effects of randomness on the exponential dimension of quantum state space. More recently, the relationship between entanglement and barrenness has been explored by quantum scrambling studies [23] and in terms of visible and hidden units [24]. However, we still lack comprehensive understanding of how entanglement induces barren plateaus with respect to cost function register

*taylorpatti@g.harvard.edu

size, qubit connectivity, and circuit depth. As a result, barren plateau mitigation strategies that rely on these insights have yet to be developed.

In this work, we give a detailed account of how random entanglement leads to barren plateau formation. We define "random entanglement" as the entanglement that arises as a quantum state undergoes unitary transformation by matrices with randomly sampled, statistically independent rotation angles, such as those routinely employed in quantum circuits [9–15]. In Ref. [9], it was shown that quantum circuits parametrized in this way match the Haar distribution up to at least the second moment with even moderate circuit depth. While barren plateaus can be both noise-independent and noise-induced [25], we consider here only the former variety. In particular, we derive the relationship between cost function barrenness and qubit entanglement, including the rate at which barrenness scales with circuit depth. As our findings quantify barrenness via the entanglement of specific qubit subsets, we develop partitioning methods that initially or continuously restrict such entanglement. This generates nonbarren cost function landscapes and thus improves circuit learning. We find that initially partitioned circuits not only learn faster, but often produce less entangled solutions. Because entangled states are more sensitive to decoherence, this factorizability can decrease the number of measurements required to accurately estimate the cost function, potentially reducing the problematic number of expectation values required for each circuit iteration [26–28].

To verify and exploit these findings, we design a classical meta-learning protocol that avoids barren plateaus while generating an arbitrary circuit with rich entanglement structure. In contrast with other QNN meta-learning proposals [29] which address specific problem classes, ours is suitable for general PQCs. Moreover, because our meta-learning technique does not pretrain circuit output, it is itself immune to barren plateaus. Furthermore, we model a real-time regularization process that penalizes forms of entanglement that are potentially problematic and show that this method ameliorates barren landscapes, decreasing both training time and error. We also make the novel identification of barren plateaus as a form of Langevin noise in the circuit parameter space and demonstrate the effectiveness of injecting additional Langevin noise into the training process, a technique that has been used to combat overfitting in deep classical neural networks [30]. Finally, we draw a parallel between entanglement dynamics and the improved performance of QNNs in certain measurement bases.

## II. VARIATIONAL ALGORITHMS IN LAYERED ONE-DIMENSIONAL QUANTUM CIRCUITS

Before characterizing the relationship between entanglement and barren plateaus, we provide a brief overview of hybrid quantum-classical variational algorithms in one-dimensional (1D) circuits. Examples of such circuits are shown in Figs. 1(a) and 1(b). These circuits have a total number of $n$ qubits partitioned into two registers: the cost function register $\mathcal{R}_C$, whose qubits are measured with some observable $\mathcal{M}_C$, and the non-cost function register $\mathcal{R}_N$, with qubits that

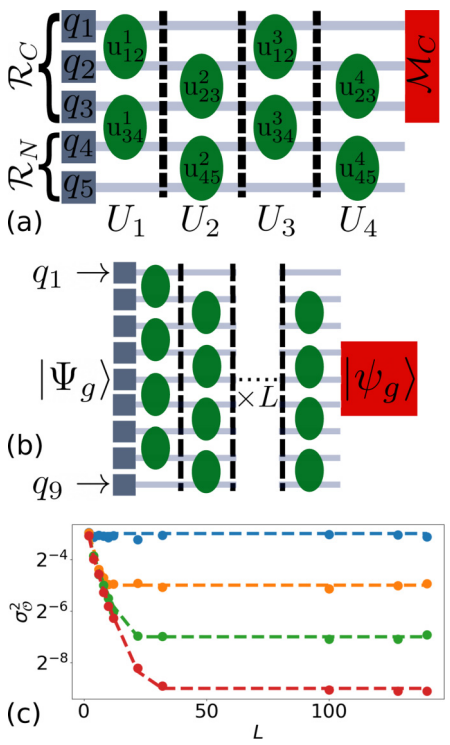


FIG. 1. (a) Diagram of a linear circuit with $n_C = 3$, $n_N = 2$, and $L = 4$. Qubits $q_i$ and $q_j$ interact in layers $k$ through two-qubit unitaries $u_{ij}^k$, which comprise layer unitaries $U_k$, and ultimately form total unitary $U$. The qubits of $\mathcal{R}_C$ are then read out by the cost function operator $\mathcal{M}_C$. (b) Ground-state compressor for randomly generated 9-qubit long-range interaction Hamiltonian [Eq. (8)] ground states. The circuit learns to represent the ground states $|\Psi_g\rangle$ as 3-qubit representations $|\psi_g\rangle$. (c) An illustration of barren plateaus with $\mathcal{L} = \langle \sigma_1^z \sigma_2^z \rangle$ ($n_C = 2$) and $n = 3, 5, 7, 9$ (blue, orange, green, red), with each data point sampling two-thousand distinct circuit iterations. The variance $\sigma_{\mathcal{O}}^2$ of the partial cost function derivative $\mathcal{O}$ is known to decrease rapidly with increasing circuit layers $L$ until ultimately reaching the plateau magnitude $\sigma_B^2 \propto 2^{-n}$.

are not directly measured. These registers have $n_C$ and $n_N$ qubits, respectively, such that $n = n_C + n_N$. In a 1D system, the qubits interact only with their nearest neighbors via two-qubit unitaries, denoted $u_{ij}^k$ for interactions between the $i$th and $j$th qubit in layer $k$. As this work considers pure states, all wave functions respect time-reversal symmetry and can be generally parametrized as completely real. This simplification renders the subset SO(4) of SU(4) to be fully expressive for any two-qubit unitary $u_{ij}^k$ [31], such that it can be fully described with six rotation angles $\hat{\theta}_{ijk} = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]$ as [32]

$$u_{ij}^k(\hat{\theta}_{ijk}) = R_{34}(\theta_6) R_{23}(\theta_5) R_{12}(\theta_4) R_{34}(\theta_3) R_{23}(\theta_2) R_{34}(\theta_1), \tag{1}$$

where $R_{ij}(\theta)$ is a sinusoidal rotation matrix on axes $i$ and $j$ that can be expressed as $R_{ij}(\theta) = \exp(-i\theta K_{ij})$. We initialize the circuits by normally and independently assigning each $\theta$ from the interval $[0, 2\pi]$. Here, $K_{ij}$ is a Hermitian matrix that is equal to $\pm i$ at elements $ij$ and $ji$ and zero elsewhere.

The universal nature of this parametrization distinguishes our work from studies that impose a restricted unitary structure [14,15] and ensures that, for sufficient depth, our unitaries are random enough to generate barren plateaus [33,34]. We note that the $\theta_i$ which correspond to $u^k_{n_C,n_C+1}$ are especially significant, as they entangle registers $\mathcal{R}_C$ and $\mathcal{R}_N$, and we denote them $\theta^E_i$ when relevant.

These two-qubit interactions are then organized into full layer unitaries

$$U_k = \prod_{m=0}^{(n-1)/2} u^k_{q+2m,q+2m+1}, \tag{2}$$

where $q$ is the remainder of $k/2$. As all interactions are pairwise, the $u^k_{ij}$ in each single-layer unitary $U_k$ commute.

We describe the unitary of the full system as

$$U = \prod_{i=1}^{L} U_i, \tag{3}$$

with total number of gate layers $L$. Figure 1(a) illustrates a generic example of such a circuit for $n = 5$, $n_C = 3$, and $L = 4$.

In hybrid quantum-classical variational algorithms, circuit training is described by a cost function $\mathcal{L}$, which is some function $f$ of expectation value

$$\langle \mathcal{M}_C \rangle = \langle \psi_{\text{out}} | \mathcal{M}_C | \psi_{\text{out}} \rangle, \tag{4}$$

such that $\mathcal{L} = f[\langle \mathcal{M}_C \rangle]$ and where $|\psi_{\text{out}}\rangle = U|\psi_{\text{in}}\rangle$ is the output of the quantum circuit $U$ with input state $|\psi_{\text{in}}\rangle$. Unless otherwise specified, we take $|\psi_{\text{in}}\rangle = |0\rangle$. The classical learning algorithm then minimizes $\mathcal{L}$ by updating the parameters $\theta_i$ through use of the partial derivatives

$$\mathcal{O}_i = \frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{\partial f}{\partial \langle \mathcal{M}_C \rangle} \frac{\partial \langle \mathcal{M}_C \rangle}{\partial \theta_i}. \tag{5}$$

Figure 1(b) illustrates a specific learning example: a ground-state compressor. The ground-state compressor is a circuit that takes $n = 9$ qubit ground states $|\Psi^g_i\rangle$ and their average $z$-axis magnetization

$$\langle M_i \rangle = \langle \Psi^g_i | \frac{1}{n} \sum_i^n \sigma^z_i | \Psi^g_i \rangle, \tag{6}$$

as training data, where $\sigma^b_i$ is the Pauli operator along axis $b$ acting on qubit $i$. The circuit then learns to compress $|\Psi^g_i\rangle$ into $n_C = 3$ qubit equivalents $|\psi^g_i\rangle$ in the $x$ basis by using their average $x$-axis magnetization

$$\langle m_i \rangle = \langle \psi^g_i | \frac{1}{n_C} \sum_i^{n_C} \sigma^x_i | \psi^g_i \rangle \tag{7}$$

as training labels. Here, we generate $N_g$ different $|\Psi^g_i\rangle$ from randomly parametrized long-range interaction Hamiltonians

$$H = \sum_{i,j=1}^{9} \left( J^z_{ij} \sigma^z_i \sigma^z_j + J^x_{ij} \sigma^x_i \sigma^x_j \right) + \sum_{i=1}^{9} \left( w_i \sigma^x_i + v \sigma^z \right), \tag{8}$$

with uniformly sampled variables $J^z_{ij} \in [-1, 0]$, $J^x_{ij} \in [-1, 1]$, $w_i \in [-0.04, 0.04]$, and $v \in [-6, 6]$, chosen so as to study the effects of barren plateaus in an otherwise successful quantum machine-learning algorithm [31]. In this case, $\langle \mathcal{M}_C \rangle$ is a series of $\langle m_i \rangle$ and we choose $\mathcal{L}$ as the L1 loss between the training output and labels

$$\mathcal{L}_g = \sum_i^{N_g} |\langle m_i \rangle - \langle M_i \rangle|. \tag{9}$$

This circuit is an extension of that used in Ref. [31]. We remark that this task is inherently global, requiring magnetization information from both $\mathcal{R}_C$ qubits 4–6 and $\mathcal{R}_N$ qubits 1–3 and 7–9.

## III. THE EFFECT OF ENTANGLEMENT ON BARREN PLATEAUS

Barren plateaus are a manifestation of concentration of measure [35], meaning that they arise from the tendency of high-dimensional, random distributions to cluster about their mean. In a PQC, the measurement expectation value $\langle \mathcal{M}_C \rangle$ of the quantum circuit is determined by parameters $\theta_i$. For random circuit initialization, as the number of these parameters grows, the impact of the individual parameter uncertainties becomes small and, for the vast majority of parameter sets $\theta_i$, $\langle \mathcal{M}_C \rangle$ approaches its mean with very low variance such that $\frac{\partial \langle \mathcal{M}_C \rangle}{\partial \theta_i} \to 0$. In the interest of building intuition, we can draw an analogy between the collective effects of parameters $\theta_i$ on $\langle \mathcal{M}_C \rangle$ and the behavior of an average of $M$ Gaussian distributions $X = \frac{1}{M} \sum_i^M \mathcal{N}_i(\mu, \sigma^2)$, where $\mathcal{N}_i$ are Gaussian distributions with mean $\mu$ and variance $\sigma^2$. Assuming that all $\mathcal{N}_i$ are independent, the uncertainty of individual $\mathcal{N}_i$ are washed out and $X = \mathcal{N}(\mu, \sigma^2/M)$. That is, the probability that $X$ deviates from $\mu$ vanishes polynomially in $M$.

We emphasize that both concentration of measure and the barren plateaus that they produce are a product of randomness in large-dimensional systems, not large dimensionality alone. For this reason, barren plateaus are typically discussed in the context of random PQCs and quantified in terms of unitary $t$ designs [34,36,37], or probability distributions that approximate the average of polynomial functions of degree $\leqslant t$. Figure 1(c) illustrates the characteristic behavior of these features, as detailed in Ref. [9]. Figure 1(c), like all numerical quantities in this work, is calculated with data from two-thousand distinct random circuit parametrizations. As previously explained, such randomly parametrized circuits have statistical moments that match those of the Haar distribution up to at least second order (unitary 2-design) for even moderate circuit depth [9]. We define the mean of $\mathcal{O}_i$ over the probability distribution of all Haar random unitary matrices $U$ as $\mu_{\mathcal{O}_i}$. Assuming that $U$ is at least as random as a quantum 1-design, $\mu_{\mathcal{O}_i} = 0$ [9] and the training dynamics rely solely on the variance of this quantity. For relatively shallow circuit depth $L$, it is known that the unitary approaches a quantum 2-design [34]. As such, the variance of the gradient with respect to this unitary ensemble $\sigma^2_{\mathcal{O}_i} = \text{var}(\mathcal{O}_i)$ decreases rapidly in $L$, ultimately reaching the steady-state 2-design value $\approx 2^{-n}$ [9]. As $n$ becomes large, randomly initialized circuit parameters cease to update and training fails. In what follows, we use the omitted subscript $\mathcal{O}$ to refer in general to arbitrary parameters $\theta_i$, using the subscripted version $\mathcal{O}_i$ to specify only when the

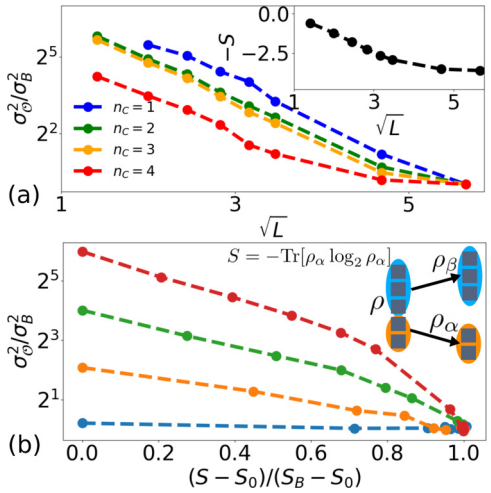


FIG. 2. (a) Variance $\sigma^2_{\mathcal{O}}$ of cost function derivative $\mathcal{O}$ for $\mathcal{L} = \langle \prod_{i=1}^{n_C} \sigma_i^z \rangle$ in units of its barren plateau value $\sigma^2_B$ [horizontal asymptote for each $n$ in Fig. 1(c)] vs number of gate layers, $L$, for $n = 9$ and various $n_C$. The inset shows the change in entanglement entropy $S$ vs circuit depth $L$ for a 2–3 qubit bipartition, as defined in Eq. (11) and illustrated in the inset of panel (b). As the $n_C = 4$ system has an entropy bipartition which is exactly aligned with the $\mathcal{R}_C$-$\mathcal{R}_N$ partitioning, $S$ describes the entanglement growth analytically for this case, while it is only an approximation for $n_C = 1, 2, 3$. This approximation could be improved by using an initial $n_C$-$n_N$ bipartition. (b) Variance $\sigma^2_{\mathcal{O}}$ of cost function derivative $\mathcal{O}$ in units of its barren plateau value $\sigma^2_B$ vs normalized change in entropy $S$ for $\mathcal{L} = \langle \sigma_1^z \sigma_2^z \rangle$ ($n_C = 2$) and $n = 3, 5, 7, 9$ (blue, orange, green, and red). Larger values of $n$ experience greater relative suppression of $\sigma^2_{\mathcal{O}}$ as $\sigma^2_{\mathcal{O}}/\sigma^2_B \propto S_B = 2^{-S}$, which is the analytical solution for $n = 5$ (orange) and an approximation for other $n$. The inset shows a schematic of bipartition entropy of entanglement $S$ for the smaller system $n = 5$. Full density matrix $\rho$ broken into two subsets $\mathcal{R}_\alpha$ and $\mathcal{R}_\beta$, where $\mathcal{R}_\alpha$ always contains as much of $\mathcal{R}_C$ as possible. In all subfigures, each data point sampling two-thousand distinct circuit iterations.

distinction is relevant. For instance, our numerical data are calculated with $\mathcal{O}_1$.

To intuitively understand how random entanglement causes barren plateaus, we point out that, for a randomly initialized parameter $\theta_i$ to contribute to the concentration of $\langle \mathcal{M}_C \rangle$ and thus to the vanishing of $\mathcal{O}$, it must have some form of influence over the qubits of $\mathcal{R}_C$. For the qubits of $\mathcal{R}_N$, this interaction occurs via $U$ and results in entanglement between the two registers. According to this reasoning, barren plateau emergence should be proportional to the spread of random entanglement. Figure 2(a) shows the emergence of barren plateaus vs circuit depth $L$. As $L$ increases, $\sigma^2_{\mathcal{O}}$ decreases exponentially with $\sqrt{L}$ until approaching its asymptotic limit $\sigma^2_B$. While shallow circuits with smaller cost function registers $\mathcal{R}_C$ initially enjoy greater $\sigma^2_{\mathcal{O}}$, $n$ determines $\sigma^2_B$ for deep circuits and we will later conjecture that this asymptote corresponds to entanglement saturation between all qubits on the random circuit. As circuit depth is a form of discretized interaction time $\tau$, this scaling is equivalent to the $\tau$ dependence of late-time entanglement growth of two-level quantum systems in 1D [38–41].

To describe these entanglement dynamics quantitatively, we consider the density matrix of the output qubits

$$\rho = |\psi\rangle\langle\psi| = U|0\rangle\langle 0|U^\dagger. \tag{10}$$

In a compromise between simplicity and generality, in this work we describe the spread of circuit entanglement with the bipartite entanglement entropy

$$S = -\mathrm{Tr}[\rho_\alpha \log_2 \rho_\alpha], \tag{11}$$

where $\rho_\alpha$ is the reduced density matrix of $(n-1)/2$ connected qubits of register $\mathcal{R}_\alpha$, taken so as to contain as many cost function qubits as possible. The remaining $(n+1)/2$ qubits are in $\mathcal{R}_\beta$, such that $\rho_\alpha = \mathrm{Tr}_\beta[\rho]$, as illustrated in the inset of Fig. 2(b). We here partition the qubits as $\mathcal{R}_\alpha$ and $\mathcal{R}_\beta$ to discuss the dynamic spread of entanglement during state evolution, as opposed to the static specification of $\mathcal{R}_C$ and $\mathcal{R}_N$, which are determined by the measurement scheme. For pure states, this entropy is symmetric and $S = -\mathrm{Tr}[\rho_\beta \log_2 \rho_\beta]$ is equivalent. As the bipartite entanglement entropy measures the entanglement between $\mathcal{R}_\alpha$ and $\mathcal{R}_\beta$, it is most accurate in describing the spread of entanglement between $\mathcal{R}_C$ and $\mathcal{R}_N$ when $\mathcal{R}_C = \mathcal{R}_\alpha$. Figure 2(a) displays $\sigma^2_{\mathcal{O}}$ vs circuit depth for a variety of $n_C$ in an $n = 9$ system. While all $n_C$ scale roughly as $2^{-S}$ (inset), $n_C = 4$ is most accurately characterized as, for that case, $\mathcal{R}_C = \mathcal{R}_\alpha$.

Thus, while a single such partitioning is adequate for describing entanglement spread in configurations with $|\mathcal{R}_C| \sim |\mathcal{R}_\alpha| = (n-1)/2$, various such partitions may be used to track short-term entanglement growth when $|\mathcal{R}_C| \ll |\mathcal{R}_\alpha|$ or long-term entanglement growth when $|\mathcal{R}_C| \gg |\mathcal{R}_\alpha|$, such that the entanglement entropy does not temporarily stagnate, like Fig. 2(b), $n = 9$ (red), or rapidly saturate, such as Fig. 2(b), $n = 3$ (blue). The plot is scaled from initial entanglement $S_0$ and normalized to asymptotic difference $S_B - S_0$. In particular, $n = 3$ (blue) is initially saturated because it is nearly fully entangled with the minimal number of gates $L = 2$, while as $|\mathcal{R}_C| < |\mathcal{R}_\alpha|$, $n = 7, 9$ (green, red) have superlogarithmic scaling for $S \to S_B$. At the expense of computational simplicity, more general metrics could be adopted, such as a $n_N$-fold sum of bipartite mutual information $I_2$

$$S_N = \sum_{q \in \mathcal{R}_N} I_2(\mathcal{R}_C, \mathcal{R}_q), \tag{12}$$

where $\mathcal{R}_q$ is the single-qubit subspace for each qubit $q \in \mathcal{R}_N$.

We now derive the relationship between $S$ and $\sigma^2_{\mathcal{O}}$. In particular, we consider $\mathcal{R}_E$, the subspace of qubits that are either directly measured by, or entangled with the qubits measured by (causal to) the cost function. We begin by proving that $\sigma^2_{\mathcal{O}}$ is dependent on the dimension $d_E$ of $\mathcal{R}_E$ and then establish the link between $d_E$ and $S$. Let us assume that the circuit input is a product state, here specifically $|0\rangle$. Then the output state $\rho = |\psi\rangle\langle\psi|$ can be written as $\rho = \rho^E \otimes \rho^D$, where $\rho^E$ belongs to $\mathcal{R}_E$ (measured in $\mathcal{M}_C$ or entangled with the qubits that are) and $\rho^D$ does not. For simplicity, in the following proofs we assume that a given qubit is either completely entangled or disentangled. A similar result for the more general case of partial entanglement follows straightforwardly by taking general $|\psi\rangle = \sum_i c_i |\psi_E^i\rangle |\psi_D^i\rangle$, following the above steps, and

repartitioning each component of the sum

$$\rho = \sum_{i,j} c_i c_j^* |\psi_E^i\rangle|\psi_D^i\rangle\langle\psi_E^j|\langle\psi_D^j| \tag{13}$$

into $|\psi_{E'}^k\rangle\langle\psi_{E'}^k| \otimes |\psi_{D'}^k\rangle\langle\psi_{D'}^k|$ such that the dimension of $|\psi_{E'}^k\rangle$ is maximal under the factorization constraint. The factorization $\rho = \rho^E \otimes \rho^D$ implies that

$$\rho^E \otimes \rho^D = U|0\rangle\langle 0|U^\dagger \to U = U_E \otimes U_D, \tag{14}$$

where $U_E|0_E\rangle\langle 0_E|U_E^\dagger = \rho^E$ and $U_D|0_D\rangle\langle 0_D|U_D^\dagger = \rho^D$. Then, the expectation value of our observable $\langle\mathcal{M}_C\rangle$ becomes

$$\langle 0_E|U_E^\dagger \mathcal{M}_C U_E|0_E\rangle\langle 0_D|U_D^\dagger U_D|0_D\rangle = \langle 0_E|U_E^\dagger \mathcal{M}_C U_E|0_E\rangle, \tag{15}$$

rendering its derivative $\frac{\partial\langle\mathcal{M}_C\rangle}{\partial\theta_i}$ for $\theta_i$ in layer $l$

$$\frac{\partial\langle\mathcal{M}_C\rangle}{\partial\theta_i} = i\langle 0_E|U_R^\dagger[K, U_L^\dagger \mathcal{M}_C U_L]U_R|0_E\rangle, \tag{16}$$

where $U_R$ and $U_L$ are the products of unitaries $U_k$ for $k < l$ and $k \geqslant l$, respectively and where $K$ is the rotation generator for $\theta_i$. Assuming a randomly initialized circuit, this reduces the problem to that of Ref. [9], where using the Haar measure it is shown that $\mu_{\mathcal{O}_i} = 0$ with respect to any $\theta_i$, with variance $\sigma_{\mathcal{O}}^2 \sim 1/d_E$, where $d_E = 2^{n_E}$ is the dimensionality of the entangled subspace $\rho^E$. We contrast this with the barrenness of a fully entangled circuit $\sim 1/d = 2^{-(n_E+n_D)} = 2^{-n}$, which can, for many applications, be numerous orders of magnitude smaller.

To implicate $S$ in this barren plateau process, we note that, for $\rho = \rho^E \otimes \rho^D$,

$$S = -\mathrm{Tr}[\mathrm{Tr}_\beta[\rho^E \otimes \rho^D]\log_2(\mathrm{Tr}_\beta[\rho^E \otimes \rho^D])]. \tag{17}$$

Given that, if $n_E < n_\alpha$, we need to describe early entanglement spread with a smaller bipartition of $S$, we can assume that $\rho^D$ is fully contained in $\rho_\beta$ such that $\mathrm{Tr}_\beta[\rho^E \otimes \rho^D] = \mathrm{Tr}_{\beta_E}[\rho^E]\mathrm{Tr}[\rho^D] = \mathrm{Tr}_{\beta_E}[\rho^E]$, where $\beta_E$ is the entangled portion of $\mathcal{R}_\beta$. Then

$$\begin{aligned} S &= -\mathrm{Tr}[\mathrm{Tr}_{\beta_E}[\rho^E]\log_2 \mathrm{Tr}_{\beta_E}[\rho^E]] \\ &= -\mathrm{Tr}[(\rho^E)_\alpha \log_2(\rho^E)_\alpha] = n_{\beta_E} = n_E - n_\alpha, \end{aligned} \tag{18}$$

because this is precisely the definition of the number of entangled qubits shared between $\rho_\alpha$ and $\rho_\beta$. Then, the total number of $\mathcal{R}_C$ entangled qubits is $n_{\beta_E} + n_\alpha = n_E$ such that $d_E = 2^{n_E}$. Therefore, $\sigma_{\mathcal{O}}^2 \propto 2^{-n_E}$ and changes proportionally to $2^{-S}$.

This mapping between the number and degree of cost function-entangled qubits and plateau barrenness highlights that circuit connectivity, and not simply overall circuit depth, is an accurate indicator for the barrenness of the training landscape. Figure 3(a) is a proof of principle illustration of this point. We remove the register connecting the $u_{2,3}^k$ (unitaries between qubits $q_2$ and $q_3$) gates from each layer $k$ for circuits where $n_C = 2$ and $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\rangle$, permanently separating, or partitioning, the registers $\mathcal{R}_C$ and $\mathcal{R}_N$. As circuit depth grows, entanglement with the qubits of $\mathcal{R}_N$ suppresses $\sigma_{\mathcal{O}}^2$ much faster than its partitioned counterpart $\sigma_{\mathcal{O}\mathcal{P}}^2$, which never exceeds the variance of circuit of total $n = 2$ and is therefore numerous orders of magnitude larger than the variance $\sigma_{\mathcal{O}}^2$ of the fully entangled system. While insightful, permanent partitioning is clearly not a practical solution for barren plateaus because it limits not only the barrenness but also the *expressibility* of the circuit to that of only $n_C$ qubits.

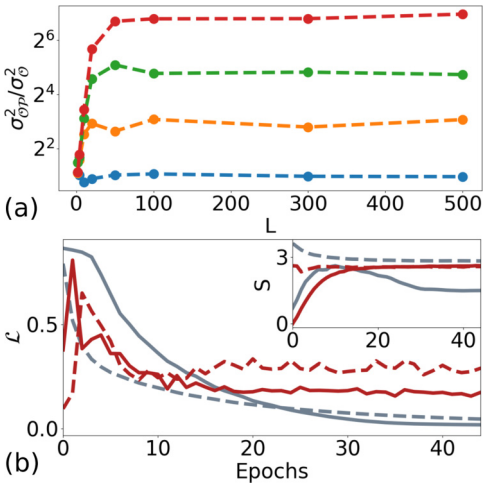


FIG. 3. (a) Partitioned cost function derivative variance $\sigma_{\mathcal{O}\mathcal{P}}^2$ for $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\rangle$ in units of its nonpartitioned value $\sigma_{\mathcal{O}}^2$ vs number of gate layers $L$ for $n = 3, 5, 7, 9$ (blue, orange, green, and red) and $n_C = 2$. As expected, $\sigma_{\mathcal{O}\mathcal{P}}^2$ is approximately a factor of $2^{n_N}$ larger than $\sigma_{\mathcal{O}}^2$ due to the absence of random $\mathcal{R}_C$-$\mathcal{R}_N$ entanglement. (b) Training loss $\mathcal{L} = \mathcal{L}_g$ of Eq. (9) for ground-state compressor in Fig. 1(b) (gray) and $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ (red) vs training epochs for $L = 200$. Solid lines are initially partitioned circuits and dashed lines are fully random initializations, with increased learning performance in the latter. The corresponding evolution of $S$ is shown in the inset. In both subfigures, each data point sampling two-thousand distinct circuit iterations.

The various methods of barren plateau mitigation presented throughout this work serve to regularize, either explicitly or implicitly, the circuit cost function by moderating the entanglement generated by random unitaries. This cost function regularization can also be seen as a penalization of high connectivity in the circuit, such that qubit connectivity is selectively attenuated when not providing outsized benefit to cost function minimization.

## IV. INITIALIZATION TECHNIQUES FOR BARREN PLATEAU MITIGATION

While permanent partitioning is tantamount to simply employing a circuit of smaller $n$, initial parameter restrictions can improve circuit trainability without reducing circuit expressibility. Intuitively, the advantages of this method stem from the role of entropy as a thermodynamic arrow that drives statistical processes forward. As a toy example, let us imagine a classical machine-learning protocol where we would like to create a gaseous mixture with optimized concentrations of two gases. If we initially partition the gases, the learning algorithm can simply allow the gases to mix themselves by passing through a vent in the partition, sealing the vent when the ideal concentration is reached on one side. This process occurs independently, driven forward by entropic considerations. If, however, the gases are initially mixed and therefore have a maximum entropy configuration, the learning algorithm

cannot succeed by simply unsealing a vent. The problem has been complicated and learning will fail unless more heroic measures are taken.

In this section, we explore methods for quantum equivalents of such entropy-limiting initialization schemes and in Sec. V we detail some of these "more heroic" measures.

### A. Initial entanglement partitioning

One way to avoid barren plateaus without suppressing expressibility is to initially partition the circuit, like in Fig. 3(a), but then to allow $\mathcal{R}_C$-$\mathcal{R}_N$ entanglement throughout the training process. This method is fundamentally distinct from Refs. [15,19] because we only initialize a subset of two-qubit gates $u_{ij}^k$ to the identity and have devised a cost function and entanglement-based strategy to motivate this choice. Furthermore, our treatment applies to universal PQCs of potentially great depth, not restricted subspaces of $U$ [15]. Figure 3(b) displays the $n = 9$ ground-state compressor loss $\mathcal{L} = \mathcal{L}_g$ of Eq. (9) (gray) and training of $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ (red) for both initially partitioned (solid lines) and fully random (dashed) initializations with $L = 200$. These loss functions were chosen for their relevance to various approaches in quantum machine learning, with the loss functions $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ and $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ serving as simple yet astoundingly effective representations of solution-encoded state preparation, i.e., standard implementations of generic variational quantum eigensolver, and the ground-state condenser as a minimal but complete example of generalization problems in quantum circuits. Throughout this work, the AMSGrad gradient descent algorithm is used for circuit parameter update [42]. The corresponding bipartite entanglement entropies $S$ are in the inset and an in-depth discussion of the behavior of $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ is given in Sec. V.

At first, initially partitioned circuits suffer a bout of decreased accuracy, which corresponds to a period of low yet rapidly growing entanglement $S$ (inset) that is either insufficient to express the target state of interest (gray) or simply lower than those of many of the degenerate solutions (red). Later, initially partitioned circuits can produce lower error and require fewer training epochs, which we hypothesize stems from the responsiveness of the gradient during the initial phase of low entanglement, enabling the system to train unfettered by barren plateaus and driving interactions forward through entropy growth. We emphasize that, while the benefits of initial partitioning for $n = 9$ qubits in Fig. 3(b) are definitive yet moderate, they become vital to training for larger networks [9], where barren plateaus can completely preclude circuit learning due to the exponential suppression of the training gradient. In particular, Fig. 4(b) illustrates that initial partitioning combats the vanishing gradient with matched exponential scaling, rendering it an effective method at-scale. Finally, although initially partitioned training of the ground-state compressor (gray) is initially delayed, such training outperforms random initialization at finding high-accuracy solutions, which is the ultimate goal of such machine-learning generalization tasks. Strategic initializations may be sufficient to avoid barren plateaus throughout training. We emphasize that the relationship between plateau barrenness and $S$ (or alternatively, in other works, $n$ and $L$) is a product of circuit

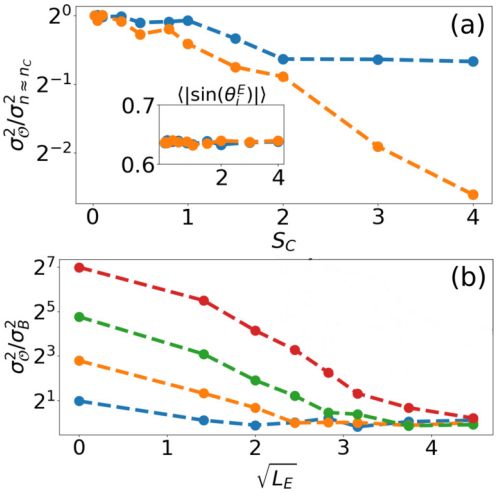


FIG. 4. (a) A deep circuit ($L = 100$) pretraining procedure for $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\rangle$ that minimizes collective entanglement $S_C$ [Eq. (21)], the entanglement entropy between both the input and output registers of $\mathcal{R}_C$ and $\mathcal{R}_N$ for $n = 3, 5$ (blue, orange). As random entanglement decreases with $S_C$, $\sigma_{\mathcal{O}}^2$ increases. Crucially, this pretraining procedure is unique from partitioned initialization as it permits nontrivial interaction at the level of individual circuit layers (inset), with magnitude of inter-register interaction remaining $2/\pi \approx 0.637$, which is consistent with that of random $\theta_i$. (b) Derivative variance $\sigma_{\mathcal{O}}^2$ for $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\rangle$ ($n_C = 2$) and initially partitioned registers $\mathcal{R}_C$ and $\mathcal{R}_N$ in units of its nonpartitioned value $\sigma_B^2$ vs number of register entangling gate layers $L_E$. Here, $L = 200$ and $n = 3, 5, 7, 9$ (blue, orange, green, and red). In both subfigures, each data point sampling two-thousand distinct circuit iterations.

parameter randomness of at least a quantum 2-design and can thus only be assumed for random circuit initializations. That is, such randomness cannot generally be assumed throughout the training process, as this represents an inherently structured organizing of circuit parameters.

When initially partitioned cost functions are learned with high accuracy [here, ground-state compression, but also in both the partitioned training of $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ in Fig. 5(c)], $S$ peaks towards the end of the rapid training period before dropping down to a lower steady-state value. This indicates that the initially partitioned circuit identifies an appropriate solution that is less entangled with the unmeasured qubits of $\mathcal{R}_N$, a potentially desirable quality as widespread entanglement can lower the coherence time of qubits. Moreover, an extension of this technique could be used to partially factor the cost function registers themselves, resulting generally in fewer required readouts for a given cost function determination and ameliorating the so-called "measurement problem" [26–28].

We indicate that this ultimate drop in bipartite entanglement is reminiscent of the late-stage decrease in tripartite mutual information noted in Ref. [31]. The two phenomena are not in conflict, however, the former indicating the disentanglement of measured and unmeasured qubits *after* the necessary information from those qubits had been collected, while the latter signals that the global features of the *input information* are learned towards the end of the training process. In fact, both observations suggest that information locality is the salient feature of late-stage hybrid quantum-classical

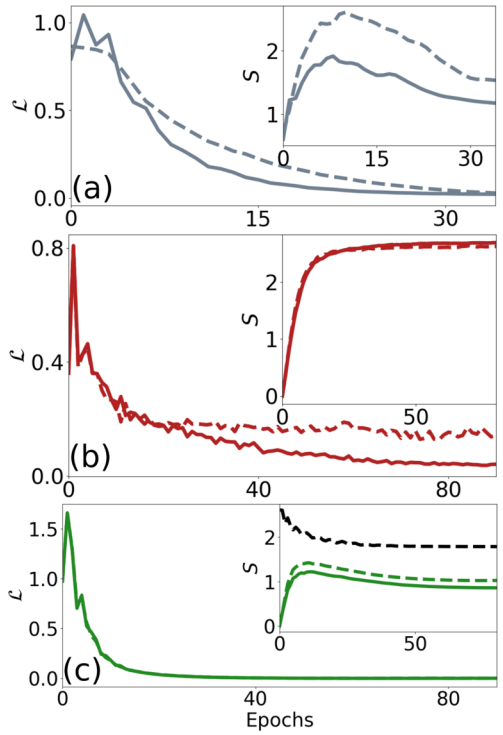


FIG. 5. Loss function vs epochs averaged over two-thousand iterations of both entanglement regularized and initially partitioned circuits (solid lines) and initially partitioned only circuits (dashed lines) of (a) ground-state compressor $\mathcal{L} = \mathcal{L}_g$ [Eq. (9)], (b) $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$, and (c) $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ with $L = 200$. In all cases, nonzero regularization terms lead to (c) equal or (a), (b) faster and more accurate learning with decreased $S$, mitigating the effects of barren plateaus. Moreover, the behavior of $S$ is reflective of the overall training difficulty (insets), whereas the cost functions of panels (a) and (c) can be learned rather rapidly and accurately and do so with naturally lower levels of bipartite entanglement $S$ that are more responsive to regularization, that of panel (b) trains slower and with less accuracy, with entanglement growth that is controlled very little by regularization. The black dashed line in panel (c) corresponds to $S$ for unregularized, unpartitioned $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$, which learns with equal effectiveness as its partitioned and regularized counterparts, highlighting the increased learnability eigenstate learning.

learning algorithms. Finally, we indicate that a reduction in $S$ also occurs for nonpartitioned circuits that can be learned with high accuracy [dashed gray in Fig. 3(b) and dashed black Fig. 5(c)], whereas it is absent from low-accuracy circuits (red). Indeed, some degree of automatic $\mathcal{R}_C$-$\mathcal{R}_N$ factorization appears to be a natural feature of high-accuracy QNN training.

Finally, we comment that certain cost functions [see $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ in Fig. 3(c)] train rapidly when initialized to a barren plateau, and neither the time nor accuracy are improved by mitigating these barren plateaus via initial partitioning. This suggests that certain classes of cost functions (in this case observables that target one of their eigenstates, as will be discussed in Sec. V) may be naturally resistant to barren plateaus. This could potentially be due to an accelerating ordering process, wherein even initially slow learning (small initial gradient variation) reduces entanglement and, in turn, leads to progressively larger gradients and faster learning. In this manner, even small $\sigma_{\mathcal{O}}^2$ might quickly navigate $\mathcal{L}$ to a nonbarren region of its landscape.

### B. Entanglement meta-learning as circuit pretraining

A similar yet more sophisticated solution for nonbarren initializations is classical pretraining of the circuit gates to control $\mathcal{R}_C$-$\mathcal{R}_N$ entanglement. This process is a form of meta-learning [43], a branch of machine-learning algorithms directed at optimizing the learning process of other algorithms.

We must be careful, however, in our choice of pretraining cost function. Simply minimizing $S$ would itself be a form of randomly parametrized gradient descent algorithm on the output qubits and would, thus, like previous meta-learning techniques, tend to generate the concentration of parameters that lead to barren plateaus [29]. We reiterate that this observation is not in conflict with our claim that $\sigma_{\mathcal{O}}^2$ vanishes $\propto 2^{-S}$ for randomly initiated PQCs, because this relation is not universal, but rather applies to circuit unitaries that are Haar distributed, and can therefore only be assumed in random, not pretrained, circuits.

As an alternative to $S$, we can combat barrenness by minimizing the *collective* entanglement $S_C$ of registers $\mathcal{R}_C$ and $\mathcal{R}_N$, which considers the $2n$-qubit space of both input and output registers. To define $S_C$, we boost into the $2n$-qubit pure state

$$|\Psi\rangle = \sum_{i,j=0}^{2^n-1} \frac{\langle\psi_i|U|\psi_j\rangle}{\sqrt{2^n}} |\psi_j\rangle|\psi_i\rangle, \tag{19}$$

where $|\psi_i\rangle$ represent some set basis vectors in the $n$-dimensional Hilbert space. We can then define the density-matrix operator of the full $2n$ collective qubit system

$$\mathrm{P} = |\Psi\rangle\langle\Psi|. \tag{20}$$

Now the reduced density matrix and corresponding entropy of entanglement of $\mathcal{R}_C$ are defined over its $2n_C$ input *and* output qubits as

$$\mathrm{P}_C = \mathrm{Tr}_N[\mathrm{P}], \quad S_C = -\mathrm{Tr}[\mathrm{P}_C \log_2 \mathrm{P}_C], \tag{21}$$

where $\mathrm{Tr}_N$ is a trace over the $2n_N$ input and output qubits of $\mathcal{R}_N$. Figure 4(a) shows the pretraining process for $n = 3, 5$ (blue, orange), $n_C = 2$, and $L = 100$, wherein the loss function is set as $\mathcal{L} = S_C$ and circuit parameters are updated according to gradient descent. As the $S_C$ entanglement is minimized, $\sigma_{\mathcal{O}}^2$ draws closer to its partitioned value $\sigma_{n\approx n_C}^2$, reducing the initialization problem of the barren plateau from $O(2^{-n})$ to approximately $O(2^{-n_C})$ as the initialization $S_C$ of the circuit is minimized through training. This comparison is approximate because the pretraining method, while quite general, implies some inherent ordering that may distinguish it from a bipartition of 2-designs. As $S_C \to 0$ and $\mathcal{R}_C$ and $\mathcal{R}_N$ become factorized, $\sigma_{\mathcal{O}}^2$ grows more similar to the variance of an $n_C$ qubit system, reducing the barren plateau effect by $\approx 2^{n-n_C}$ orders of magnitude. We note that as $S_C \leqslant 2\min(n_C, n_N)$, $\sigma_{\mathcal{O}}^2$ of $n = 3$ is constant for $S_C > 2$.

Critically, the average magnitude of $\mathcal{R}_C$-$\mathcal{R}_N$ interaction on a given layer $k$ is not reduced. To see this, consider a rough

metric of inter-register mixing

$$\langle|\sin(\theta_i^E)|\rangle = \frac{1}{3L}\sum_{\theta_i^E}|\sin(\theta_i^E)|, \tag{22}$$

where $\theta_i^E$ are the $3L$ rotation angles of the $u^k_{n_C,n_C+1}$ gates that entangle the registers $\mathcal{R}_C$ and $\mathcal{R}_N$. $\langle|\sin(\theta_i^E)|\rangle$ describes these interactions because it is the average of off-diagonal (or rotating) elements in the two-qubit rotation matrices $u^k_{n_C,n_C+1}$. The inset of Fig. 4 demonstrates that this quantity remains at its uniformly distributed value $2/\pi$, even as the collective registers become increasingly factored. This indicates that, while the total entanglement of collective $\mathcal{R}_C$ and $\mathcal{R}_N$ is reduced, the average inter-register interaction at any given layer $k$ remains unaffected, providing a highly nontrivial circuit initialization. This method is distinct from Ref. [19] because it does not produce a network of identity-producing blocks but rather a nearly arbitrary initialization with the sole yet crucial constraint of adjustable factorizability along a single connection. This distinction may be particularly important for deep circuits [44]. What is more, this method assumes no specification of problem structure [29], making it generally applicable. Due to this generality, once a pretrained set of parameters is learned, it may be stored and redeployed for various cost functions.

Although classical pretraining is untenable for circuits of large $n$, it can serve as meta-learning for the fundamental effect of entanglement on PQCs and their optimal initializations. As such, it may lead to some scalable generalization of the procedure. It could also be applied iteratively on subsets of large circuits, i.e., on the qubits which form the border between $\mathcal{R}_C$ and $\mathcal{R}_N$. Most promisingly, recent advances in efficient subsystem entanglement measuring techniques, such as random measurements [45] and fidelity out-of-time correlators [46], may pave the way for on-hardware hybrid quantum-classical variational minimization of collective entanglement $S_C$, or some analogous measure, enabling full-circuit, high-variance gradient initializations for arbitrary-size PQCs.

## V. DYNAMIC CONTROL OF BARREN PLATEAUS

As the difficulties of training the relatively simple cost function $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ for deep circuits in Fig. 3(b) alludes, initialization techniques can be insufficient for complete mitigation of barren plateaus. To combat this, we now propose a variety of methods to directly manage the long-term entanglement of the $\mathcal{R}_C$ and $\mathcal{R}_N$ output registers. Returning to the analogy of optimal mixing of a classical bipartite gas in Sec. IV, this section details quantum analogies to the more "heroic" methods that we can take to dynamically control the gaseous mixture's entropy, such as regularization of (penalization of the learning algorithm) or the introduction of additional dynamics into the system.

### A. Hard limit on $\mathcal{R}_C$-$\mathcal{R}_N$ entangling gates

The simplest of these dynamic methods is imposing a hard limit on the number of $\mathcal{R}_C$-$\mathcal{R}_N$ entangling layers $L_E$ in an otherwise deep circuit. This is explored in Fig. 4(b). Although total depth $L = 200$, relatively large gradients can still be achieved while still permitting a considerable number of $\mathcal{R}_C$-$\mathcal{R}_N$ interactions. As $L_E$ grows, $\sigma^2_{\mathcal{O}}$ decays with a similar scaling in $L_E$ as unrestricted circuits do in total gate number $L$, corroborating that barren plateaus indeed arise with the spread of cost function entanglement, not circuit depth itself. This method could be particularly fruitful when using a reinforcement learning algorithm [29,47] because the circuit could learn to process and extract the most relevant portions of $\mathcal{R}_N$ before ultimately transferring them to $\mathcal{R}_C$ in a limited number of $L_E$.

### B. Entanglement regularization

A yet more dynamic method for limiting entanglement is with regularization of $\mathcal{R}_C$-$\mathcal{R}_N$ gates. Regularization adds a penalizing term with adjustable scale parameter $\lambda$,

$$\eta = \lambda\sum_i|\sin(\theta_i^E)|\mathcal{L}, \tag{23}$$

to the original cost function $\mathcal{L}$ in order to implicitly limit the amount of cross-register entanglement. $\sum_i|\sin(\theta_i^E)|$ is proportional to the inter-register mixing measure $\langle|\sin(\theta_i^E)|\rangle$ of Sec. IV and serves to limit entanglement-generating interactions. Moreover, as the values $\theta_i$ are already stored within the classical learning algorithm, this metric does not require additional queries to the quantum hardware. Scaling $\eta$ by $\mathcal{L}$ results in an adaptive regularization process that resists entanglement in regions of poor solutions while relaxing to the original learning problem as $\mathcal{L}$ approaches zero. This adaptivity can be even more fruitful by making $\lambda$ a decreasing function of $\mathcal{L}$, such that $\eta$ disturbs the learning process even less near optimal solutions.

The regularized gradient is then

$$\mathcal{O}_i = \left(1+\lambda\sum_i|\sin(\theta_i^E)|\right)\frac{\partial\mathcal{L}}{\partial\theta_i} + \lambda\cos(\theta_i)\mathrm{sign}(\sin(\theta_i))\mathcal{L}. \tag{24}$$

During portions of the training process that are still largely random, the average of $\frac{\partial\mathcal{L}}{\partial\theta_i}$ over all Haar random unitaries $\mu_{\mathcal{O}_i} = 0$, as we can assume by concentration of measure for deep circuits that $\sum_i|\sin(\theta_i^E)|$ is approximately constant. The variance, however, does increase to

$$\sigma^2_{\mathcal{O}_i} \to \left(1+\lambda\frac{6L}{\pi}\right)^2\sigma^2_{\mathcal{O}_i}. \tag{25}$$

We highlight that, although regularization only *directly* augments the variance of parameters $\theta_i^E$ upon which it acts, we observe a similar increase in the unregularized angles, indicating that its mitigation of barren plateaus is a system-wide effect. Furthermore, we note that $\lambda$ is adjustable and that the regularized variance grows quadratically in circuit depth, whereas $\sigma^2_{\mathcal{O}}$ is constant for deep circuits with a given number of qubits $n$. While effective for relatively small ($n = 9$) quantum circuits (see Fig. 5), the necessity of regularization or other techniques to combat barren plateaus becomes exponentially more vital for large networks.

Figure 5 displays this learning process for an initially partitioned circuit trained with a $\lambda$ which is

piecewise-adaptive in (solid line) comparison with an algorithm using only initial partitioning; that is, $\lambda = 0$ (dashed) for $L = 200$. Three different loss functions are used: (a, gray) ground-state compressor $\mathcal{L} = \mathcal{L}_g$ [Eq. (9)], (b, red) $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$, and (c, green) $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$. Ground-state compression can be achieved both faster and with greater factorization of the output solution, while solutions to $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ have greatly improved accuracy. The persistence of high entanglement for $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ further supports our claim that entanglement from randomly parametrized quantum circuits is the source of barren plateau-related training problems, correlating learning difficulty with susceptibility to entanglement growth. The mechanism by which entanglement regularization increases the accuracy learned for $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ without reducing entanglement is derived in the following section by a novel reframing of barren plateaus in terms of Langevin noise, and the results are used to design a new, general-purpose barren plateau amelioration technique. Finally, although $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ is rapidly learned both with and without regularization and/or initial partitioning, its regularized solutions still benefit from the increased factorizability, with the dashed black line in the inset of Fig. 5(c) showing its unpartitioned $S$.

As discussed in Sec. IV, we again comment on the seeming resilience from barren plateaus of $\mathcal{L} = \langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle$ and other cost function measurements that target their eigenstates. While still beginning in a barren landscape with overwhelming probability for random circuit initializations, these algorithms learn equally well as without barren plateau mitigation.

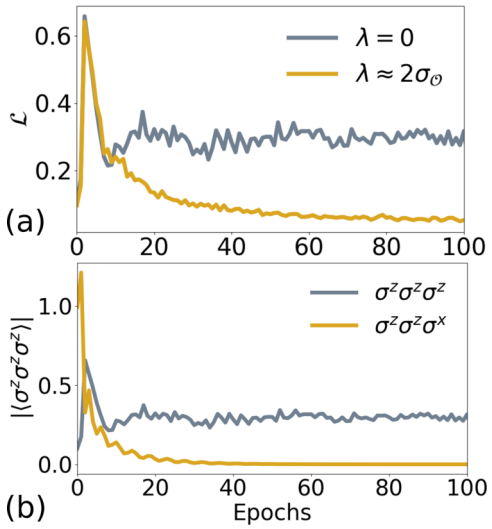


FIG. 6. Preparing a state $|\psi\rangle$ under barren plateau conditions (random initialization of $L = 200$) for $n = 9$ and cost function $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ through both (a) the addition of Langevin noise on a subset of parameters and (b) substitution of measurement basis for which the target state is an eigenstate. (a) Additional Langevin noise term $\lambda\sum_i^N |\phi_i|\mathcal{L}$ increases the gradient variance $\sigma^2_{\mathcal{O}_i} \to (1 + \lambda N\pi)^2\sigma^2_{\mathcal{O}_i}$ with respect to parameters $\phi_i$, thus helping to navigate barren plateau landscapes. This can be viewed as an increase in diffusion constant $2\mathcal{D}$. (b) By substituting the cost function $\langle\sigma_1^z\sigma_2^z\sigma_3^x\rangle$, for which our target state $|\psi\rangle$ is an eigenstate (or alternatively, choosing a "natural" cost function basis), we obtain a faster, more accurate learning process. Both experiments are the average over two-thousand distinct randomly parametrized circuits.

### C. Langevin noise as gradient supplement

The results of Fig. 5(b) raise an interesting point: the addition of regularization terms to the cost function can improve accuracy without significantly decreasing entanglement. We hypothesize that this is because $\eta$, while sometimes successfully limiting entanglement, is always providing additional perturbation in the form of noise. Langevin noise in particular has proven fruitful in classical machine learning and has been used to prevent overfitting in classical neural networks [30].

To motivate this hypothesis, we make the observation that the cost function gradient in barren plateaus can be conceptualized as a form of Langevin noise in circuit parameter space. Typically, Langevin noise is defined for functions $g$ that vary with time $\tau$. Then $g(\tau)$ satisfies the conditions that $\langle g(\tau)\rangle_{\text{Lan}} \equiv \int g(\tau)d\tau = 0$ and $\langle g(\tau)g(\tau')\rangle_{\text{Lan}} = 2\mathcal{D}\delta(\tau - \tau')$, where $\delta$ is the Dirac $\delta$ function and $\mathcal{D}$ is some finite, nonzero diffusion constant [48]. In the case of $\mathcal{O}$, the moments are not integrals over time, but rather over the parameters $\theta_i$ as described by the Haar measure, such that $\mathcal{D} = \sigma^2_{\mathcal{O}}/2$.

By introducing this novel Langevin noise formulation, we reframe barren plateaus as an entanglement-induced diminution of $\mathcal{D}$. In this framework it becomes apparent that, even when a network's random entanglement cannot be restricted, barren plateaus can still be mitigated by introducing additional Langevin noise $\lambda\sum_i^N |\phi_i|\mathcal{L}$ to the original loss function such that

$$G = \left(1 + \lambda\sum_i^N |\phi_i|\right)\mathcal{L}, \tag{26}$$

with derivatives $g_i = \frac{\partial G}{\partial\phi_i}$ and where $\phi_i$ are an arbitrarily chosen subset of circuit parameters of size $N$. For uniformly distributed $\phi_i \in \hat{\phi}$ on the interval $[0, 2\pi)$, this yields the equivalent relation

$$\langle g_i(\hat{\phi})g_i(\hat{\phi})\rangle_{\text{Lan}\phi} \equiv 2\mathcal{D} = (1 + \lambda N\pi)^2\sigma^2_{\mathcal{O}_i}. \tag{27}$$

Figure 6(a) illustrates the effectiveness of such Langevin noise in barren landscapes, producing a high-accuracy solution for $\mathcal{L} = |\langle\sigma_1^z\sigma_2^z\sigma_3^z\rangle|$ despite fully random initialization for $n = 9$ on a deep circuit ($L = 200$). We note that, like the increased variance of entanglement regularization, angles that are not directly perturbed by added Langevin noise still enjoy an increase in variance from the system-wide effect of the technique.

### D. Natural cost function bases

Finally, we discuss our repeated observation that successful learning in initially barren landscapes is greatly facilitated when the target output is an eigenstate of the cost function observables, a basis choice that we refer to as "natural." This observation of a natural basis has been made for other product-state PQC objective functions, such as in the basis

transformations of electronic structure reference states in quantum chemistry [2]. Not only do such configurations learn more rapidly, their training rate and accuracy are not impacted by otherwise successful barren plateau mitigation techniques.

We have suggested a potential link between this trainability of natural basis cost functions and the tendency of these circuits to limit their own entanglement, navigating out of the barren landscapes of random matrices and into a tractable configuration. In particular, the variational preparation of measurement eigenstates has several entropy-based advantages, such as a vanishing gradient variance when circuit approaches an optimal solution. As discussed in Sec. IV, this effect may also be due to a rapidly accelerating ordering process, wherein even small $\sigma_{\mathcal{O}}^2$ lead to $\mathcal{L}$ efficiently escaping into nonbarren regions of its landscape.

Regardless of the origin of its effect, rotating cost function measurements into natural bases can be a strong barren plateau mitigating strategy. Figure 6(b) illustrates that, by substituting $\mathcal{L} = \langle \sigma_1^z \sigma_2^z \sigma_3^x \rangle$ for $\mathcal{L} = |\langle \sigma_1^z \sigma_2^z \sigma_3^z \rangle|$, we can obtain a desired solution $|\psi\rangle$ such that $\langle \psi | \sigma_1^z \sigma_2^z \sigma_3^z | \psi \rangle = 0$ much more effectively. The replacement of a single $\sigma^z$ with the operator $\sigma^x$ reduces the problem to an eigenstate optimization and results in a learning process that trains quickly and automatically limits entanglement [entanglement behavior analogous to the black dashed line in Fig. 5(c)], in contrast to the difficulties of the original problem [red dashed line in Fig. 3(b)].

## VI. CONCLUSION

We have demonstrated the relationship between total qubit-cost function random entanglement and the barrenness of a learning landscape both analytically and numerically and oriented these findings within the context of many-body entanglement dynamics. Based on these results, we established various metrics for barren plateau prediction, both in terms of entanglement and, for a 1D system, circuit depth. We also proposed an input-output entanglement metric, whose minimization we suggest is key to circuit learnability. Using this knowledge, we went on to propose various mitigation schemes, including initial partitioning of cost function and non-cost function registers, meta-learning of low-entanglement high-interaction PQC initializations, limiting inter-register interaction, entanglement regularization, the addition of Langevin noise, and utilizing natural cost function bases. We demonstrated the effectiveness of these techniques, elucidating the role that entanglement minimization plays in both the assisted and unassisted training of QNNs and emphasizing that, as existing barren plateau proofs assume sufficiently random parametrizations which do not apply under all circumstance, barren plateaus can potentially be avoided or escaped in generic PQCs.

While these findings imply that QNN learning must strike a nontrivial balance between randomness, expressibility, and barrenness, they lay the groundwork for numerous mitigation techniques that may facilitate large-scale quantum circuit learning. Furthermore, these methods furnish various secondary benefits, such as solution factorization, novel paradigms of quantum meta-learning, and increased understanding of circuit optimization, to name a few. Furthermore, they suggest that the growth of circuit entanglement could potentially be harnessed to drive the learning process.

Oftentimes, the presence of barren plateaus in PQCs is interpreted as an absolute impasse, because it is typically believed to preclude learning. However, this work emphasizes that not only can barren plateaus be ameliorated through entanglement considerations, they should be understood as manifestations of circuit randomness that have not been proved to apply to more organized configurations, such as those which may manifest during the learning process. To understand the relationship between cost function barrenness and total circuit learnability, the evolution of circuit parameter distributions throughout the learning process should be characterized. Such a statistical characterization will also shed further light on the viability of barren plateau mitigation methods.

## ACKNOWLEDGMENTS

S.F.Y. and T.L.P. would like to thank the AFOSR and the NSF for funding through the CUA-PFC grant. T.L.P. acknowledges that this material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1745303. X.G. is supported by the Postdoctoral Fellowship in Quantum Science of the Harvard-MPQ Center for Quantum Optics, the Templeton Religion Trust grant TRT 0159, and by the Army Research Office under Grant W911NF1910302 and MURI Grant W911NF-20-1-0082.

[1] J. Preskill, Quantum **2**, 79 (2018).
[2] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New J. Phys. **18**, 023023 (2016).
[3] E. Farhi, J. Goldstone, and S. Gutmann, arXiv:1411.4028.
[4] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, Nat. Commun. **5**, 4213 (2014).
[5] M.-H. Yung, J. Casanova, A. Mezzacapo, J. McClean, L. Lamata, A. Aspuru-Guzik, and E. Solano, Sci. Rep. **4**, 3589 (2014).
[6] D. Wecker, M. B. Hastings, and M. Troyer, Phys. Rev. A **92**, 042303 (2015).
[7] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Nature **549**, 242 (2017).
[8] E. Farhi and H. Neven, arXiv:1802.06002.
[9] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Nat. Commun. **9**, 4812 (2018).
[10] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Nat. Commun. **12**, 1791 (2021).
[11] A. Uvarov and J. Biamonte, J. Phys. A: Math. Theor. **54**, 245301 (2021).
[12] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, arXiv:2011.12245.

[13] A. Anand, M. Degroote, and A. Aspuru-Guzik, Sci. Technol. **2** 045012 (2021)

[14] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, arXiv:2005.12458.

[15] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, PRX Quantum **1**, 020319 (2020).

[16] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, arXiv:2006.14904.

[17] E. Fontana, M. Cerezo, A. Arrasmith, I. Rungger, and P. J. Coles, arXiv:2011.08763.

[18] T. Volkoff and P. J. Coles, Quantum Sci. Technol. **6**, 025008 (2021).

[19] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, Quantum **3**, 214 (2019).

[20] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, arXiv:2011.02966.

[21] M. J. Bremner, C. Mora, and A. Winter, Phys. Rev. Lett. **102**, 190502 (2009).

[22] D. Gross, S. T. Flammia, and J. Eisert, Phys. Rev. Lett. **102**, 190501 (2009).

[23] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht, and A. T. Sornborger, Phys. Rev. Lett. **126**, 190501 (2021).

[24] C. O. Marrero, M. Kieferová, and N. Wiebe, arXiv:2010.15968.

[25] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, arXiv:2007.14384.

[26] V. Verteletskyi, T.-C. Yen, and A. F. Izmaylov, J. Chem. Phys. **152**, 124114 (2020).

[27] P. Gokhale, O. Angiuli, Y. Ding, K. Gui, T. Tomesh, M. Suchara, M. Martonosi, and F. T. Chong, arXiv:1907.13623.

[28] A. F. Izmaylov, T.-C. Yen, R. A. Lang, and V. Verteletskyi, arXiv:1907.09040.

[29] G. Verdon, M. Broughton, J. R. McClean, K. J. Sung, R. Babbush, Z. Jiang, H. Neven, and M. Mohseni, arXiv:1907.05415.

[30] M. Welling and Y. W. Teh, *Proceedings of the 28th International Conference on International Conference on Machine Learning* (2011).

[31] H. Shen, P. Zhang, Y.-Z. You, and H. Zhai, Phys. Rev. Lett. **124**, 200504 (2020).

[32] P. Dita, J. Phys. A: Math. Gen. **36**, 2781 (2003).

[33] Z. Puchala and J. Miszczak, Bull. Pol. Acad. Sci.: Tech. Sci. **65**, 21 (2017).

[34] A. W. Harrow and R. A. Low, Commun. Math. Phys. **291**, 257 (2009).

[35] M. Walters, Ph.D. thesis, University of Rochester, 2015 (unpublished).

[36] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, J. Math. Phys. **45**, 2171 (2004).

[37] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Phys. Rev. A **80**, 012304 (2009).

[38] M. Žndarič, Commun. Phys. **3**, 100 (2020).

[39] A. Nahum, J. Ruhman, S. Vijay, and J. Haah, Phys. Rev. X **7**, 031016 (2017).

[40] A. Nahum, S. Vijay, and J. Haah, Phys. Rev. X **8**, 021014 (2018).

[41] C. W. von Keyserlingk, T. Rakovszky, F. Pollmann, and S. L. Sondhi, Phys. Rev. X **8**, 021013 (2018).

[42] S. J. Reddi, S. Kale, and S. Kumar, *International Conference on Learning Representations* (2018).

[43] S. Thrun and L. Pratt, in *Learning to Learn* (Springer Science+Business Media, New York, 1998), pp. 3–17.

[44] E. Campos, A. Nasrallah, and J. Biamonte, Phys. Rev. A **103**, 032607 (2021).

[45] T. Brydges, A. Elben, P. Jurcevic, B. Vermersch, C. Maier, B. P. Lanyon, P. Zoller, R. Blatt, and C. F. Roos, Science **364**, 260 (2019).

[46] R. J. Lewis-Swan, A. Safavi-Naini, J. J. Bollinger, and A. M. Rey, Nat. Commun. **10**, 1581 (2019).

[47] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, npj Quantum Inf. **5**, 85 (2019).

[48] W. Coffey, Y. P. Kalmykov, and J. T. Waldron, *The Langevin Equation: With Applications in Physics, Chemistry and Electrical Engineering* (World Scientific, Singapore, 1996).