Avoiding Barren Plateaus Using Classical Shadows

Stefan H. Sack,[1,*,‡] Raimel A. Medina,[1,†,‡] Alexios A. Michailidis,[1,2] Richard Kueng,[3] and Maksym Serbyn[1]

[1] *IST Austria, Am Campus 1, Klosterneuburg 3400, Austria*
[2] *Department of Theoretical Physics, University of Geneva, 24 quai Ernest-Ansermet, Geneva 1211, Switzerland*
[3] *Institute for Integrated Circuits, Johannes Kepler University Linz, Altenberger Straße 69, Austria*

(Received 1 February 2022; revised 12 April 2022; accepted 20 May 2022; published 29 June 2022)

Variational quantum algorithms are promising algorithms for achieving quantum advantage on near-term devices. The quantum hardware is used to implement a variational wave function and measure observables, whereas the classical computer is used to store and update the variational parameters. The optimization landscape of expressive variational ansätze is however dominated by large regions in parameter space, known as barren plateaus, with vanishing gradients, which prevents efficient optimization. In this work we propose a general algorithm to avoid barren plateaus in the initialization and throughout the optimization. To this end we define a notion of *weak barren plateaus* (WBPs) based on the entropies of local reduced density matrices. The presence of WBPs can be efficiently quantified using recently introduced shadow tomography of the quantum state with a classical computer. We demonstrate that avoidance of WBPs suffices to ensure sizable gradients in the initialization. In addition, we demonstrate that decreasing the gradient step size, guided by the entropies allows WBPs to be avoided during the optimization process. This paves the way for efficient barren plateau-free optimization on near-term devices.

DOI: 10.1103/PRXQuantum.3.020365

I. INTRODUCTION

In recent years the field of quantum computation has seen rapid growth fueled by the arrival of the first generation of quantum computers, dubbed noisy intermediate-scale quantum devices (NISQ) [1]. The NISQ era is characterized by quantum computers with a small number of qubits and limited control. The number of coherent operations that can be performed is small and the implementation of famous algorithms with proven quantum speedups, such as Shor's algorithm [2], remains out of reach. To make use of the current generation of quantum computers, the so-called variational hybrid approach [3] was proposed. The idea is to use the quantum computer in a feedback loop with a classical computer, where it implements a variational wave function that is measured to compute the value of the so-called cost function. This information is then fed into a classical computer where it is processed and the variational wave function is subsequently updated aiming to find a minimum of the cost function, which provides an (approximate) solution to the computationally hard problem. The variational hybrid approach has seen a wide range of proof-of-concept applications on NISQ devices ranging from quantum chemistry [4,5] to quantum optimization [6,7] and quantum machine learning [8,9].

Despite the large number of suggested applications, the variational approach encountered also a number of obstacles, that have to be overcome for the future success of the method. In particular, the infamous emergence of *barren plateaus* (BPs) implies that expressive variational ansätze tend to be exponentially hard to optimize [10]. The main obstacle on the way to optimization lies in the fact that gradients of the cost function are on average zero and deviations vanish exponentially in system size, thus precluding any potential quantum advantage. Moreover, it has been shown that the classical optimization problem is generally NP-hard and plagued with many local minima [11].

The problem of BPs attracted significant attention, and numerous approaches were proposed in the literature. In particular, the early research focused on avoidance of BP at the *initialization stage* of variational algorithms [12–16]. In a different direction, the relation between occurrence of BPs and the structure of the cost function was studied [17,18]. Also notions of so-called

[*] stefan.sack@ist.ac.at
[†] raimel.medina@ist.ac.at
[‡] These authors contributed equally.

entanglement-induced [19] and noise-induced [20] BPs were introduced. The relation between BPs and entanglement has lead to various proposals that suggest controlling entanglement to mitigate BPs [21–24]. However, measuring entanglement is hard, therefore making these approaches impractical on a real quantum device.

In this work we introduce the notion of *weak barren plateaus* (WBPs), in order to diagnose and avoid BPs in variational quantum optimization. WBPs emerge when the entanglement of a local subsystem exceeds a certain threshold identified by the entanglement of a fully scrambled state. In contrast to BPs, WBPs can be efficiently diagnosed using the few-body density matrices and we show that their absence is a sufficient condition for avoiding BPs. Based on the notion of WBPs, we propose an algorithm that can be readily implemented on available NISQ devices. The algorithm employs *classical shadow* estimation [25] during the optimization process in order to efficiently estimate the expectation value of the cost function, its gradients, and the second Rényi entropy of small subsystems. The tracking of the second Rényi entropy enabled by the classical shadows protocol allows for an efficient diagnosis of the WBP both at the initialization step and during the optimization process of variational parameters. If the algorithm encounters a WBP, as witnessed by a certain subregion having a sufficiently large Rényi entropy, the algorithm restarts the optimization process with a decreased value of the update step (controlled by the so-called learning rate). We support the proposed procedure by rigorous results and numerical simulations. The structure of the paper is as follows.

In Sec. II we introduce the theoretical framework and present our main results. In Sec. II A we introduce the framework of variational quantum eigensolvers (VQEs). Section II B introduces the phenomenon of BPs, which dramatically hinders the performance of VQEs. In Sec. II C we demonstrate WBPs to be a precursor to BPs. We explain why and how WBPs can be efficiently diagnosed in experiments and contrast this with much harder task of detecting BPs. Finally we propose a modification to the VQE algorithms, which allows prevention of the occurrence of BPs by avoiding WBPs.

In Sec. III we present a bound for the expectation value of the second Rényi entropy in quantum circuits drawn from a unitary ensembles forming a 2-design. This bound allows us to use the second Rényi entropy, which is much easier to estimate, instead of the entanglement entropy. In Sec. III A we provide a formal definition of WBPs according to the value of the second Rényi entropy of the subsystem and prove that the occurrence of a BP implies the occurrence of a WBP. From this argument it follows that the absence of a WBP precludes the occurrence of a BP. In addition, we provide an upper bound (whose proof is found in Appendix A) for the measurement budget require in order to estimate a WBP using classical shadows. Finally, in Sec. III B we demonstrate numerically how the avoidance of WBPs precludes the presence of a BP using the popular BP-free small-angle initialization [15,26].

In Sec. IV, we explore how BPs and WBPs emerge at different stages in the VQE optimization and perform a systematic performance analysis. Next, in Sec. IV A we explore the relation of the learning rate and entropy growth for a single update of the VQE algorithm. We analytically and numerically illustrate how a large learning rate leads to an uncontrolled growth in subsystem entropies, essentially driving optimization to a WBP region. In Sec. IV B we explore the performance of the WBP-free VQE algorithm in different settings for the Heisenberg model on a chain. Finally, in Sec. IV C, we show that our approach allows for the efficient convergence to both, area- and volume-law entangled ground states and compare it to layerwise optimization [13], which is a popular heuristic for BP avoidance. This is illustrated using the Heisenberg model on a random 3-regular graph, additional results for Sachdev-Ye-Kitaev (SYK) model can be found in the Appendix E which exhibits a nearly maximally entangled ground state.

Finally, in Sec. V we summarize our results, discuss their implications, and outline open questions.

II. AVOIDING BARREN PLATEAUS IN VARIATIONAL QUANTUM OPTIMIZATION

In this section we first introduce the framework of VQEs, i.e., the unitary ensemble, the cost functions, and the optimization algorithm, and discuss the BP problem. After this, we present our main result—a specific modification of the VQE that avoids the issue of BPs.

A. Variational quantum eigensolver

The aim of the VQE, initially introduced by Peruzzo *et al.* [27], is to approximate the ground state $|\mathrm{GS}\rangle$ of a Hamiltonian H with a variational wave function $|\psi(\boldsymbol{\theta})\rangle$. A quantum computer is used to prepare the variational function via the action of a set of unitary gates, $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\,|\psi_0\rangle$, where $|\psi_0\rangle$ is the initial state that is typically assumed to be a product state. The variational parameters are then iteratively updated to minimize the expectation value of the Hamiltonian, also called cost function $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,H\,|\psi(\boldsymbol{\theta})\rangle$.

We consider a unitary circuit $U(\boldsymbol{\theta})$ of the form of the so-called "hardware-efficient" ansatz [4]

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{p} W_l \left(\prod_{i=1}^{N} R_l^i(\theta_l^i) \right), \tag{1}$$

where $\theta_l^i \in [-\pi, \pi)$ are pN variational angles, concisely denoted as $\boldsymbol{\theta}$. In this expression the product goes over spatial dimension $i = 1, \ldots, N$ that labels individual qubits and "time dimension," $l = 1, \ldots, p$ with p specifying a number of layers, see Fig. 1 (a). We choose the single-qubit gates to be rotations $R_l^i(\theta_l^i) = \exp\left(-\frac{i}{2}\theta_l^i G_{l,i}\right)$ with random directions given by $G_{l,i} \in \{\sigma^x, \sigma^y, \sigma^y\}$. W_l is an entangling layer that consists of two-qubit entangling gates represented by nearest-neighbor controlled-Z (CZ) gates with periodic boundary conditions, see Fig. 1(a) for an illustration.

We focus our study on k-local Hamiltonians H, defined as sum of terms each containing at most k Pauli matrices. We take k to be finite and fixed, while the number of qubits $N \gg k$. A particular example of a 2-local Hamiltonian from the many-body physics is provided by the Heisenberg (XXX) model subject to a magnetic field

$$H_{XXX} = \sum_{i,j\in V_{\mathcal{G}}} J\left(\sigma_i^z\sigma_j^z + \sigma_i^y\sigma_j^y + \sigma_i^x\sigma_j^x\right) + h_z\sum_{i=1}^{N}\sigma_i^z, \quad (2)$$

where $V_{\mathcal{G}}$ refers to the vertex set of the graph $\mathcal{G}$ and, couplings are fixed $J = h_z = 1$. In our simulations we consider two different graphs: a ring corresponding to a one-dimensional (1D) chain with periodic boundary condition, and a random 3-regular graph. The $U(1)$ symmetry related to the conservation of the z component of the spin under the action of H, as well as translational invariance present for chains with periodic boundary condition, can be explored to decrease the space of parameters by using a suitable gate set respecting this symmetry. However, for the sake of generality we focus on the hardware-efficient unitary ansatz defined in Eq. (1).

Obtaining the energy expectation value $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})| H |\psi(\boldsymbol{\theta})\rangle$ requires measuring a subset or all qubits in the circuit as we schematically show in Fig. 1(a). For our example of a 2-local Hamiltonian on the 1D chain, the required measurements include the value of the σ^z operator on all sites along with the $\sigma_i^a\sigma_{i+1}^a$ expectation values of all $i = 1, \ldots N$ (periodic boundary condition is assumed, so that bits 1 and $N+1$ are identified) and $a = x, y, z$. Finding the optimal parameters $\boldsymbol{\theta}^\star$ requires minimization of the Hamiltonian expectation value $E(\boldsymbol{\theta}^\star) = \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ performed by a classical computer.

There is a plethora of sophisticated classical optimization algorithms that were applied to this minimization problem [28–31]. We use the plain gradient-descent (GD) algorithm due to its simplicity, which makes analytical considerations easier. A GD update step is given by

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\nabla_{\boldsymbol{\theta}}E(\boldsymbol{\theta}), \quad (3)$$

where η is the *learning rate*, which controls the update magnitude. This update step is repeated until convergence of $E(\boldsymbol{\theta})$, which results from finding a (local) minimum of $E(\boldsymbol{\theta})$.

The resulting VQE algorithm is shown schematically in Fig. 1(b) by solid lines. Following the initialization of the variational angles $\boldsymbol{\theta}$, that may be chosen to be real random numbers, the quantum computer is used to prepare the variational state and provide quantum measurement results. The classical computer uses the measurements to estimate the value of the cost function and its gradient, and performs an update of the variational parameters controlled by the learning rate η.

B. Barren plateaus and entanglement

Whilst the VQE described above is a promising framework for near-term quantum computing due to its modest

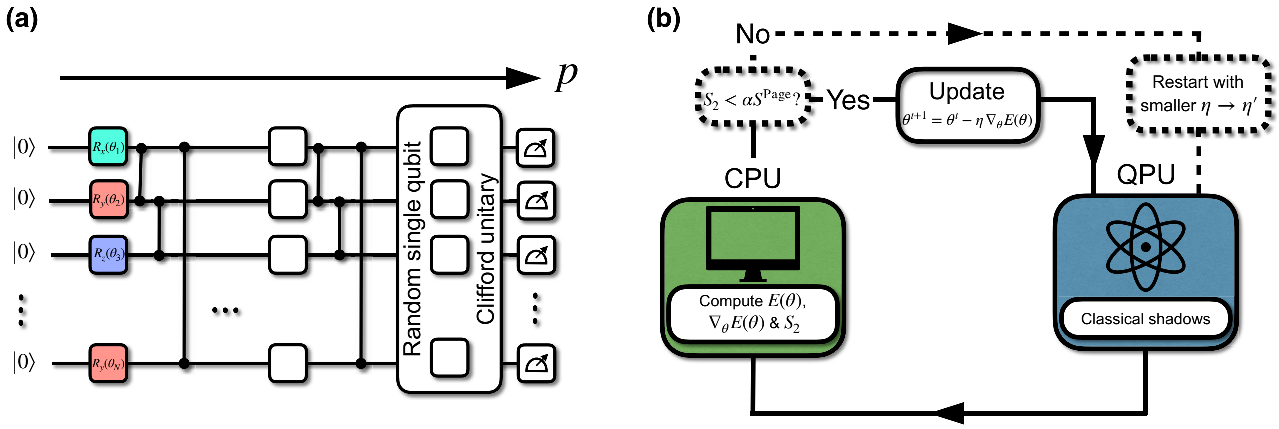
(a)
p
|0⟩
|0⟩
|0⟩
|0⟩
$R_x(\theta_1)$
$R_y(\theta_2)$
$R_z(\theta_3)$
$R_y(\theta_N)$
...
Random single qubit Clifford unitary
(b)
No
$S_2 < \alpha S^{\rm Page}$?
Yes
Update
$\theta^{t+1} = \theta^t - \eta\nabla_\theta E(\theta)$
Restart with smaller $\eta \to \eta'$
CPU
QPU
Compute $E(\theta)$, $\nabla_\theta E(\theta)$ & S_2
Classical shadows

FIG. 1. (a) Illustration of the variational quantum circuit $U(\boldsymbol{\theta})\,|0\rangle$ that is considered in the main text followed by the shadow tomography scheme [25]. The variational circuit consists of alternating layers of single-qubit rotations represented as boxes and entangling CZ gates shown by lines. The measurements at the end are used to estimate values of the cost function, its gradients, and other quantities. (b) The original hybrid variational quantum algorithm shown by solid boxes can be modified without incurring significant overhead as is shown by the dashed lines and boxes. The modified algorithm tracks entanglement of small subregions and restarts the algorithm if it exceeds the fraction of the Page value that is set by parameter α. The full algorithm is efficient; rigorous sample complexity bounds are provided in Appendix A.

hardware requirements, its performance may be ruined by the issue of barren plateaus [10,15,17]. Specifically, the BPs are defined as regions in the space of variational parameters where the variance of the cost-function gradient (and consequently its typical value) vanishes exponentially in the number of qubits [10]:

$$\mathrm{Var}[\partial_{i,l}E(\boldsymbol{\theta})] \sim \mathcal{O}\left(\frac{1}{2^{2N}}\right). \tag{4}$$

McClean *et al.* [10] were among the first to theoretically investigate BPs. They showed that the appearance of a BP can be related to the circuit matching the Haar random distribution up to the second moment. More precisely, they showed that BPs are a consequence of the unitary ensemble $\mathcal{E} \sim \{U(\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ forming a so-called 2-design [10] (see Appendix B for details and the definition of a t-design). To understand the different circuit depth at which BPs are encountered, the authors in Ref. [17] introduced the concept of cost-function-dependent BPs. In particular, they argued that the emergence of BP occurs at different circuit depths, depending on the nature of the cost function.

In contrast, for a so-called global cost function, exemplified by the fidelity, Ref. [17] found that BPs already occur at very modest circuit depths $p \sim \mathcal{O}(1)$. The emergence of BP for the fidelity is naturally related to “orthogonality catastrophe” in many-body physics: even a small global unitary rotation applied to the many-body wave function results in it becoming nearly orthogonal to itself. In terms of fidelity, this implies that it vanishes exponentially in the number of qubits. Moreover, most global state features—such as expectation values of general operators, fidelities with general states and global purities—cannot be efficiently accessed on NISQ devices, and are therefore not practical from an algorithmic point of view [25,32–34]. Therefore, in what follows we do not consider the global cost functions and corresponding BPs.

Local cost functions, that are the focus of the present work are characterized by a later onset of BPs. Specifically, for a k-local cost function where k is fixed, the BPs will occur for circuit depth $p \sim \mathcal{O}(\mathrm{poly}(N))$ that increases polynomially in system size [10,17]. In other words, for a large enough p the VQE algorithm will also suffer from a BP already at the very first step of the GD optimization, provided random choice of variational angles $\boldsymbol{\theta}$. We also note that gradient-free optimization strategies do not circumvent the BP problem since the optimization landscape is inherently flat [35].

The potential emergence of BPs at the initialization stage of the VQE and other algorithms spurred the investigation of different initializations strategies that avoid BPs. Until now, several BP-free initializations were considered in the literature. Reference [12] suggests to initialize the circuit with blocks of identities, Ref. [13] suggests to optimize the ansatz layer by layer, and Ref. [14] suggests to use a matrix-product-state ansatz that is optimized by a separate algorithm [36] and map that to a quantum circuit. In this work we focus on small single-qubit rotation as suggested in Ref. [15].

More recently, it was observed that the entanglement entropy defined as a trace of the reduced density matrix, $S = -\mathrm{tr}\rho_A \ln \rho_A$ (where $\rho_A = \mathrm{tr}_B \rho$ is the reduced density matrix where A is the subset of qubits that are measured and B is the rest of the system) is another source for the occurrence of BPs [19]. The community has subsequently dubbed this kind of BP, *entanglement-induced* BP [19,21,23,24]. In this work, we however show that entanglement-induced BPs and BPs for local cost functions, are in fact one and the same. Avoiding entanglement-induced BPs is equivalent to avoiding BPs for local cost functions, the details are presented in Sec. III.

Experimentally probing a BP is a hard task. The estimation of the exponentially small gradient in Eq. (4) requires a number of measurements that is exponential in the number of qubits, and therefore invalidates any potential quantum speedup. Moreover, small values of gradient encountered in BP have to be distinguished from the case when gradient vanishes due to convergence to a local minimum. Experimentally diagnosing BPs via entanglement is also impractical. For example, quantum circuits that implement 2-design and thus lead to BPs for local cost functions are characterized by typical volume-law entanglement that approaches nearly maximal values. Checking volume-law entanglement scaling on any device is a formidable challenge.

In the process of variational quantum optimization, the majority of approaches to mitigate BPs apply to the initialization stage [12,37,38] and not during the optimization. In Sec. IV, we illustrate the importance of BP mitigation during the optimization. This motivates the need to devise a BP mitigation strategy for the initialization and during the optimization procedure that is efficient. This procedure is discussed in the algorithm proposed below.

C. Weak barren plateaus and improved algorithm

In order to devise an efficient algorithm for BP-free initialization and optimization of the VQE we introduce the notion of WBPs. Specifically, for a Hamiltonian that is k local, we define the WBP as the point where the second Rényi entropy $S_2 = -\ln \mathrm{tr}\rho_A^2$ of any subregion of k qubits satisfies $S_2 \geq \alpha S^{\mathrm{Page}}(k, N)$, where the Page entropy in the limit $k \ll N$ corresponds to the (nearly) maximal possible entanglement of subregion A,

$$S^{\mathrm{Page}}(k, N) \simeq k \ln 2 - \frac{1}{2^{N-2k+1}}, \tag{5}$$

where we explicitly use that the Hilbert-space dimension of region A is 2^k and its complement B has Hilbert-space dimension 2^{N-k}. The naive choice for the parameter α is

$\alpha = 1$. Given some *a priori* knowledge of the entanglement structure of the target state $|\mathrm{GS}\rangle$, the choice can however be more informed to help avoid large entanglement local minima, see Sec. III.

The notion of WBP is practical since it is defined by k-body density matrices, being much easier to access on a real NISQ device. The fact that the prevention of a WBP is sufficient for avoiding the BP may be understood by the intuition from quantum many-body dynamics and the process of thermalization or scrambling of quantum information. In the thermalization process the small subsystems are first to become strongly entangled, as is witnessed by the proximity of their density matrix to the infinite temperature density matrix. This intuition suggests that it is enough to keep in check the density matrices of small subsets of qubits. If their entanglement or other properties are far away from thermal, the system overall is still far away from the BP.

Practically, the WBP can be diagnosed using the shadow tomography scheme [25]. This scheme enables an efficient way of representing a classical snapshot of a quantum wave function on a classical computer. In essence, the shadow tomography replaces the measurements performed in the computational basis with a more general measurements, that turns out to be sufficient for reconstructing linear and nonlinear function of the state, such as expectation values of few-body observables and second Rényi entropy of few-body reduced density matrices, respectively.

Relying on the shadow tomography, we propose the following modification of the VQE shown by dashed lines in Fig. 1(b). In essence, we suggest to use the tomography to *simultaneously* measure the cost function value and the k-body second Rényi entropy. For the derivative we require an additional $2pN$ tomographies (two for each parameter) to compute the gradient exactly using the parameter shift rule [39,40], a detailed derivation of the computational cost for each operation is presented in Appendix A. Access to the second Rényi entropy allows prevention of the appearance of WBPs not only at the initialization step, but throughout the optimization cycle. The explicit algorithm works as follows.

If a WBP is diagnosed at the initialization, one may have to take a different initial value of the variational angles or change the initialization ensemble. These aspects are discussed in detail in Sec. III. In addition, the WBP can occur in the optimization loop. This can be mitigated by keeping track of the second Rényi entropies in the optimization process. If the WBP condition is fulfilled, one must restart the algorithm with a smaller learning rate. In Sec. IV we discuss the optimization process in greater details. In particular, we show how the learning rate is related to the potential change in entanglement entropy, which implies that a smaller learning rate is generally better at avoiding WBPs.

```
Choose $\alpha$, default is $\alpha = 1$          ▷ see Sec. III A for details
Choose $\boldsymbol{\theta}$ such that $S_2 < \alpha S^{\mathrm{Page}}(k, N)$
Choose learning rate $\eta$
repeat                                         ▷ see Appendix A for details
    Obtain classical shadows $\hat{\rho}^{(t)}(\boldsymbol{\theta})$
    Use them to compute $E(\boldsymbol{\theta})$, $\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ and $S_2(\boldsymbol{\theta})$
    if $S_2 < \alpha S^{\mathrm{Page}}(k, N)$ then
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$
    else
        Start again with smaller $\eta \leftarrow \eta'$
    end if
until convergence of $E(\boldsymbol{\theta})$
```

Algorithm 1. WBP-free optimization with classical shadows

III. WEAK BARREN PLATEAUS AND INITIALIZATION OF VQE

A. Definition and relation to barren plateaus

As mentioned in the above, BPs for local cost functions are a consequence of the unitary ensemble $\mathcal{E} \sim \{U(\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ forming a 2-design [10,17], which leads to an exponentially vanishing gradient variance, i.e., a BP. What is important to note is that the exponential decay is simply a witness of the emergence of a 2-design. Another, equivalent witness is the second Rényi entropy, where we have the following.

Theorem 1. *(2-design and Rényi entropy) If the unitary ensemble $\mathcal{E} \sim \{U(\boldsymbol{\theta})\}$ forms a 2-design, then for typical instances the second Rényi of the state ρ_A concentrates around the Page value*

$$S^{\mathrm{Page}}(k, N) - \frac{1}{2^{N-2k+1}} \leq \mathbb{E}_{\mathcal{E}}\big[S_2(\rho_A)\big] \leq S^{\mathrm{Page}}(k, N),$$

for all subregions A of size $k \ll N$.

These results are known in the literature, and in the context of random quantum circuits, can be found in Refs. [41–43]. However, for completeness we also provide a proof in Appendix C.

The theorem above implies that a large amount of entanglement naturally follows from the similarity between the considered circuit and a random unitary (2-design). Such similarity also gives rise to the vanishing variance of local cost-function gradients that define BPs. Therefore, so-called entanglement-induced BPs [19] and BPs for local cost functions are the same. In fact, entanglement provides an intuitive picture for the emergence of BPs and its circuit-depth dependence. Every entangling layer in the circuit typically increases entanglement of the resulting wave function, until it saturates to its maximal value for any subregion of k qubits at a circuit depth $p \sim \mathcal{O}(\mathrm{poly}(N))$. If the second Rényi entropy for half of the subsystem $k = N/2$ has saturated, it has saturated for all smaller subsystem sizes and is thus a sufficient check for a BP. Computing the second Rényi is however typically exponentially hard in subsystem size on NISQ devices (for single-copy access this was recently proven in Ref. [33,34]). It is therefore only practical to check a small subregion where k is small and independent of system size.

The above considerations naturally lead us to introduce the notion of WBPs as a modification of the BP that is computationally efficient to diagnose on NISQ devices. More formally we have as follows.

Definition 2. *(Weak barren plateaus) Let H be an N-qubit Hamiltonian, and A is a region containing k qubits. We define a weak barren plateau by the second Rényi entropy of the reduced density matrix ρ_A satisfying $S_2 \geq \alpha S^{Page}(k,N)$ with $\alpha \in [0,1)$.*

This definition works for any k, however it is reasonable to use k that corresponds to the number of spins involved in interaction terms in the Hamiltonian H since it provides a natural length scale. Moreover, in such a case the reduced density matrix of subregion with k spins contains all necessary information needed to extract the expectation values of Hamiltonian terms localized inside this region.

While a WBP is a necessary condition for a BP, it is however not sufficient (which motivates the term *weak*). From a practical perspective we are actually interested only in avoiding a BP. For this, WBPs provide a powerful tool, since the following holds.

Corollary 2.1. *If we find a particular subregion A such that ρ_A does not satisfy the weak barren plateau condition, i.e., Definition 2, it is on average also not in a barren plateau where the variance is exponentially small.*

Proof. This assertion immediately follows from negating Theorem 1. ■

The corollary above formalizes the intuition behind the dynamics of entanglement in a circuit: if the state restricted to the smaller subsystem has not scrambled, then neither has the state restricted to a larger subregion. In practice, using classical shadows we can efficiently check one subregion of size k with a total measurement budget

$$T \geq \frac{4^{k+1}\mathrm{tr}\rho_A^2}{\epsilon^2\delta}, \tag{6}$$

where ϵ is a desired accuracy and δ is a failure probability (over the randomized measurement process). Parameters ϵ and δ do not depend on the number of qubits, whereas the factor $\mathrm{tr}\rho_A^2$ is upper bounded by one for weakly entangled states and can be as small as 2^{-k} when entanglement is large. Moreover, checking all size k subregions incurs an additional overhead of only $k \ln N$. A derivation of this result is presented in Appendix A, see Eq. (A6). Provided that k is small and does not scale with system size, N, this can be efficiently implemented on NISQ devices.

If any of these subregions avoids the WBP condition, we are guaranteed to also avoid an actual BP. For simplicity, in the numerical results below we check for the WBP condition for a particular region containing the first k qubits, i.e., $A = \{1,\ldots,k\}$.

This argument is also intuitive to see by considering a causal cone (blue region) that indicates the extent of the so-called scrambled region (i.e., extend of a subregion with entropy close to the maximal value) in the circuit, see Fig. 2(a). Such a scrambled region grows with every consecutive entangling layer W_l [see Eq. (1)]. When this region extends beyond k qubits, the WBP is reached (left orange dashed line). Later, when the "scrambling lightcone" has extended to the full system, the BP is reached (right orange dashed line). Once the BP is reached all smaller regions are also fully entangled and will satisfy the WBP condition on average.

Figure 2(b) to (d) provides a numerical illustration for the Corollary 2.1 stated above. We use the hardware-efficient circuit, presented in Eq. (1), and compute the gradient variance and second Rényi entropy as a function of circuit depth p for different system sizes N. We fix $|\psi_0\rangle = |0\rangle$ as the initial state, which is simply all qubits in the zero state. Panel Fig. 2(b) shows the exponential decay of the gradient variance that is usually used to diagnose a BP. Panel Fig. 2(c) shows the corresponding bipartite second Rényi entropy. We see that it indeed approaches the Page value (gray dashed line). The Page value is not fully reached since we are considering the second Rényi instead of the von Neumann entanglement entropy, this difference however becomes negligible once the subsystem size is decreased. This numerically illustrates that when the 2-design is reached both the gradient variance and bipartite second Rényi entropy have converged. In panel Fig. 2(d) we consider a smaller region of two qubits and see that the second Rényi for this region saturates to its maximal value at a significantly lower circuit depth. This illustrates the emergence of the WBP that precedes the onset of the BP after a few more entangling layers. Before the WBP is reached, gradients are well behaved and do not decrease exponentially with the system size.

Finally, we address the effects of the control parameter α, that enters in Definition 2 of the WBP. The naive choice is $\alpha = 1$, which means that a WBP is reached if the subregion is maximally entangled with the rest of the system. However, in the case when some *a* priori knowledge about the entanglement properties of the target state $|\mathrm{GS}\rangle$ is available, it can be used to set a smaller value of α. If, for instance, the ground state is only weakly

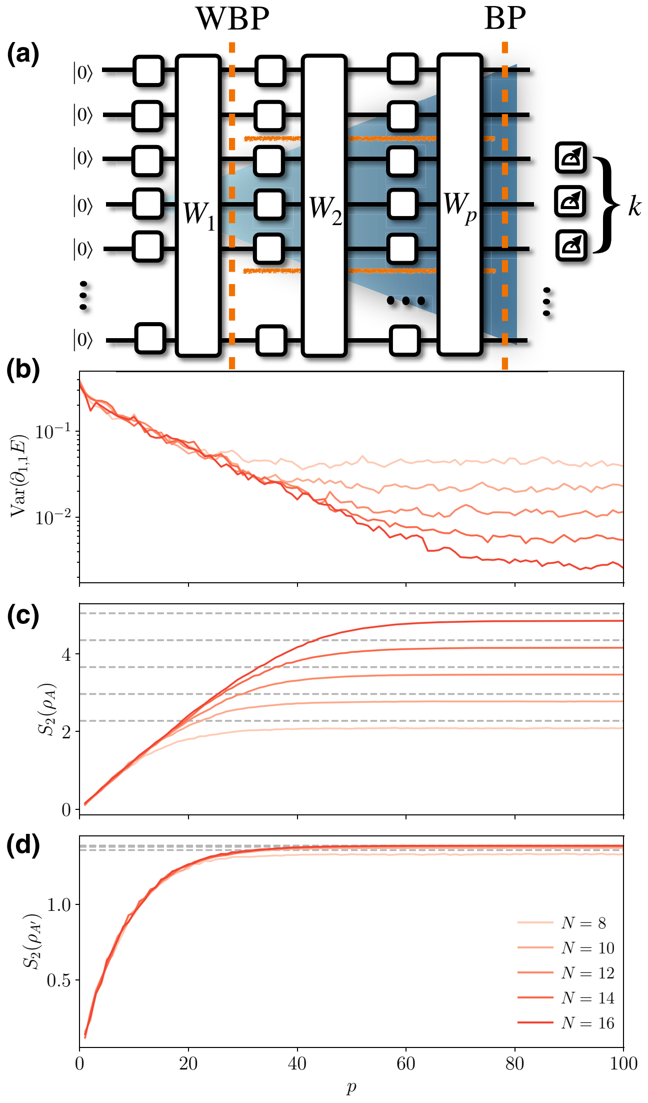

FIG. 2. (a) Sketch of the circuit, where the blue color shows the scrambling lightcone. The lightcone first extends over k qubits, where the WBP occurs, and for larger circuit depths extends to the full system size where the BP occurs. (b) The saturation of the gradient variance $\mathrm{Var}[\partial_{1,1}E]$ and (c) saturation of the bipartite second Rényi entropy $S_2(\rho_A)$ of the region A consisting of qubits $1,\ldots,N/2$ nearly to the Page value happen at the similar circuit depths p, that increases with the system size N. (d) In contrast, the saturation of the second Rényi for two qubits ($A'=\{1,2\}$) is system-size independent, illustrating that WBP precedes the onset of a BP. Data is averaged over 100 random initializations. Gradient variance is computed for the local term $\sigma_1^z\sigma_2^z$, typically used in BP illustrations. Gradient variance for the full Heisenberg Hamiltonian, Eq. (2), looks similar.

entangled, a choice of $\alpha \ll 1$ may be appropriate. In this way Algorithm 1 in Sec. II C can also help in avoiding convergence to highly entangled local minima. We discuss this in more detail in Sec. IV B.

B. Illustration of WBP-free initialization

In order to illustrate the notion of WBP in a more specific setting we apply it to the initialization process of the VQE. Specifically, we focus on the family of initializations that was proposed earlier in order to avoid the issue of BPs [15,26]. The one-parametric family of initializations restricts the single-qubit rotation angles from ansatz Eq. (1) as $\theta_l^i \in \epsilon_\theta[-\pi,\pi)$, where $\epsilon_\theta \in [0,1)$ is the control parameter. This strategy allows the onset of the BP to be delayed to arbitrary circuit depths by tuning ϵ_θ accordingly.

Similarly, it allows the onset of WBPs to be delayed. Depending on the parameter ϵ_θ one can afford a deeper circuit without encountering a WPB in the initialization when compared to the full parameter range ($\epsilon_\theta = 1$). It is straightforward to see that for $\epsilon_\theta = 0$, the ansatz is WBP free for all circuit depths. Indeed, in the absence of the single-qubit rotations, the entangling gates in W_l do not create any entanglement [since the CZ gates used in Eq. (1) are diagonal in the computational basis], leaving $|0\rangle$ invariant. Note that, for example, the *identity block* initialization, proposed by Grant *et al.* [12] works in a similar way in that the unitary is constructed such that it also implements the identity and one is equally left with the zero state.

In Fig. 3 we numerically illustrate the influence of ϵ_θ on the growth of entanglement and its relation to the gradient variance. Panel (a) illustrates the growth of the second

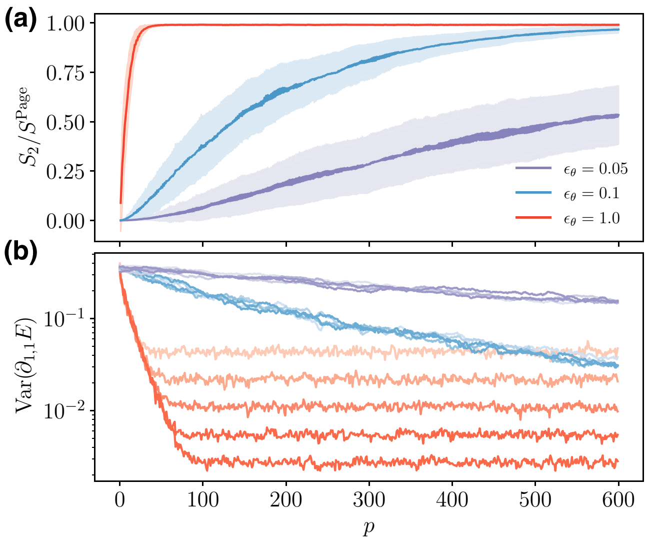

FIG. 3. (a) Decreasing parameter ϵ_θ from 1 slows down the growth of the second Rényi entropy with the circuit depth p. The chosen region contains two qubits. (b) The encounter of BP in the variance of the gradient of the cost function is visible only for the case $\epsilon_\theta = 1$, and it is preceded by the onset of a WBP. We use a system size of $N = 16$ for (a) and $N = 8,\ldots,16$ for (b), color intensity corresponds to system size, same as in Fig. 2. Data is averaged over 100 random instances, variance is for the local term $\sigma_1^z\sigma_2^z$.

Rényi entropy in the circuit for three different small-angle parameters ϵ_θ and panel (b) shows the corresponding gradient variance. Outside of the WBP the gradient variance vanishes at most polynomially in system size N. This illustrates that the avoidance of a WBP is sufficient for avoiding a BP and thus allows for a simple strategy for constructing BP-free initializations.

IV. ENTANGLEMENT CONTROL DURING OPTIMIZATION

A. Bounding entanglement increase at a single optimization step

In Sec. II we presented how the general VQE can be extended with minimal overhead to avoid WBPs in the optimization procedure. The learning rate, as presented in Algorithm 1, hereby plays a crucial role. A smaller learning rate, as observed in Figs. 1(c)–1(e) is more likely to avoid a WBP. To understand this phenomenological observation on more rigorous grounds, let us consider a sufficiently deep circuit (with a polynomial number of layers in system size), so that the optimization landscape is dominated by WBPs. Careful selection of the parameters allows for an initialization outside of a WBP. However, to remain in the WBP-free region, the optimization has to be performed in a controlled manner, such that the optimizer does not leave the region of low entanglement due to large learning rate and does not end in a WBP.

Since WBPs are defined in terms of the second Rényi entropy S_2, we need to bound the change in S_2 between iteration steps t and $t+1$. For practical purposes, we instead use the purity ($\mathrm{tr}\rho_A^2 = e^{-S_2}$). The change in purity is upper bounded by [44]

$$\left|\mathrm{tr}\rho_A^2(t+1) - \mathrm{tr}\rho_A^2(t)\right| \leq 1 - (1 - T_A(t))^2 - \frac{T_A^2(t)}{2^k - 1}, \tag{7}$$

where $T_A(t) \equiv T(\rho_A(t), \rho_A(t+1))$ is the trace distance between the reduced density matrices at iteration steps t and $t+1$, and we assume that region A has k qubits.

Assuming that the states at consecutive update steps of gradient descent are pertubatively close (see Appendix D for details), as measured by the trace distance, one can show that

$$T(\rho_A(t+1), \rho_A(t)) \lesssim \sqrt{\frac{\eta^2}{4}(\nabla_{\boldsymbol{\theta}} E)^T \mathcal{F}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} E}, \tag{8}$$

where $\mathcal{F}_{i,j}(\boldsymbol{\theta}) = 4\,\mathrm{Re}[\langle \partial_i \psi | \partial_j \psi \rangle - \langle \partial_i \psi | \psi \rangle \langle \psi | \partial_j \psi \rangle]$ is the quantum Fisher information matrix (QFIM) [45] and η is the learning rate. Inequalities (7)–(8) imply that the learning rate η can be used to limit the maximal possible change of the purity [46]. Provided that the change in purity is sufficiently small, the Taylor expansion can be used to argue that the corresponding change in the second Rényi entropy S_2, related to the purity as $e^{-S_2} = \mathrm{tr}\rho_A^2$, also remains controlled. Therefore, the choice of an appropriately small learning rate can guarantee the avoidance of a WBP at $t+1$, provided the absence of one at t.

To illustrate the bound numerically, we prepare an initialization outside of the WBP using a small angle parameter $\epsilon_{\boldsymbol{\theta}}$ and compute the change in the purity $\mathrm{tr}\rho_A^2$ after one GD update step for different learning rates η. The results of this procedure for four different learning rates are shown in Fig. 4. We see that larger learning rates correspond to a bigger change in purity and are thus more prone to encounter a WBP. At the same time, all data points are below the theoretical bound. While up to the best of our knowledge the bound Eq. (7) is not proven to be tight, we observe that points corresponding to the extreme learning rates closely approach the theoretical line.

Using Eq. (8), the bound can be efficiently approximated on NISQ hardware: the QFIM can be estimated efficiently on a quantum device using techniques suggested in Ref. [31] or Ref. [47] using classical shadows. For the computation of the gradient one can use the parameter shift rule [39,40] also with shadow tomography. The expression can thus be efficiently evaluated on a real device and used together with the continuity bound to estimate a suitable learning rate η. However, in practice this might not be needed and simply following Algorithm 1 could be more efficient and easier to implement.

B. Optimization performance with learning rate

Finally, we illustrate Algorithm 1 in practice. To this end we first prepare a WBP-free initial state using small

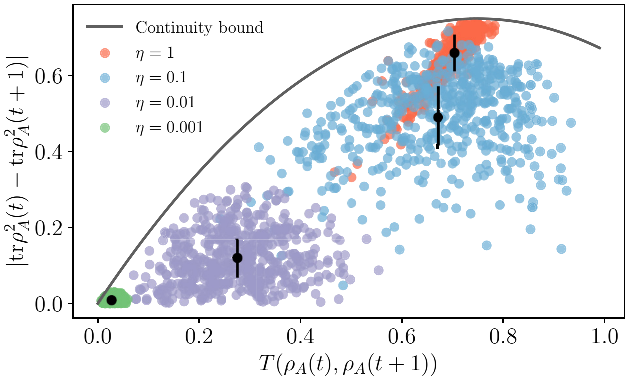

FIG. 4. We numerically illustrate the continuity bound Eq. (7) and its relation to the learning rate η for $t = 0$, i.e., at the beginning of the optimization schedule. This shows that one should be careful with the choice of the learning rate since a large learning rate leads to a big change in the trace distance and change in purity. We use a system size of $N = 10$ and a random circuit with circuit depth $p = 100$ and small qubit rotations ($\epsilon_\theta = 0.05$) to generate a BP-free initialization. Data is averaged over 500 random instances.

qubit rotation angles and compare the performance of GD optimization with different learning rates. If we start with a large learning rate, $\eta = 1$, corresponding to red lines in Figs. 5(a)–5(c), we see that the energy expectation value in Fig. 5(a) rapidly (within one or two update steps) converges to a value far away from the target ground state energy E_{GS}. At the same time, panel (b) reveals that this can be attributed to an onset of a WBP, as the second Rényi entropy spikes up to the Page value. Finally, panel (c) shows that the gradient norm also is convergent, though at non-zero value. We attribute this to the fact that the system gets trapped in the WBP region.

As suggested by Algorithm 1, we thus decrease the learning rate to $\eta = 0.1$ and start again. This time a WBP is avoided, the algorithm however gets stuck in a local minimum with large entanglement entropy. In this instance a choice of parameter α that defines an onset of a WBP in Definition 2 being smaller than one may be beneficial. For instance, setting $\alpha = 0.5$ could help avoiding the suboptimal local minima characterized by large entanglement, see gray dashed line in Fig. 5(b). Note that the large gradient persistent after many iterations for the blue line in Fig. 5(c) may also indicate that the learning rate is chosen too large for the width of the local minima.

Provided that our algorithm uses $\alpha = 0.5$, the system would satisfy a WBP condition even for learning rate $\eta = 0.1$, forcing us to restart the algorithm with an even smaller learning rate. Setting $\eta = 0.01$, we see that the algorithm is now able to converge very close to the true ground-state energy [violet line in Figs. 5(a)–5(c)]. In particular, the norm of the gradient assumes the smallest value among all learning rates. We note, that the further decrease of the learning rate (i.e., to $\eta = 0.001$) degrades the performance of GD. While WBPs are not encountered during the optimization process, the GD optimization converges slower and within the considered number of iterations leads to a larger energy expectation value. This highlights the fact that it is best to choose the highest possible learning rate, that still avoids a WBP. We speculate, that an optimization strategy that adapts the learning rate at each optimization step would give the best performance, though testing this assumption is beyond the scope of the present work.

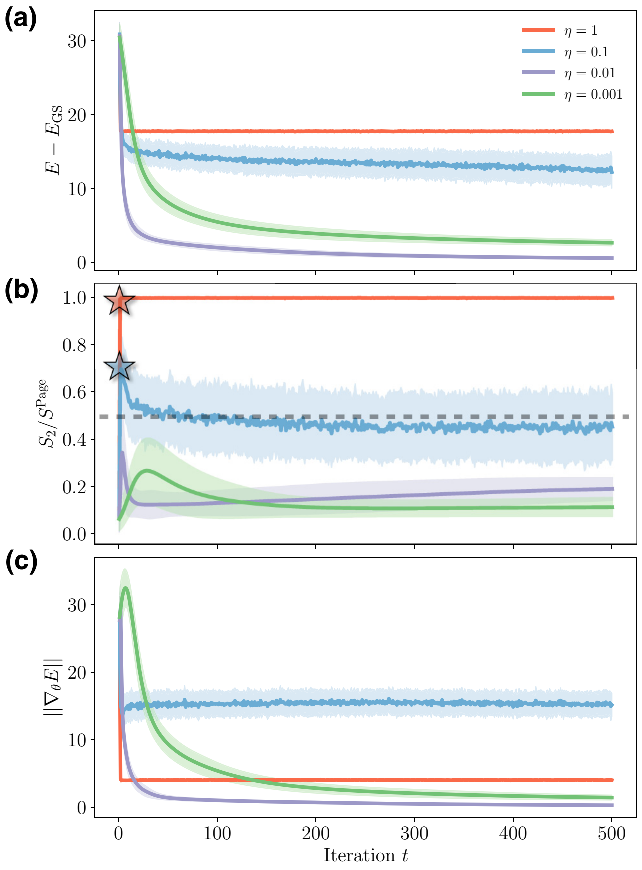

FIG. 5. (a)–(c) The application of the proposed algorithm to the problem of finding the ground state of the Heisenberg model. For large learning rates $\eta = 1$ and 0.1 (red and blue lines) the optimization gets into a large entanglement region as is shown in (b), indicated by colored stars, forcing the restart of the optimization with smaller value of η. For $\eta = 0.01$ the algorithm avoids large entanglement region and gets a good approximation for the ground state. Finally, setting even smaller learning rate (green lines) degrades the performance. The normalized second Rényi entropy of the true ground state is $S_2/S^{\mathrm{Page}}(k, N) \approx 0.246$. (c) Shows the corresponding gradient norm. A small gradient norm equally corresponds to the BP and the good local minima found with $\eta = 0.01$ and 0.001. We use a system size of $N = 10$, subsystem size $k = 2$, and a random circuit [see Eq. (1)] with circuit depth $p = 100$ and small qubit rotations ($\epsilon_\theta = 0.05$) to generate a BP-free initialization. Here we choose $\alpha = 0.5$ indicated by the gray dashed line, see the last paragraph of Sec. III A for a discussion on the choice of α. Data is averaged over 100 random instances.

C. Classical simulatability and performance comparison

Now that we have illustrated the procedure outlined in Algorithm 1 in detail, let us comment on the restrictions that our algorithm imposes, its relation to classical simulatability and finally compare our method with other common means for mitigating BPs.

To avoid WBPs and thus BPs we require that the second Rényi entropy of a small subregion is less than a fraction α of the Page value, where $\alpha \in (0, 1]$ and the default choice is $\alpha = 1$. While this restriction does place a limitation on the entanglement generated by the circuit for a region of k qubits, it does not imply classical simulatability of the circuit. Indeed, it is the scaling of the entanglement entropy with system size that is important for classical simulatability of a quantum system. Only in the special case when the entanglement entropy of the quantum state scales poly-logarithmically with the number of qubits, we can simulate the states on a classical computer in polynomial time [48–50]. In contrast, the criteria for WBP, Definition 2 is generally consistent with volume-law entanglement as we illustrate below, thus allowing our algorithm to be applied to systems that cannot be efficiently simulated on a classical computer.

Here we focus on two types of systems: namely systems where the ground state satisfies area law, which implies that the entanglement entropy of an arbitary bipartition of the state scales with the size of the boundary $S(\rho_A) \sim |\partial A|$, as well as volume law, which implies that it scales with the volume, $S(\rho_A) \sim |A|$ (see Ref. [51] for a review on these concepts). For area-law states in 1D the entanglement entropy is constant and therefore allows for an efficient classical representation using techniques such as matrix product states [52]. The 1D Heisenberg model, considered in the previous subsection, is an example for such a system.

The Heisenberg model, however, can be made hard to simulate classically by considering a random-graph geometry illustrated in Fig. 6(a), instead of a 1D chain. This leads to nonlocal interactions and a volume-law entanglement scaling for a typical bipartite cut. Due to the nonlocal nature of the model we choose $\alpha = 1$ since we have no prior knowledge on the entanglement properties of the ground state. We again use the small-angle initialization [15,26] to generate a BP-free initial state. We compare this with layerwise optimization [13], which is another common heuristic for avoiding BPs. There the circuit is initialized with a single layer, which is optimized, the circuit is then grown by one layer at a time and optimized while keeping the parameters in the previous layers constant.

Figures 6(b) and 6(c) reveal that for the Heisenberg model on a graph layerwise optimization ends up in a WBP during the optimization for both learning rates that we considered. The small-angle initialization successfully avoids the WBP for both learning rates, however good convergence is only achieved with $\eta = 0.01$. This is similar to the situation encountered in the Heisenberg model in 1D, see Fig. 7 where a too large learning rate prevents convergence to the basin of attraction of the local minimum. Likewise to the case of the 1D Heisenberg model, the fact that learning rate $\eta = 0.1$ does not lead to convergence to a minimum can be revealed through the norm of the gradient, which stays large even after 500 iterations.

In addition to the Heisenberg model on the random graph, we also considered the SYK model [53] that features a volume-law entangled ground state [54]. In Appendix E we illustrate that our method is also successful in preventing the BP occurrence and results in finding the SYK ground state.

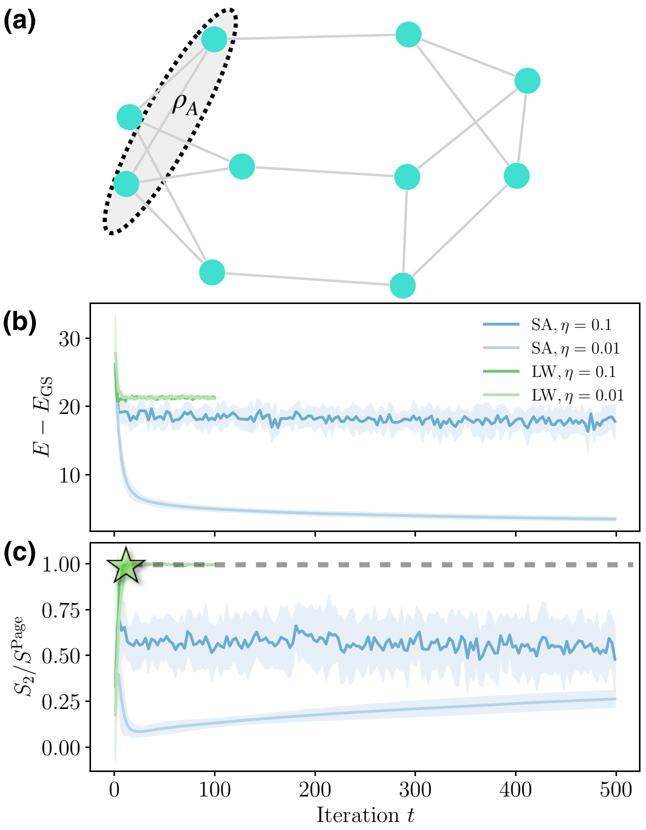

FIG. 6. Application of our algorithm to the problem of finding the ground state for the Heisenberg model on a 3-regular random graph depicted in (a). Panel (b) shows the energy as a function of GD iterations t and panel (c) illustrates the second Rényi entropy of two-spin region A with $k = 2$ shown in panel (a). Since the interactions are now nonlocal and we do not have any prior knowledge on the entanglement properties of the target state we set $\alpha = 1$ (gray dashed line). For the initialization we use the small-angle initialization (SA) with $\epsilon_\theta = 0.1$ and compare it to layerwise optimization (LW). LW encounters a WBP for both learning rates that we consider (green star). In contrast, SA avoids the WBP for both learning rates. Good performance and further convergence in the local minimum is only achieved through a smaller learning rate of $\eta = 0.01$. We use a system size of $N = 10$ and a random circuit from Eq. (1) with circuit depth $p = 100$. Data is averaged over 100 random instances.

V. SUMMARY AND DISCUSSION

The main result of this work is the introduction of the concept of WBPs, which in essence provides an efficiently detectable version of BPs. In particular, we propose to use the classical shadows protocol to estimate the second Rényi entropy of small subregions that are independent of system size. If these subregions avoid nearly maximal entanglement—a condition sufficient for avoiding WBPs—the system also avoids conventional BPs. Building on this definition of the WBP, we proposed an algorithm that is capable of avoiding BPs on NISQ devices without requiring a computational overhead that scales exponentially in system size.

In order to illustrate the notion of WBPs and the proposed algorithm, we studied a particular BP-free initialization of the variational quantum eigensolver. Furthermore, we considered an optimization procedure that uses gradient descent. Phenomenologically, we observed that the encounter of a BP during the optimization crucially depends on the learning rate, which controls the parameter update magnitude between consecutive optimization steps. A smaller learning rate is less likely to lead to the encounter of a BP during the optimization. However, choosing the learning rate to be very small degrades the performance of GD. These results support the feasibility of the proposed algorithm for efficiently avoiding BPs on NISQ devices. While our results and numerical simulations are focused on VQEs, they readily extend to other variational hybrid algorithms, such as quantum machine learning [8,55,56], quantum optimization [6,57,58], or variational time evolution [59,60].

Although the issue of avoiding BPs at the circuit initialization is a subject of active research [12–16], the influence and role of BPs in the optimization process has received much less attention [61]. Our results indicate that entanglement, in addition to playing a crucial role for circumventing BPs at the launch of the VQE, is also important for achieving a good optimization performance. In addition, our heuristic results in Sec. IV suggest that postselection based on the entanglement of small subregions may help to avoid low-quality local minima that are characterized by higher entanglement. Algorithm 1 allows for such postselection by appropriately tuning the value of α. Doing so, however, requires some prior knowledge about the entanglement structure of the target state. This may be inferred from the structure of the Hamiltonian (for instance, for a Hamiltonian that is diagonal in the computational basis, the eigenstates are product states with no entanglement), or by targeting small instances of the computational problem using exact diagonalization.

Beyond that, one could imagine an algorithm where the learning rate is not only adapted when a WBP is encountered, but dynamically adjusted at every step of the optimization process. This may allow for efficiently maneuvering complicated optimization landscapes by staying clear of highly entangled local minima. VQE, for instance, is known to have many local minima [11], but a systematic study of their entanglement structure, required for devising such a dynamic entanglement postselection procedure, has yet to be done.

Another important question concerns the effect of noise, which has been suggested to be an additional source for the emergence of BPs [20]. Noise cannot be avoided on NISQ machines and has a profound impact on any near-term quantum algorithm, which is difficult to analyze analytically. Fortunately, none of the tools we propose are especially susceptible to noise corruption. In fact, both the classical shadow protocol and the estimation of observables and purities are stable with respect to the addition of a small but finite amount of noise, and there have even been some proposals for noise mitigation techniques [62,63].

Finally, we comment on the possibility of testing Algorithm 1 on a real NISQ device. While the shadows protocol can readily be implemented on near-term devices to diagnose WBPs, whether a variational circuit with enough entangling layers that lead to a BP can be realized on a NISQ device is not entirely clear at this stage. Nevertheless recent results of Ref. [64] observed convergence of the out-of-time correlators to zero, indicating that a 2-design might already have been reached. This implies that large entanglement, as present in a BP, could be realizable on available NISQ devices, and opens the door to experimental studies of the effect of entanglement on the optimization performance on current NISQ machines using the proposed shadows protocol.

ACKNOWLEDGMENTS

We thank Marco Cerezo, Zoe Holmes, and Nicholas Hunter-Jones for fruitful discussion and valuable feedback. We also acknowledge Adam Smith, Johannes Jakob Meyer, and Victor V. Albert for comments on the paper. The simulations were performed in the Julia programming language [65] using the Yao module [66]. S.H.S., R.A.M., A.A.M. and M.S. acknowledge support by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 850899).

APPENDIX A: CLASSICAL SHADOWS AND IMPLEMENTATION DETAILS

Shadow tomography attempts to directly estimate interesting properties of an unknown state without performing full state tomography as an intermediate step. Aaronson [67] and Aaronson and Rothblum [68] showcased that such a direct estimation protocol can be exponentially more efficient, both in terms of Hilbert-space dimension (2^N in our case) and in the number of target properties (we use L to denote this cardinality). These techniques do, however, require copies of the underlying quantum state to be stored in parallel within a quantum memory and highly entangled gates to be performed on all copies simultaneously. This is too demanding for current and near-term quantum devices.

Huang *et al.* [25] developed a more near-term friendly variant of this general idea known as prediction with *classical shadows*. Similar ideas have been independently proposed by Paini and Kalev [69] and Morris and Dakić [70], respectively. As explained in detail below, the key idea is to sequentially generate state copies and perform randomly selected single-qubit Pauli measurements. Such measurements can be routinely implemented in current quantum hardware and enable the prediction of many (linear and polynomial) properties of the underlying quantum state. Importantly, the measurement budget (number of required measurements) still scales logarithmically in the number of target properties L, but it may scale exponentially in the support size k of these properties. This is not a problem for local features, like subsystem purities or terms in a quantum many-body Hamiltonian, but does prevent us from directly estimating global state features like fidelity estimation.

The general measurement budget that is required to simultaneously estimate L local observables using classical shadows, necessary for the energy expectation value estimation, is provided in Theorem 3. Typically the estimation of L observables would scale linearly in L (essentially every term is estimated individually). This is traded with a $\ln L$ dependence instead and an exponential dependence on the support k of the operators. The cost for estimating the subsystem purities and thus second Rényi entanglement entropies is provided in Eq. (A6) and is exponential in k (this dependence was recently proven to be unavoidable [34]). However since for the WBP check outlined in the main text k is small, this is generally an efficient operation. Lastly, the cost for estimating the gradients is given in Eq. (A8). The efficiency of using classical shadows to estimate the energy expectation value and gradients is system dependent (see Ref. [25] for the application of classical shadow tomography to the lattice Schwinger model). For the estimation of the purities, the shadow protocol, however, generally provides the most efficient technique currently available [71]. One possibility to circumvent these restrictions is to use a hybrid scheme where the energy and gradients are estimated with either classical shadows or the usual approach dependent on the structure of the Hamiltonian while the second Rényi entropies for the WBP check are always estimated using classical shadows.

1. Data acquisition via classical shadows

We use randomized single-qubit measurements to extract information about a variational N-qubit state represented by a density matrix

$$\rho(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})| \quad \text{with} \boldsymbol{\theta} \in \mathbb{R}^m.$$

To this end, we repeat the following procedure a total of T times. For $1 \leq t \leq T$ we carry out the following.

1. Prepare quantum state $\rho(\boldsymbol{\theta})$ on the NISQ device.
2. Select N single-qubit Pauli observables independently and uniformly at random.
3. Perform the associated N-qubit Pauli measurement (single shot) to obtain N classical bits (0 if we measure "spin down" and 1 if we measure "spin up").
4. Store N single-qubit "postmeasurement" states, $|s_i^{(t)}\rangle$, where an ith qubit measurement outcome, s_i, can take six possible values denoted as $|0\rangle$, $|1\rangle$ if qubit is measured in z basis, $|+\rangle$ and $|-\rangle$ for x basis, and, finally, $|+\mathrm{i}\rangle$ and $|-\mathrm{i}\rangle$ for y basis. Here, $|\pm\rangle = (|0\rangle \pm |1\rangle)/\sqrt{2}$ denote Pauli-x matrix eigenstates and $|\pm \mathrm{i}\rangle = (|0\rangle \pm \mathrm{i}|1\rangle)/\sqrt{2}$ are two Pauli-y eigenstates. In practice, this is achieved by applying random single-qubit Clifford gates that effectively implement a change of basis such that the usual z-basis measurement can be used, see Fig. 1 (a) for a visualization.
5. (Implicitly) Construct the N-qubit *classical shadow*

$$\hat{\rho}^{(t)}(\boldsymbol{\theta}) = \bigotimes_{i=1}^{N} \left(3|s_i^{(t)}\rangle\langle s_i^{(t)}| - \mathbb{I}\right). \tag{A1}$$

Repeating this procedure a total of T times provides us with T classical shadows $\rho^{(1)}(\boldsymbol{\theta}), \ldots, \rho^{(T)}(\boldsymbol{\theta})$. These are random matrices that are statistically independent (because they are constructed from independent quantum measurements). By construction, each classical shadow reproduces the true underlying state in expectation (over both the choice of Pauli observable and the observed spin direction):

$$\mathbb{E}\left[\hat{\rho}^{(t)}(\boldsymbol{\theta})\right] = \rho(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|, \tag{A2}$$

see, e.g., Ref. [25, Proposition S.2]. We can now approximate this ideal expectation value by empirical averaging over all samples:

$$\rho(\boldsymbol{\theta}) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{\rho}^{(t)}(\boldsymbol{\theta}).$$

This approximation becomes exact in the limit $T \to \infty$ of infinitely many measurement repetitions. But the main results in Refs. [25,69] highlight that convergence actually happens much more rapidly.

This is, in particular, true for subsystem density matrices. The tensor product structure of classical shadows, Eq. (A1), plays nicely with taking partial traces. Let $A \subseteq \{1, \ldots, N\}$ be a collection of $|A| = k$ qubits. Then,

$$\hat{\rho}_A^{(t)}(\boldsymbol{\theta}) = \mathrm{tr}_{\neg A}\left(\hat{\rho}_A^{(t)}\right) \tag{A3}$$

is a k qubit shadow that can be used to approximate the associated subsystem density matrix. More precisely, Eq. (A2) asserts

$$\mathbb{E}\left[\rho_A^{(t)}(\boldsymbol{\theta})\right] = \mathrm{tr}_{\neg A}\left(\mathbb{E}\left[\hat{\rho}^{(t)}(\boldsymbol{\theta})\right]\right) = \mathrm{tr}_{\neg A}(\rho(\boldsymbol{\theta})) = \rho_A(\boldsymbol{\theta}), \tag{A4}$$

which can (and should) form the basis of empirical averaging directly for the subsystem in question. Here is a mathematically rigorous result in this direction. In what follows, the range (or weight) of an observable is the number of qubits on which it acts nontrivially. For example, coupling terms in the Heisenberg Hamiltonian (2) have range $k=2$, while the external field terms have range $k=1$.

Theorem 3. *Fix a collection of L range-k observables O_l, as well as parameters $\epsilon, \delta > 0$. Then, with probability (at least) $1-\delta$, classical shadows of size*

$$T \geq \frac{4^{k+1}\ln(2L/\delta)}{\epsilon^2}$$

suffice to jointly estimate all L expectation values up to additive accuracy ϵ. In other words,

$$\hat{\rho}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\hat{\rho}^{(t)}(\boldsymbol{\theta}) \text{ obeys } \left|\mathrm{tr}\left(O_l\hat{\rho}(\boldsymbol{\theta})\right)\right. \\ \left. - \mathrm{tr}\left(O_l\rho(\boldsymbol{\theta})\right)\right| \leq \epsilon,$$

for all $1 \leq l \leq L$.

We emphasize that it is not necessary to form global shadow approximations. If O_l only acts nontrivially on subsystem $A_l \subseteq \{1,\ldots,N\}$ ($O_l = \tilde{O}_l \otimes \mathbb{I}_{\neg A_l}$), then $\mathrm{tr}\left(O_l\hat{\rho}(\boldsymbol{\theta})\right) = \mathrm{tr}\left(\hat{O}_l\hat{\rho}_{A_l}\right)$. Theorem 3 is slightly stronger than a related result in Ref. [25] (it does not require median-of-means estimation). Conceptually similar results have been established in Refs. [72] and [73,74]. Notably, the authors of Ref. [75] pointed out to us that they provided a similar statement as in Theorem 3 in their work. We present a formal proof in Appendix A 5 below.

2. Estimating subsystem purities

Suppose we are interested of estimating a collection of multiple subsystem purities

$$p_A(\boldsymbol{\theta}) = \mathrm{tr}\left(\rho_A(\boldsymbol{\theta})^2\right) = \mathrm{tr}\left(\rho_A(\boldsymbol{\theta})\rho_A(\boldsymbol{\theta})\right), \tag{A5}$$

where $A \subseteq \{1,\ldots,N\}$ labels different subsystems of size $|A| = k$ each. Then, we can use the corresponding subsystem shadows, Eq. (A3), to approximate each p_A by empirical averaging:

$$\hat{p}_A(\boldsymbol{\theta}) = \frac{1}{T(T-1)}\sum_{t\neq t'}\mathrm{tr}\left(\hat{\rho}_A^t\hat{\rho}_A^{t'}\right). \tag{A6}$$

It is important that we restrict our averaging operation to distinct pairs of classical shadows ($t \neq t'$). This guarantees that the expectation values factorize, i.e.,

$$\mathbb{E}\left[\hat{\rho}_A^t\hat{\rho}_A^{t'}\right] = \mathbb{E}\left[\hat{\rho}_A^t\right]\mathbb{E}[\hat{\rho}_A^{t'}] = \rho_A^2,$$

where the last equality is due to Eq. (A3). Formula (A5) is an empirical average over all distinct shadow pairs contained in the data set. It converges to the true average $p_A(\boldsymbol{\theta}) = \mathbb{E}\left[\hat{p}_A(\boldsymbol{\theta})\right]$, and the speed of convergence is governed by the variance. As data size T increases, this variance decays as

$$\mathrm{Var}\left[\hat{p}_A(\boldsymbol{\theta})\right] \leq \frac{2}{T}\left(2\times 4^k p_2(\boldsymbol{\theta}) + \frac{1}{T-1}2^{4k}\right),$$

see, e.g., Ref. [76, SM Eq. (12)]. In the large-T limit, this expression is dominated by the first term in parentheses, $4\times 2^k p_2(\boldsymbol{\theta})/T$, and Chebyshev's inequality allows us to bound the probability of a large approximation error. For $\epsilon > 0$,

$$\Pr\left[\left|\hat{p}_A(\boldsymbol{\theta}) - \mathrm{tr}\left(\rho_A(\boldsymbol{\theta})^2\right)\right| \geq \epsilon\right] \lesssim \frac{4^{k+1}\mathrm{tr}\left(\rho_A^2\right)}{T\epsilon^2},$$

provided that the total number of measurements T is large enough to suppress the higher-order contribution in the variance bound (this is why we write $\lesssim$). In this regime, a measurement budget that scales as

$$T \geq \frac{4^{k+1}\mathrm{tr}\left(\rho_A^2\right)}{\epsilon^2\delta} \tag{A7}$$

suppresses the probability of a sizable approximation error ($\geq \epsilon$) below δ. It is worthwhile to point out that this bound depends on the subsystem purity under consideration. Smaller purities are cheaper to estimate than large ones. It is also important to note that the accuracy parameter ϵ has to be small enough in order to accurately capture the purity in the WBP regime, which decays exponentially fast, but only with the subsystem size k.

The δ dependence in Eq. (A6) can be further improved to $\ln(1/\delta)$ by replacing simple empirical averaging in Eq. (A5) by median-of-means estimation [25]. Doing so would allow us to estimate all possible $L = \binom{N}{k} \leq N^k$ size-k subsystem purities with only a $k\ln N$ overhead. Median-of-means estimation does, however, worsen the dependence on ϵ by a constant amount. Empirical studies conducted in Ref. [77] showcase that such a trade-off only becomes viable if one wishes to approximate polynomially many subsystem purities.

3. Estimating gradients

To perform the GD update step suggested in Algorithm 1 we require the knowledge of gradient $\nabla_{\boldsymbol{\theta}}E(\boldsymbol{\theta})$, which consists of pN derivatives $\partial_{i,l}E(\boldsymbol{\theta})$. The derivative can naively be approximated using finite difference, though for variational single-qubit rotation gates, as used in the main text [see Eq. (1)], we can use the parameter-shift rule to compute the gradients exactly (up to finite sampling errors) [39,40]. The parameter-shift rule is given by

$$\partial_{i,l}E\left(\boldsymbol{\theta}\right) = \frac{1}{2}\left(E\left(\boldsymbol{\theta} + (\pi/2)\boldsymbol{e}_{i,l}\right) - E\left(\boldsymbol{\theta} - (\pi/2)\boldsymbol{e}_{i,l}\right)\right),$$

where i labels the qubits and l cycles through all circuit layers, and $\boldsymbol{e}_{i,l}$ is the unit vector. In order to approximate a single gradient, we need to estimate the difference of two energy expectation values $E(\boldsymbol{\theta}_+) = \langle\psi(\boldsymbol{\theta}_+)|H|\psi(\boldsymbol{\theta}_+)\rangle$ with $\boldsymbol{\theta}_+ = \boldsymbol{\theta} + (\pi/2)\boldsymbol{e}_{i,l}$ and $E(\boldsymbol{\theta}_-) = \langle\psi(\boldsymbol{\theta}_-)|H|\psi(\boldsymbol{\theta}_-)\rangle$ with $\boldsymbol{\theta}_- = \boldsymbol{\theta} - (\pi/2)\boldsymbol{e}_{i,l}$ (we suppress i and l indices in $\boldsymbol{\theta}_\pm$ for the sake of brevity). Typically, the Hamiltonian itself can be decomposed into a sum of L "simple" terms: $H = \sum_{l=1}^{L} h_l$, where often L can be proportional to the number of qubits, N. This allows expression of the gradient as a linear combination of $2L$ expectation values,

$$\begin{aligned}\partial_{i,l}E\left(\boldsymbol{\theta}\right) = \frac{1}{2}\sum_{l=1}^{L}&\left(\langle\psi(\boldsymbol{\theta}_+)|h_l|\psi(\boldsymbol{\theta}_+)\rangle\right.\\ &\left.-\ \langle\psi(\boldsymbol{\theta}_-)|h_l|\psi(\boldsymbol{\theta}_-)\rangle\right),\end{aligned} \tag{A8}$$

each of which can be estimated by performing a collection of single-qubit Pauli measurements. If each term h_l is supported on (at most) k qubits, then Theorem 3 applies. Performing $T \approx 4^k \ln(L/\delta)/\epsilon^2$ randomized Pauli measurements on state $\rho(\boldsymbol{\theta}_+)$ and $\rho(\boldsymbol{\theta}_-)$ each allows us to ϵ approximate all $2L$ simple terms in Eq. (A7).

Unfortunately, approximation errors may accumulate when taking the sum over all $2L$ terms. Suppose that we obtain ϵ-accurate estimators $\hat{E}_l(\boldsymbol{\theta}_\pm)$ of contribution of the local Hamiltonian term to the energy $E_l(\boldsymbol{\theta}_\pm) = \langle\psi(\boldsymbol{\theta}_\pm)|h_l|\psi(\boldsymbol{\theta}_\pm)\rangle$. A triangle inequality over all approximation errors then produces only

$$\begin{aligned}&\left|\partial_{i,l}E(\boldsymbol{\theta}) - \hat{\partial}_{i,l}E(\boldsymbol{\theta})\right|\\ &\quad= \frac{1}{2}\left|\sum_{l=1}^{L}\left(\hat{E}_l(\boldsymbol{\theta}_+) - E_l(\boldsymbol{\theta}_+) - \hat{E}_l(\boldsymbol{\theta}_-) + E_l(\boldsymbol{\theta}_-\right)\right|\\ &\quad\leq \frac{1}{2}\sum_{l=1}^{L}\left|\hat{E}_l(\boldsymbol{\theta}_+) - E_l(\boldsymbol{\theta}_+)\right| + \frac{1}{2}\sum_{l=1}^{L}\left|\hat{E}_l(\boldsymbol{\theta}_-) - E_l(\boldsymbol{\theta}_-)\right|\\ &\quad= L\epsilon.\end{aligned}$$

This upper bound equals only ϵ if we rescale the accuracy of original approximation to ϵ/L. Inserting this rescaled accuracy into Theorem 3 produces an overall measurement cost of

$$T \geq \frac{4^{k+1}L^2\ln(2L/\delta)}{\epsilon^2}. \tag{A9}$$

The number L of terms in the Hamiltonian typically scales (at least) linearly in the number of qubits N. This implies that the measurement budget, Eq. (A8), required to (conservatively) estimate gradients scales quadratically in the system size and thus is parametrically larger than the (conservative) cost of estimating purities of size-k subsystems, Eq. (A6). To obtain the full gradient $\nabla_{\boldsymbol{\theta}}E(\boldsymbol{\theta})$ the procedure has to be repeated pN times since the parameter-shift rule has to implemented for every variational parameter. It should be noted though, that in principle this can be computed in parallel, provided large enough (quantum) computational resources. For example, different NISQ computers could be used to estimate different gradient components at the same time.

4. Example of error accumulation in an Ising model

The extra scaling with L^2 in Eq. (A8) is a consequence of error accumulation. If we use the same measurement data to jointly estimate many Hamiltonian terms, then all these estimators become highly correlated. And the effect of outlier corruption—which occurs naturally in empirical estimation—becomes amplified.

Here, we illustrate this subtlety by means of a simple example. Let $H = -J\sum_{i=1}^{N-1}\sigma_i^z\sigma_{i+1}^z$ be the Ising Hamiltonian on a 1D chain comprised of N qubits ($L = N-1$). Let us also assume that N is even. This Hamiltonian is diagonal in the z basis $|i_1,\ldots,i_N\rangle = |i_1\rangle\otimes\cdots\otimes|i_N\rangle$ with $i_1,\ldots,i_N \in \{0,1\}$. So, in order to estimate H, it suffices to perform measurements solely in this basis. Born's rule asserts, that we observe bitstring $\hat{s}_1,\ldots,\hat{s}_N$ with probability

$$\Pr\left[\hat{s}_1,\ldots,\hat{s}_N\right] = \langle\hat{s}_1,\ldots,\hat{s}_N|\rho|\hat{s}_1,\ldots,\hat{s}_N\rangle,$$

where ρ denotes the underlying N-qubit state. And, we can use these outcomes to directly estimate the total energy. It is easy to check that

$$\begin{aligned}\hat{E} =&\langle\hat{s}_1,\ldots,\hat{s}_N|H|\hat{s}_1,\ldots,\hat{s}_N\rangle\\ =&-J\sum_{i=1}^{N}\langle\hat{s}_i|\sigma_i^z|\hat{s}_i\rangle\langle\hat{s}_{i+1}|\sigma_{i+1}^z|\hat{s}_{i+1}\rangle\end{aligned}$$

obeys $\mathbb{E}\left[\hat{E}\right] = \mathrm{tr}\left(H\rho\right)$, regardless of the quantum state ρ in question. Also, estimating individual terms in this sum is both cheap and easy. Convergence of the sum, however, does depend on the underlying quantum state and the correlations within. To illustrate this, we choose $\lambda \in (0,1)$ and set

$$\rho(\lambda) = (1-\lambda)|\psi\rangle\langle\psi| + \lambda|\phi\rangle\langle\phi|,$$

where $|\psi\rangle = |00\cdots 00\rangle$ is the Ising ground state and $|\phi\rangle = |01\cdots 01\rangle$ is a Néel state. These states obey $\langle\psi|H|\psi\rangle = -J(N-1)$ (ground state) and $\langle\phi|H|\phi\rangle = +J(N-1)$ (highest excited state), so

$$\mathrm{tr}(H\rho(\lambda)) = -J(n-1)\,(1-2\lambda)\,.$$

The task is to approximate this expectation value based on computational basis measurements. For each measurement, we either obtain outcome $0\cdots 0$ (with probability $1-p$) or outcome $01\cdots 01$ (with probability p). This dichotomy extends to our estimator

$$\hat{E} = \begin{cases} \langle\psi|H|\psi\rangle = -J(N-1) & \text{with prob. } 1-\lambda, \\ \langle\phi|H|\phi\rangle = +J(N-1) & \text{with prob. } \lambda, \end{cases}$$

and we are effectively faced with estimating the (rescaled) expectation value of a biased coin. The associated variance of such a coin toss can be easily computed and amounts to

$$\mathrm{Var}\left[\hat{E}\right] = \mathbb{E}\left[\hat{E}^2\right] - \left(\mathbb{E}\left[\hat{E}\right]\right)^2 = 4J^2(N-1)^2\lambda(1-\lambda).$$

Unless $\lambda \neq 0, 1$ (where the variance vanishes), this variance it is proportional to $L^2 = (N-1)^2$ and controls the rate of convergence. Asymptotically, a total number of

$$T \geq \mathrm{Var}\left[\hat{E}\right]/\epsilon^2 = 4J^2L^2\lambda(1-\lambda)/\epsilon^2 = \Omega(L^2/\epsilon^2)$$

independent coin tosses are necessary (and sufficient) to ϵ approximate the true expectation value $\mathbb{E}\left[\hat{E}\right] = \mathrm{tr}\,(\rho(\lambda)H)$. This is a consequence of the central limit theorem and showcases that a measurement budget scaling with the number L of Hamiltonian terms is unavoidable in general.

We emphasize that this is a contrived worst-case argument that showcases how correlated measurements can affect the approximation quality of a sum of many simple terms, while each term individually is cheap and easy to evaluate. A generalization to the Heisenberg Hamiltonian considerable in the main text, see Eq. (2), is straightforward.

5. Proof of Theorem 3

Theorem 3 is a consequence of the following concentration inequality. Let $\|O\|_\infty$ denote the operator and spectral norm of an observable. We also use $\|\cdot\|_1$ to denote the trace norm.

Theorem 4. *Fix a collection of L range-k observables O_l with $\|O_l\|_\infty \leq 1$, a quantum state ρ and let $\hat{\rho} = 1/T\sum_{t=1}^T \hat{\rho}^{(t)}$ be a classical shadow estimate thereof. Then, for $\epsilon \in (0,1)$,*

$$\Pr\left[\max_{1\leq l\leq L}\left|\mathrm{tr}\left(O_l\hat{\rho}\right) - \mathrm{tr}\left(O_l\rho\right)\right| \geq \epsilon\right] \leq 2L\exp\left(-\frac{\epsilon^2 T}{4^{k+1}}\right).$$

This large deviation bound is a consequence of another well-known tail bound, see, e.g., Ref. [78, Theorem 7.30].

Theorem 5 ((Bernstein inequality))**.** *Let $X^{(1)},\ldots,X^{(T)}$ be independent, centered (i.e., $\mathbb{E}[X_t] = 0$) random variables that obey $|X^{(t)}| \leq R$ almost surely. Then, for $\epsilon > 0$*

$$\Pr\left[\left|\frac{1}{T}\sum_{t=1}^T X^{(t)}\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2T^2/2}{\sigma^2 + RT\epsilon}\right),$$

where $\sigma^2 = \sum_{t=1}^T \mathbb{E}\left[\left(X^{(t)}\right)^2\right]$.

Proof of Theorem 4. Fix an observable $O = O_l$ with $1 \leq l \leq L$ and define $X^{(t)} = \mathrm{tr}\left(O\hat{\rho}^{(t)}\right) - \mathrm{tr}\left(O\rho\right)$. Then, by construction of classical shadows, each $X^{(t)}$ is an independent random variable that also obeys $\mathbb{E}\left[X^{(t)}\right] = 0$, courtesy of Eq. (A2). Next, let $A \subseteq \{1,\ldots,N\}$ with $|A| = k$ be the subsystem on which the range-k observable acts nontrivially, i.e., $O = O_A \otimes \mathbb{I}_{\neg A}$ and $\|O\|_\infty = \|O_A\|_\infty \leq 1$. Then, Hoelder's inequality ($|\mathrm{tr}\,(O_A\rho_A)| \leq \|O_A\|_\infty\|\rho_A\|_1$) asserts

$$\begin{aligned} \left|X^{(t)}\right| &= \left|\mathrm{tr}\left(O_A\hat{\rho}_A^{(t)}\right) - \mathrm{tr}\left(O_A\rho_A\right)\right| \\ &\leq \|O_A\|_\infty\left(\|\rho_A\|_1 + \left\|\hat{\rho}_A^{(t)}\right\|_1\right) \\ &= \|O_A\|_\infty\left(1 + \prod_{a\in A}\left\|3|s_a^{(t)}\rangle\langle s_a^{(t)}| - \mathbb{I}\right\|_1\right) \\ &\leq \left(1 + 2^{|A|}\right) = 1 + 2^k = R, \end{aligned}$$

where we also use $\|\rho_A\|_1 = \mathrm{tr}(\rho_A) = 1$ and the specific form of subsystem classical shadows, Eq. (A3), that factorizes nicely into tensor products. Estimating the variance is more difficult by comparison. However, Ref. [25, Proposition S3] asserts

$$\mathbb{E}\left[\left(X^{(t)}\right)^2\right] \leq \|O\|_{\mathrm{shadow}}^2 \leq 4^k\|O\|_\infty = 4^k.$$

In turn, $\sigma^2 \leq T4^k$ and we conclude

$$\begin{aligned}
&\Pr\left[\left|\operatorname{tr}\left(O\hat{\rho}\right) - \operatorname{tr}\left(O\rho\right)\right| \geq \epsilon\right] \\
&= \Pr\left[\left|\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right| \geq \epsilon\right] \\
&\leq 2\exp\left(-\frac{\epsilon^2 T^2/2}{T4^k + (1+2^k)T\epsilon}\right) \\
&\leq 2\exp\left(-\frac{\epsilon^2 T}{4^{k+1}}\right),
\end{aligned}$$

where the last line is a rough simplification of the exponent. Such a tail bound is valid for any $O = O_l$ and the advertised statement follows from taking a union bound (also known as Boole's inequality) over all possible deviations:

$$\begin{aligned}
&\Pr\left[\max_{1\leq l\leq L}\left|\operatorname{tr}\left(O_l\hat{\rho}\right)\right) - \operatorname{tr}\left(O_l\rho\right)\right| \geq \epsilon\right] \\
&\leq \sum_{l=1}^{L}\Pr\left[\left|\operatorname{tr}\left(O_l\hat{\rho}\right)\right) - \operatorname{tr}\left(O_l\rho\right)\right| \geq \epsilon\right] \\
&\leq 2L\exp\left(-\frac{\epsilon^2 T}{4^{k+1}}\right).
\end{aligned}$$

■

APPENDIX B: UNITARY *t*-DESIGNS

Here, we briefly review the notion of unitary *t*-designs. The Haar measure is the unique left and right invariant measure on the unitary group $U(d)$, where d here stands for the dimension of the full Hilbert space, $d = 2^N$. Unitary t-designs are ensembles of unitaries that approximate moments of the Haar measure. More precisely, let $\mathcal{E}$ be an ensemble of unitaries, i.e., a subset of $U(d)$ equipped with a probability measure. For an operator O acting on the t-fold Hilbert space $\mathcal{H}^{\otimes t}$, the t-fold channel with respect to $\mathcal{E}$ is defined as

$$\Phi^t_{\mathcal{E}}(O) = \int_{\mathcal{E}} dU U^{\otimes t}(O)U^{\dagger\otimes t}. \tag{B1}$$

Essentially, we are asking when the average of an operator O over the ensemble $\mathcal{E}$ equals an average over the full unitary group. A unitary t-design [79,80] is an ensemble $\mathcal{E}$ for which the t-fold channels are equal for all operators O,

$$\Phi^t_{\mathcal{E}}(O) = \Phi^t_{\text{Haar}}(O).$$

Being a t-design means we exactly capture the first t moments of the Haar measure with larger t better approximating the full unitary group. There are known constructions of t-designs for $t = 2$ and $t = 3$ [79,81–84]. For $t = 1$, it is known that any basis for the algebra of operators of $\mathcal{H}$, including the Pauli group, is a 1-design. In practice, one is more interested in when the ensemble of unitaries is close to forming a t-design. With this, given a tolerance $\epsilon_t > 0$ one refers to the ensemble $\mathcal{E}$ as being an approximate t-design if

$$\|\Phi^t_{\mathcal{E}} - \Phi^t_{\text{Haar}}\|_\diamond \leq \epsilon_t,$$

where $\|\cdot\|_\diamond$ is the diamond norm—a worst-case distance measure that is very popular in quantum information theory, see, e.g., [85]. In the quantum-machine-learning literature the distance between the two t-fold channels is known as the expressibility of the ensemble $\mathcal{E}$ [15], the smaller the distance the more expressive the ensemble is.

APPENDIX C: ENTANGLEMENT AND UNITARY 2-DESIGNS

Random unitary operators have often been used to approximate late-time quantum dynamics. In the crudest approximation, it is assumed that the unitary matrix is directly drawn from the Haar measure. Although modeling quantum dynamics by random unitaries is an approximation, it has led to new insights into black-hole physics [86–88] and produced computable models of information spreading and entanglement dynamics [89–92].

In what follows, we consider a weaker situation where the random unitary operator is drawn from an ensemble $\mathcal{E}$ forming a 2-design, and focus on the entanglement properties of N-qubits random pure states

$$|\psi\rangle = U|\psi_0\rangle, \tag{C1}$$

with $U \sim \mathcal{E}$. These results have been previously obtained, see, for example, Refs. [41–43] and references therein.

Given a bipartition $(A, \neg A)$ of the system, we begin by studying the distance of the reduced density matrix ρ_A to the maximally entangled state $\rho_A^\infty = \mathbb{I}_A/d_A$, where d_A is the dimension of the Hilbert space $\mathcal{H}_A$ associated with region A. The full Hilbert-space dimension is denoted by $d = 2^N$.

1. Bounding the expected trace distance

Let us recall the following inequality relating the 1-norm (trace distance) $\|M\|_1 = \operatorname{tr}\sqrt{M^\dagger M}$, and the 2-norm (Frobenius norm) $\|M\|_2 = \sqrt{\operatorname{tr}(M^\dagger M)}$

$$\|M\|_2 \leq \|M\|_1 \leq \sqrt{d}\|M\|_2. \tag{C2}$$

We are interested in bounding $\mathbb{E}_{\mathcal{E}}\left(\|\rho_A - \mathbb{I}_A/d_A\|_1\right)^2$. To do so we first use Jensen's inequality and afterwards employ the inequality (C2),

$$\begin{aligned}\mathbb{E}_{\mathcal{E}}\big(\|\rho_A-\mathbb{I}_A/d_A\|_1\big)^2 &\le \mathbb{E}_{\mathcal{E}}\big(\|\rho_A-\mathbb{I}_A/d_A\|_1^2\big)\\ &\le d_A\mathbb{E}_{\mathcal{E}}(\|\rho_A-\mathbb{I}_A/d_A\|_2^2). \qquad \text{(C3)}\end{aligned}$$

The last term on the right-hand side is related to the purity:

$$\mathbb{E}_{\mathcal{E}}(\|\rho_A-\mathbb{I}_A/d_A\|_2^2)=\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\rho_A^2)-1/d_A. \qquad \text{(C4)}$$

As we see, the only nontrivial dependence on U comes from the purity of the reduced density matrix. Let $\{|I\rangle = |i_A, j_{\neg A}\rangle\}_{i,j}$ be the computational basis for the Hilbert space $\mathcal{H}=\mathcal{H}_A\otimes\mathcal{H}_{\neg A}$ (such that it respects the bipartition).

Let us now proceed with the calculation of the average purity. We first compute the reduced density matrix ρ_A and write it as a sum over products of matrix elements of the unitary operator U:

$$\begin{aligned}\rho_A &= \sum_{j_{\neg A}}^{d_{\neg A}}\langle j_{\neg A}|\rho|j_{\neg A}\rangle = \sum_{j_{\neg A}}^{d_{\neg A}}\sum_{J,I}^{d}\rho_{I,K}\langle j_{\neg A}|I\rangle\langle K|j_{\neg A}\rangle,\\ &= \sum_{i_A,k_A}\sum_{j_{\neg A}}\rho_{(i_A,j_{\neg A}),(k_A,j_{\neg A})}|i_A\rangle\langle k_A|,\\ &= \sum_{i_A,k_A}\sum_{j_{\neg A}}U_{(i_A,j_{\neg A}),(0,0)}U^*_{(k_A,j_{\neg A}),(0,0)}|i_A\rangle\langle k_A|,\end{aligned}$$

where the last line follows from Eq. (C1).

Afterwards, it can be easily verified that $\mathrm{tr}(\rho_A^2)$ reads

$$\begin{aligned}\mathrm{tr}(\rho_A^2) = \sum_{i_A,k_A}\sum_{j_{\neg A},p_{\neg A}} & U_{(i_A,j_{\neg A}),(0,0)}U_{(k_A,p_{\neg A}),(0,0)}\\ &\times U^*_{(k_A,j_{\neg A}),(0,0)}U^*_{(i_A,p_{\neg A}),(0,0)}. \qquad \text{(C5)}\end{aligned}$$

Using the following identities for the first and second moment of the unitary group endowed with the Haar measure

$$\begin{aligned}&\int_{U(n)} dU_H U_{i,j}U^*_{i_1,j_1} = \delta_{i,i_1}\delta_{j,j_1}/d,\\ &\int_{U(n)} dU_H U_{i,j}U_{l,m}U^*_{i_1,j_1}U^*_{l_1,m_1}\\ &= \frac{1}{d^2-1}(\delta_{i,i_1}\delta_{l,l_1}\delta_{j,j_1}\delta_{m,m_1}+\delta_{i,l_1}\delta_{l,i_1}\delta_{j,j_1}\delta_{m,m_1})\\ &-\frac{1}{d(d^2-1)}(\delta_{i,i_1}\delta_{l,l_1}\delta_{j,m_1}\delta_{m,j_1}+\delta_{i,l_1}\delta_{l,i_1}\delta_{j,j_1}\delta_{m,m_1}), \qquad \text{(C6)}\end{aligned}$$

we get that the following simple expression for the expected purity

$$\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\rho_A^2)=\frac{d_A+d_{\neg A}}{1+d_Ad_{\neg A}}. \qquad \text{(C7)}$$

Finally, substituting Eq. (C7) into Eq. (C4) we obtain

$$\mathbb{E}_{\mathcal{E}}\big(\|\rho_A-\mathbb{I}_A/d_A\|_1\big)\le\sqrt{\frac{d_A^2-1}{d_Ad_{\neg A}+1}}\sim\mathcal{O}(\sqrt{d_A/d_{\neg A}}). \qquad \text{(C8)}$$

Note that the above result implies that when the complementary subsystem $\neg A$ is (significantly) larger than A, the expected deviation of ρ_A from the maximally mixed state is exponentially small.

2. Bounding the expected second Rényi entropy

Let us now explore the average value of the second Rényi entropy, which, as mentioned in the main text, can be easily estimated using the classical shadows protocol by Huang *et al.* [25].

Computing the exact average value of the second Rényi is a complicated task. Hence, we instead provide a lower and an upper bound for it. On one hand, via Jensen's inequality, we have that

$$-\ln\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\rho_A^2)\le\mathbb{E}_{\mathcal{E}}(S_2(\rho_A)), \qquad \text{(C9)}$$

which changes the focus of our attention to the expectation value of the purity of the reduced density matrix $\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\rho_A^2)$. Using the result from the previous subsection Eq. (C7) and taking the logarithm, we get the following lower bound:

$$-\ln\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\rho_A^2)=-\ln\frac{d_A+d_{\neg A}}{1+d_Ad_{\neg A}}. \qquad \text{(C10)}$$

Taking the large d limit and writing everything in terms of $d_A/d_{\neg A}$ we find

$$-\ln\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\rho_A^2)\approx\ln d_A-\frac{d_A}{d_{\neg A}}+\mathcal{O}\left(\frac{d_A^2}{d_{\neg A}^2}\right). \qquad \text{(C11)}$$

On the other hand, we have that for any state ρ_A the following inequality holds:

$$S_2(\rho_A)\le S(\rho_A)=-\ln\rho_A\mathrm{tr}\rho_A,$$

where $S(\rho_A)$ is the von Neumann entropy of ρ_A. Taking averages does not change this relation and we conclude $\mathbb{E}_{\mathcal{E}}(S_2(\rho_A))\le\mathbb{E}_{\mathcal{E}}(S(\rho_A))$. The expectation value of the von Neumann entropy is upper bounded by the *Page entropy*:

$$S^{\mathrm{Page}}(d_A,d)=\frac{1}{\ln 2}\Big(-\frac{d_A-1}{2}\frac{d_A}{d}+\sum_{j=d/d_A+1}^{d}\frac{1}{j}\Big). \qquad \text{(C12)}$$

Page [86] conjectured that this analytical formula accurately captures the von Neumann entropy of a Haar random state. This conjecture was subsequently proven in

Ref. [93]. Putting everything together, we obtain

$$-\ln\frac{d_A+d_{\neg A}}{1+d_A d_{\neg A}} \leq \mathbb{E}_{\mathcal{E}}(S_2(\rho_A)) \leq S^{\text{Page}}(d_A,d). \tag{C13}$$

Considering now that the number of qubits inside region A is equal to k and assuming that $d_A/d_{\neg A} = 1/2^{N-2k} \ll 1$ we arrive at the expression in Theorem 1, that is

$$k\ln 2 - \frac{1}{2^{N-2k}} \leq \mathbb{E}_{\mathcal{E}}(S_2) \leq k\ln 2 - \frac{1}{2}\frac{1}{2^{N-2k}}. \tag{C14}$$

We see that whenever the unitary ensemble $\mathcal{E}$ forms a 2-design, the expected value of the second Rényi entropy is close to the Page entropy.

APPENDIX D: ENTANGLEMENT GROWTH AND LEARNING RATE

Here we detail the derivation of Eq. (8). We first upper bound the trace distance via

$$T(\rho_A,\sigma_A) \leq T(|\psi\rangle,|\phi\rangle) = \sqrt{1-f(|\psi\rangle,|\phi\rangle)}, \tag{D1}$$

where f stands for the pure state fidelity $f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle) = |\langle\psi(\boldsymbol{\theta})||\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle|^2$. Taylor expanding the pure state fidelity around $\boldsymbol{\theta}$ we get

$$f(|\psi(\boldsymbol{\theta})\rangle,|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle) = 1 - \frac{1}{4}\boldsymbol{\delta}^T\mathcal{F}(\boldsymbol{\theta})\boldsymbol{\delta} + \mathcal{O}(\delta^4), \tag{D2}$$

where $\mathcal{F}(\boldsymbol{\theta})$ is the QFIM given by

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = 4\,\text{Re}\{\langle\partial_i\psi|\,|\partial_j\psi\rangle - \langle\partial_i\psi|\,|\psi\rangle\,\langle\psi|\,|\partial_j\psi\rangle\}. \tag{D3}$$

Assuming $\boldsymbol{\delta} \ll 1$ we can neglect higher-order terms in $\boldsymbol{\delta}$ and so

$$T(\rho_A,\sigma_A) \lesssim \sqrt{\frac{1}{4}\boldsymbol{\delta}^T\mathcal{F}(\boldsymbol{\theta})\boldsymbol{\delta}} = \sqrt{\frac{\eta^2}{4}(\nabla_{\boldsymbol{\theta}}E)^T\mathcal{F}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}E}, \tag{D4}$$

where in the last equality we plug in the parameter change under GD [Eq. (3)], $\boldsymbol{\delta} = -\eta\nabla_{\boldsymbol{\theta}}E$.

APPENDIX E: ALGORITHM PERFORMANCE FOR THE SYK MODEL

In this section we show the numerical results for the VQE applied to the ground-state search of the SYK model [53]. The SYK model provides a canonical example for a volume-law model where the ground state is nearly maximally entangled [54]. The nonlocal nature of the Hamiltonian does not allow for an efficient estimation of the energy expectation value of this model using classical shadows. Thus, this model may be viewed as a theoretical example that shows that application of our algorithm is not limited to area-law entangled states. We use a small-angle initialization as well as the identity-block initialization [12] to illustrate our method.

The SYK model is a quantum-mechanical model of $2N$ spinless Majorana fermions χ_i satisfying the following anticommutation relations $\{\chi_i,\chi_j\} = \delta_{ij}$. The SYK model was introduced by Kitaev [53] as a simplified variant of a model introduced by Sachdev and Ye [94]. The Hamiltonian of the model is

$$H_{\text{SYK}} = \sum_{i,j,k,l}^{2N} J_{i,j,l}\chi_i\chi_j\chi_k\chi_l, \tag{E1}$$

where the couplings $J_{i,j,k,l}$ are taken randomly from a Gaussian distribution with zero mean and variance

$$\text{var}[J_{i,j,k,l}] = \frac{3!}{(N-3)(N-2)(N-1)}J^2.$$

We can study Majorana fermions using spin-chain variables by a nonlocal change of basis known as the Jordan-Wigner transformation:

$$\chi_{2i} = \frac{1}{\sqrt{2}}\sigma_1^x\cdots\sigma_{i-1}^x\sigma_i^y, \quad \chi_{2i-1} = \frac{1}{\sqrt{2}}\sigma_1^x\cdots\sigma_{i-1}^x\sigma_i^z, \tag{E2}$$

such that $\{\chi_i,\chi_j\} = \delta_{i,j}$. With this representation, encoding $2N$ Majorana fermions requires N qubits. For our studies, we set $J = 1$ and consider a system of $N = 10$ qubits.

We study performance of VQE for SYK model using two different initializations. Figures 7(a) and 7(b) show that the WBP is avoided during optimization for if the learning rate is chosen appropriately. For a large learning rate ($\eta = 1$) both initializations encounter a WBP during the optimization (indicated by the gray and blue star). Once the learning rate is decreased ($\eta = 0.1$) the entanglement entropy slowly grows to the nearly maximal value associated with the ground state of the SYK model (dotted line) instead of uncontrollably reaching the Page value. For this model, it is important to use $\alpha = 1$ (the default value) such that the entanglement entropy can grow during the optimization. Only if there is some *a priori* knowledge of the properties of the ground state, α can be chosen to be smaller.

The identity block initialization [12] here leads to the best optimization performance. We attribute this to the fact that the identity block initialization allows for a faster growth in entanglement since the parameter values are highly fine tuned. Our results suggest that sensitivity of the initialization ansatz to small perturbations may be beneficial for the cases when the ground state is nearly maximally entangled. These results highlight the advantage of using

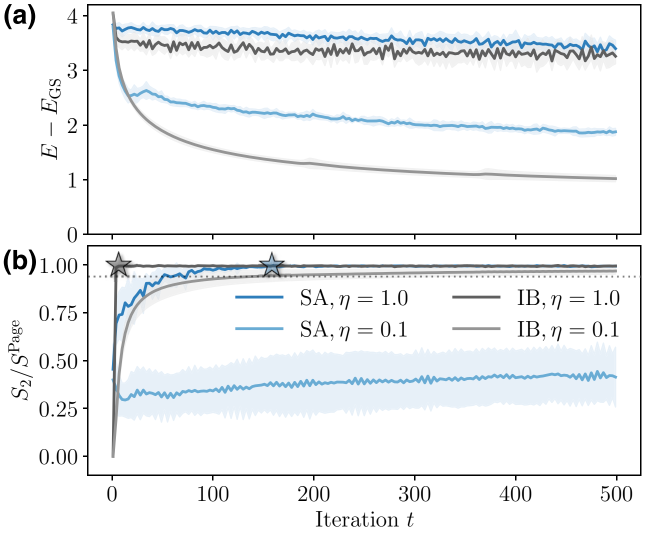

FIG. 7. (a),(b) The application of our algorithm to the problem of finding the ground state of the SYK model. For the initialization we consider the small-angle (SA) ($\epsilon_\theta = 0.1$) and identity block (IB) initialization [12] (using one block). We can see that only through the reset of the learning rate η, as suggested by Algorithm 1, WBPs are avoided during the optimization. The entanglement entropy of the target state is nearly maximal (indicated by the dotted line), we omit the WBP line for $\alpha = 1$ for improved visibility. We measure energy in units of J and use a system size of $N = 10$, subsystem size $k = 2$ and a random circuit from Eq. (1) with circuit depth $p = 100$. Data is averaged over 100 random instances.

our algorithm. The tracking of the second Rényi entanglement entropy during the optimization reveals that the larger learning rates encounter a WBP while the smaller learning rates successfully avoid it.

[1] J. Preskill, Quantum computing in the NISQ era and beyond, arXiv e-prints, arXiv:1801.00862 (2018).

[2] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, arXiv e-prints, arXiv:quant-ph/9508027 (1995).

[3] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum (NISQ) algorithms, arXiv e-prints, arXiv:2101.08448 (2021).

[4] 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 (London) **549**, 242 (2017).

[5] F. Arute *et al.*, Hartree-Fock on a superconducting qubit quantum computer, Science **369**, 1084 (2020).

[6] M. P. Harrigan *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, Nat. Phys. **17**, 332 (2021).

[7] N. Lacroix, C. Hellings, C. K. Andersen, A. Di Paolo, A. Remm, S. Lazar, S. Krinner, G. J. Norris, M. Gabureac, J. Heinsoo, A. Blais, C. Eichler, and A. Wallraff, Improving the Performance of Deep Quantum Optimization Algorithms with Continuous Gate Sets, PRX Quantum **1**, 110304 (2020).

[8] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature (London) **567**, 209 (2019).

[9] S. Johri, S. Debnath, A. Mocherla, A. Singk, A. Prakash, J. Kim, and I. Kerenidis, Nearest centroid classification on a trapped ion quantum computer, Npj Quantum Inf. **7**, 122 (2021).

[10] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[11] L. Bittel and M. Kliesch, Training Variational Quantum Algorithms Is NP-Hard, Phys. Rev. Lett. **127**, 120502 (2021).

[12] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, Quantum **3**, 214 (2019).

[13] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, Layerwise learning for quantum neural networks, arXiv e-prints, arXiv:2006.14904 (2020).

[14] J. Dborin, F. Barratt, V. Wimalaweera, L. Wright, and A. G. Green, Matrix product state pre-training for quantum machine learning, arXiv e-prints, arXiv:2106.05742 (2021).

[15] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, arXiv e-prints, arXiv:2101.02138 (2021).

[16] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, Diagnosing barren plateaus with tools from quantum optimal control, arXiv:2105.14377 [quant-ph] (2021).

[17] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1791 (2021).

[18] A. Uvarov and J. Biamonte, On barren plateaus and cost function locality in variational quantum algorithms, arXiv e-prints, arXiv:2011.10530 (2020).

[19] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement induced barren plateaus, arXiv e-prints, arXiv:2010.15968 (2020).

[20] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, arXiv e-prints, arXiv:2007.14384 (2020).

[21] J. Kim and Y. Oz, Entanglement diagnostics for efficient quantum computation, arXiv e-prints, arXiv:2102.12534 (2021).

[22] J. Kim and Y. Oz, Quantum energy landscape and VQA optimization, arXiv e-prints, arXiv:2107.10166 (2021).

[23] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, Phys. Rev. Res. **3**, 033090 (2021).

[24] R. Wiersema, C. Zhou, J. F. Carrasquilla, and Y. B. Kim, Measurement-induced entanglement phase transitions in

variational quantum circuits, arXiv e-prints, arXiv:2111.08035 (2021).

[25] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[26] T. Haug, K. Bharti, and M. S. Kim, Capacity and Quantum Geometry of Parametrized Quantum Circuits, PRX Quantum **2**, 040309 (2021).

[27] 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, Nat. Commun. **5**, 4213 (2014).

[28] M. Ostaszewski, E. Grant, and M. Benedetti, Structure optimization for parameterized quantum circuits, Quantum **5**, 391 (2021).

[29] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum Natural Gradient, Quantum **4**, 269 (2020).

[30] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv e-prints, arXiv:1412.6980 (2014).

[31] J. Gacon, C. Zoufal, G. Carleo, and S. Woerner, Simultaneous Perturbation Stochastic Approximation of the Quantum Fisher Information, Quantum **5**, 567 (2021).

[32] S. T. Flammia and Y.-K. Liu, Direct Fidelity Estimation from Few Pauli Measurements, Phys. Rev. Lett. **106**, 230501 (2011).

[33] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, and J. R. McClean, Quantum advantage in learning from experiments, arXiv e-prints, arXiv:2112.00778 (2021).

[34] S. Chen, J. Cotler, H.-Y. Huang, and J. Li, Exponential separations between learning with and without quantum memory, arXiv e-prints, arXiv:2111.05881 (2021).

[35] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, Quantum **5**, 558 (2021).

[36] I. Cirac, D. Perez-Garcia, N. Schuch, and F. Verstraete, Matrix product states and projected entangled pair states: Concepts, symmetries, and theorems, arXiv e-prints, arXiv:2011.12127 (2020).

[37] G. Verdon, M. Broughton, J. R. McClean, K. J. Sung, R. Babbush, Z. Jiang, H. Neven, and M. Mohseni, Learning to learn with quantum neural networks via classical neural networks, arXiv e-prints, arXiv:1907.05415 (2019).

[38] T. Volkoff and P. J. Coles, Large gradients via correlation in random parameterized quantum circuits, Quantum Sci. Technol. **6**, 025008 (2021).

[39] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[40] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[41] S. Popescu, A. J. Short, and A. Winter, Entanglement and the foundations of statistical mechanics, Nat. Phys. **2**, 754 (2006).

[42] R. Oliveira, O. C. O. Dahlsten, and M. B. Plenio, Generic Entanglement can be Generated Efficiently, Phys. Rev. Lett. **98**, 130502 (2007).

[43] O. C. O. Dahlsten, R. Oliveira, and M. B. Plenio, The emergence of typical entanglement in two-party random processes, J. Phys. A: Math. Theor. **40**, 8081 (2007).

[44] Z. Chen, Z. Ma, I. Nikoufar, and S.-M. Fei, Sharp continuity bounds for entropy and conditional entropy, Sci. China Phys. Mech. Astron. **60**, 020321 (2016).

[45] J. J. Meyer, Fisher information in noisy intermediate-scale quantum applications, Quantum **5**, 539 (2021).

[46] A similar continuity bound, which does not require the QFIM, can be found in terms of the maximum operator norm of the gate generators. We acknowledge Johannes Jakob Meyer for this remark.

[47] A. Rath, C. Branciard, A. Minguzzi, and B. Vermersch, Quantum Fisher Information from Randomized Measurements, Phys. Rev. Lett. **127**, 260501 (2021).

[48] G. Vidal, Efficient Classical Simulation of Slightly Entangled Quantum Computations, Phys. Rev. Lett. **91**, 147902 (2003).

[49] M. Van den Nest, W. Dür, G. Vidal, and H. J. Briegel, Classical simulation versus universality in measurement-based quantum computation, Phys. Rev. A **75**, 012337 (2007).

[50] F. G. S. L. Brandão and M. Horodecki, An area law for entanglement from exponential decay of correlations, Nat. Phys. **9**, 721 (2013).

[51] J. Eisert, M. Cramer, and M. B. Plenio, Colloquium: Area laws for the entanglement entropy, Rev. Mod. Phys. **82**, 277 (2010).

[52] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, Ann. Phys. (N. Y) **326**, 96 (2011), january 2011 Special Issue.

[53] A. Kitaev, A simple model of quantum holography (2015), talks at KITP, April 7, 2015 and May 27, 2015.

[54] Y. Huang and Y. Gu, Eigenstate entanglement in the Sachdev-Ye-Kitaev model, Phys. Rev. D **100**, 041901 (2019).

[55] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Parameterized quantum circuits as machine learning models, Quantum Sci. Technol. **4**, 043001 (2019).

[56] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).

[57] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv e-prints, arXiv:1411.4028 (2014).

[58] S. H. Sack and M. Serbyn, Quantum annealing initialization of the quantum approximate optimization algorithm, Quantum **5**, 491 (2021).

[59] S. Barison, F. Vicentini, and G. Carleo, An efficient quantum algorithm for the time evolution of parameterized circuits, Quantum **5**, 512 (2021).

[60] S.-H. Lin, R. Dilip, A. G. Green, A. Smith, and F. Pollmann, Real- and Imaginary-Time Evolution with Compressed Quantum Circuits, PRX Quantum **2**, 010342 (2021).

[61] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, Theory of overparametrization in quantum neural networks, arXiv:2109.11676 [quant-ph] (2021).

[62] S. Chen, W. Yu, P. Zeng, and S. T. Flammia, Robust Shadow Estimation, PRX Quantum **2**, 030348 (2021).

[63] D. Enshan Koh and S. Grewal, Classical shadows with noise, arXiv e-prints, arXiv:2011.11580 (2020).

[64] X. Mi *et al.*, Information scrambling in computationally complex quantum circuits, arXiv e-prints, arXiv:2101.08870 (2021).

[65] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, Julia: A fresh approach to numerical computing, SIAM Rev. **59**, 65 (2017).

[66] X.-Z. Luo, J.-G. Liu, P. Zhang, and L. Wang, Yao.jl: Extensible, efficient framework for quantum algorithm design, Quantum **4**, 341 (2020).

[67] S. Aaronson, Shadow tomography of quantum states, arXiv e-prints, arXiv:1711.01053 (2017).

[68] S. Aaronson and G. N. Rothblum, Gentle measurement of quantum states and differential privacy, arXiv e-prints, arXiv:1904.08747 (2019).

[69] M. Paini and A. Kalev, An approximate description of quantum states, arXiv e-prints, arXiv:1910.10543 (2019).

[70] J. Morris and B. Dakić, Selective quantum state tomography, arXiv e-prints, arXiv:1909.05880 (2019).

[71] A. Elben, R. Kueng, H.-Y. R. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, and B. Vermersch, Mixed-State Entanglement from Local Randomized Measurements, Phys. Rev. Lett. **125**, 200501 (2020).

[72] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Provably efficient machine learning for quantum many-body problems, arXiv e-prints, arXiv:2106.12627 (2021).

[73] T. J. Evans, R. Harper, and S. T. Flammia, Scalable Bayesian Hamiltonian learning, arXiv e-prints, arXiv:1912.07636 (2019).

[74] H.-Y. Huang, R. Kueng, and J. Preskill, Efficient Estimation of Pauli Observables by Derandomization, Phys. Rev. Lett. **127**, 030503 (2021).

[75] A. Acharya, S. Saha, and A. M. Sengupta, Informationally complete POVM-based shadow tomography, arXiv e-prints, arXiv:2105.05992 (2021).

[76] A. Neven, J. Carrasco, V. Vitale, C. Kokail, A. Elben, M. Dalmonte, P. Calabrese, P. Zoller, B. Vermersch, R. Kueng, and B. Kraus, Symmetry-resolved entanglement detection using partial transpose moments, Npj Quantum Inf. **7**, 152 (2021).

[77] A. Elben, R. Kueng, H.-Y. R. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, and B. Vermersch, Mixed-State Entanglement from Local Randomized Measurements, Phys. Rev. Lett. **125**, 200501 (2020).

[78] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Applied and Numerical Harmonic Analysis (Birkhäuser/Springer, New York, 2013), p. xviii+625.

[79] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[80] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, J. Math. Phys. **48**, 052104 (2007).

[81] R. Cleve, D. Leung, L. Liu, and C. Wang, Near-linear constructions of exact unitary 2-designs, arXiv:1501.04592 [quant-ph] (2016).

[82] R. Kueng and D. Gross, Qubit stabilizer states are complex projective 3-designs, arXiv e-prints, arXiv:1510.02767 (2015).

[83] Z. Webb, The Clifford group forms a unitary 3-design, arXiv e-prints, arXiv:1510.02769 (2015).

[84] H. Zhu, Multiqubit clifford groups are unitary 3-designs, Phys. Rev. A **96**, 062336 (2017).

[85] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, Cambridge, 2018).

[86] D. N. Page, Average Entropy of a Subsystem, Phys. Rev. Lett. **71**, 1291 (1993).

[87] P. Hayden and J. Preskill, Black holes as mirrors: Quantum information in random subsystems, J. High Energy Phys. **2007**, 120 (2007).

[88] Y. Sekino and L. Susskind, Fast scramblers, J. High Energy Phys. **2008**, 065 (2008).

[89] A. Nahum, J. Ruhman, S. Vijay, and J. Haah, Quantum Entanglement Growth under Random Unitary Dynamics, Phys. Rev. X **7**, 031016 (2017).

[90] A. Nahum, S. Vijay, and J. Haah, Operator Spreading in Random Unitary Circuits, Phys. Rev. X **8**, 021014 (2018).

[91] P. Hosur, X.-L. Qi, D. A. Roberts, and B. Yoshida, Chaos in quantum channels, J. High Energy Phys. **2016**, 4 (2016).

[92] C. von Keyserlingk, T. Rakovszky, F. Pollmann, and S. Sondhi, Operator Hydrodynamics, OTOCs, and Entanglement Growth in Systems without Conservation Laws, Phys. Rev. X **8**, 021013 (2018).

[93] S. K. Foong and S. Kanno, Proof of Page's Conjecture on the Average Entropy of a Subsystem, Phys. Rev. Lett. **72**, 1148 (1994).

[94] S. Sachdev and J. Ye, Gapless Spin-Fluid Ground State in a Random Quantum Heisenberg Magnet, Phys. Rev. Lett. **70**, 3339 (1993).