Hamiltonian variational ansatz without barren plateaus

Chae-Yeun Park and Nathan Killoran

Xanadu, Toronto, ON, M5G 2C8, Canada

Variational quantum algorithms, which combine highly expressive parameterized quantum circuits (PQCs) and optimization techniques in machine learning, are one of the most promising applications of a near-term quantum computer. Despite their huge potential, the utility of variational quantum algorithms beyond tens of qubits is still questioned. One of the central problems is the trainability of PQCs. The cost function landscape of a randomly initialized PQC is often too flat, asking for an exponential amount of quantum resources to find a solution. This problem, dubbed *barren plateaus*, has gained lots of attention recently, but a general solution is still not available. In this paper, we solve this problem for the Hamiltonian variational ansatz (HVA), which is widely studied for solving quantum many-body problems. After showing that a circuit described by a time-evolution operator generated by a local Hamiltonian does not have exponentially small gradients, we derive parameter conditions for which the HVA is well approximated by such an operator. Based on this result, we propose an initialization scheme for the variational quantum algorithms and a parameter-constrained ansatz free from barren plateaus.

1 Introduction

Recent experimental progress in controlling quantum systems has demonstrated quantum advantages in sampling tasks [1, 2, 3], and near-term quantum computers with hundreds of noisy qubits are emerging [4]. Variational quantum algorithms (VQAs) are one of the most promising applications of these near-term quantum computers. By combining highly expressive parameterized quantum circuits (PQCs) and well-established parameter optimization techniques from machine learning (ML), VQAs are relevant for many important problems, including combinatorial optimizations [5], finding the ground state of a many-body Hamiltonian [6, 7, 8, 9], and learning probability distributions [10, 11, 12, 13] (see Ref. [14] for a recent review).

VQAs solve a problem by optimizing a cost function typically defined by the expectation value of a target-problem specific observable. However, this optimization task can be challenging since the cost function landscapes are often too flat [15, 16]. This phenomenon, dubbed *barren plateaus*, is characterized by the fact that all gradient components are exponentially small with the number of qubits when parameters are randomly sampled. Given that barren plateaus are expected to be prevalent for sufficiently expressive ansätze [17], *trainability* of PQCs beyond tens of qubits is still an open question.

The issue of vanishing gradients is not entirely new, though. Classical neural networks also suffered a similar vanishing gradient problem, but theoretical and numerical advances have shown that clever neural network architectures [18, 19] or better initialization methods [20, 21] can sufficiently suppress the problem. Likewise, recent studies explored quantum circuit ansätze without barren plateaus [22, 23, 24, 25], as well as initialization techniques that provide large gradients [26, 27, 28, 29, 30]. Still, it is unclear how useful barren-plateau-free ansätze are for solving complex problems. Also, proposed initialization methods mostly rely on heuristics and do not provide strong arguments for why such parameters should yield a large gradient.

In this paper, we resolve these issues for the Hamiltonian variational ansatz (HVA) by proposing a novel parameter initialization technique. The HVA [7, 9] is widely studied for solving the ground state of a many-body Hamiltonian since it can encode adiabatic evolution. However, the HVA is still subject to the barren plateau problem [31, 32]. Even though several

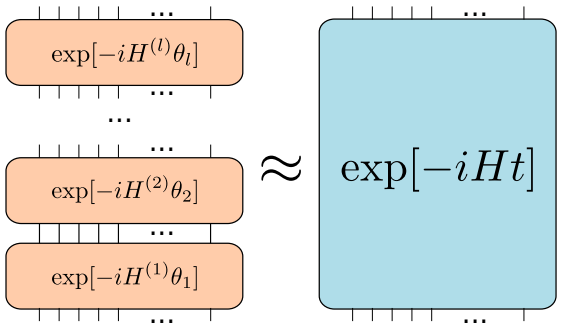

Figure 1: We find a parameter constraint such that layers of Hamiltonian evolution in the HVA (left) approximate to the time evolution under a single local Hamiltonian (right). Using the dynamical properties of local Hamiltonians, we argue that the HVA has large gradients.

initialization methods based on pre-training have been proposed to overcome this problem [29, 30], those methods not only require additional classical or quantum resources but rely on heuristics developed based on numerical results for less than 20 qubits. In contrast, our initialization scheme simply adds a constraint to the parameters and is free from additional computational resources. Moreover, we provide a rigorous argument for why this scheme yields large gradients, supported by extensive numerical results up to 28 qubits. We further propose an ansatz that imposes the constraint throughout the optimization process. Such an ansatz is expressive enough for variational time evolution [33, 34, 35, 36, 37], with the benefit that the ansatz is free from barren plateaus.

The remainder of the paper is organized as follows. After briefly introducing the problem and related concepts in Sec. 2, we show that the gradient does not decay exponentially when a circuit is described by local Hamiltonian evolution in Sec. 3. In Sec. 4, we find a parameter condition for which the HVA approximates to local Hamiltonian dynamics. We thus prove that a parameter regime for constant gradient magnitudes exists. We then introduce an initialization method based on our proof and numerically compare it to other known parameter initialization techniques in Sec. 5. We summarize our results with concluding remarks in Sec. 6.

2 Preliminaries

We consider a PQC for N qubits with l total layers, given by

$$U(\boldsymbol{\theta}) = U_l(\theta_l)\cdots U_1(\theta_1), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_l)$ is a vector of all parameters and $U_n(\theta) = e^{-i\theta G_n}$ is a unitary gate generated by G_n. In VQAs, a cost function is typically given by

$$C(\boldsymbol{\theta}) = \mathrm{Tr}[OU(\boldsymbol{\theta})\rho_0 U^\dagger(\boldsymbol{\theta})], \tag{2}$$

where O is a Hermitian operator. The cost function is then optimized with gradient-based methods. Direct computation of the gradient yields

$$\partial_n C = \frac{\partial C}{\partial \theta_n} = i\,\mathrm{Tr}[U_R\rho_0 U_R^\dagger[G_n, U_L^\dagger O U_L]], \tag{3}$$

where $U_R = U_{n-1}\cdots U_1$, $U_L = U_l\cdots U_n$, and ρ_0 is the initial state of the circuit.

For classes of PQCs, which form a 1-design, the gradient is unbiased for a given parameter set (i.e., $\mathbb{E}_{\boldsymbol{\theta}}[\partial_n C] = 0$). In this case, one can use the variance to quantify the magnitudes of gradients, which is given by

$$\begin{aligned}\mathrm{Var}[\partial_n C] &= \int d\mu(\boldsymbol{\theta}) \left(\frac{\partial C}{\partial \theta_n}\right)^2 \\ &= -\int d\mu(\boldsymbol{\theta})\,\mathrm{Tr}[U_R\rho_0 U_R^\dagger[G_n, U_L^\dagger O U_L]]^2.\end{aligned} \tag{4}$$

In the typical barren plateau scenario [15, 16], this quantity becomes close to $\mathcal{O}(1/D^2)$ [1], which is the value evaluated under the assumption that U_R or U_L is a unitary 2-design. Here, D is the total dimension of the Hilbert space, which is 2^N for a system with N qubits. Hence, the variance decays exponentially with the number of qubits, which implies that the gradient is exponentially small for most values of the parameters (can be rigorously proven by Chebyshev's inequality). Even though it is possible to optimize the cost function using a small gradient in principle, running the algorithm in real quantum hardware is extremely inefficient as estimating the gradient requires an exponential number of shots.

[1] See Appendix A for the definition of big-O and related notations

Next, we introduce the HVA. The HVA [7, 9] is a natural ansatz for solving quantum many-body Hamiltonians. After decomposing a given Hamiltonian into q terms $H = \sum_{j=1}^{q} c_j H^{(j)}$, where $\{c_j\}$ are real coefficients, the HVA is constructed as

$$|\psi(\{\theta_{i,j}\})\rangle = \prod_{i=p}^{1} [e^{-iH^{(q)}\theta_{i,q}} \cdots e^{-iH^{(2)}\theta_{i,2}} e^{-iH^{(1)}\theta_{i,1}}] |\psi_0\rangle , \tag{5}$$

where $|\psi_0\rangle$ is a quantum state that can be easily prepared. The ansatz consists of p blocks, each containing q layers. Thus, the ansatz has a total of $l = pq$ layers. We also use the notation θ_a and U_a to denote $\theta_{i,j}$ and $U_a = e^{-iH^{(j)}\theta_{i,j}}$ where $a = (i, j)$, which enables us to interpret the HVA as a PQC given by Eq. (3).

Throughout the paper, we restrict $H^{(j)}$ to be a k-local Hamiltonian in a given lattice for a constant k, i.e., each term in the Hamiltonian acts on at most k *geometrically nearby sites* in a given lattice. This condition is satisfied for most of the many-particle spin-1/2 Hamiltonians. For example, we decompose the one-dimensional transverse-field Ising model $\mathcal{H} = -\sum_i Z_i Z_{i+1} + hX_i$ into $H^{(1)} = -\sum_i Z_i Z_{i+1}$ and $H^{(2)} = -\sum_i X_i$. Then $H^{(1)}$ is 2-local and $H^{(2)}$ is 1-local.

This ansatz is powerful for solving the ground state of H, as it can encode the adiabatic evolution of the Hamiltonian [9]. Despite the usefulness of the ansatz, however, training the HVA turned out to be non-trivial. Some numerical studies have observed that the gradients decay exponentially with the system size [31, 32], although the magnitudes of gradients of the HVA are larger than one expects from a unitary 2-design [31].

In this paper, we consider the case where the HVA is well approximated by time evolution under a local Hamiltonian, i.e., there are local Hamiltonians H_L, H_R such that $U_L \approx e^{-iH_L t_L}$ and $U_R \approx e^{-iH_R t_R}$ for some $t_L, t_R \geq 0$ (see Fig. 1). With this assumption, we provide strong analytic and numerical arguments that the gradient magnitudes, Eq. (4), only decay at most polynomially in the number of qubits. Although this assumption seems unrealistic, we later show that the HVA with a certain parameter restriction can satisfy this condition.

3 Magnitudes of gradients in Hamiltonian dynamics

In this section, we study the scaling of the gradient when the circuit is given by the time evolution under a time-independent local Hamiltonian. We show that gradients in such circuits do not decay exponentially in both extreme regimes: short- and long-time evolution.

For short-time evolution, we prove a rigorous bound on the time that the gradient preserves its initial magnitudes. Thus, a circuit have large gradients for a proper initial state. On the other hand, for long-time evolution, we combine the universality of quantum thermalization [38, 39, 40] and our numerical results to argue that the gradient does not decay exponentially.

3.1 Gradient scaling for short-time evolution

In this subsection, assuming that (1) a circuit with N qubits is given by e^{-iHt} for a local Hamiltonian H and (2) the initial state has a large gradient with a value of $\Theta(1)$, we prove that there exists $t_c = \Theta(1/N)$ such that the circuit maintains the large gradient when the total evolution time is less than t_c. Our main result is the following proposition:

Proposition 1 (Quantum speed limit of gradients)**.** *For the HVA, the gradient of the cost function is given by*

$$\partial_{n,m} C = \frac{\partial C}{\partial \theta_{n,m}} = i \operatorname{Tr}[U_R \rho_0 U_R^\dagger [H^{(m)}, U_L^\dagger O U_L]] \tag{6}$$

where $U_R = e^{-iH^{(m-1)}\theta_{n,m-1}} \cdots e^{-iH^{(1)}\theta_{1,1}}$ *and* $U_L = e^{-iH^{(q)}\theta_{p,q}} \cdots e^{-iH^{(m)}\theta_{n,m}}$*. Assume that the gradient component,* $\partial_{n,m} C$*, is non-zero when the circuit is identity, i.e.,* $|\operatorname{Tr}[\rho_0 [H^{(m)}, O]]| > 0$*, and there are local Hamiltonians* H_L*,* H_R *such that* $U_L = e^{-iH_L t_L}$ *and* $U_R = e^{-iH_R t_R}$ *for some* $t_R, t_L \geq 0$*. Then,*

$$\left|\frac{\partial C}{\partial \theta_{n,m}}\right| \geq |\operatorname{Tr}[\rho_0 [H^{(m)}, O]]|/2 \tag{7}$$

for $t_R + t_L \leq t_c := |\operatorname{Tr}[\rho_0 [H^{(m)}, O]]|/(4KC)$*, where* $K = \max\{\|H_R\|, \|H^{(m)}\|\}$*,* $C = \max\{\|[H^{(m)}, O]\|, \|[H_L, O]\|\}$*, and* $\|\cdot\|$ *is the operator norm.*

Proof. Let

$$A(t_1,t_2) = i\,\mathrm{Tr}[e^{-iH_Rt_1}\rho_0 e^{iH_Rt_1}[H^{(m)}, e^{iH_Lt_2}Oe^{-iH_Lt_2}]]. \tag{8}$$

Then

$$|A(t_R,t_L)-A(0,0)| \le \int_0^{t_R} dt_1 \Big|\frac{\partial A(t_1,0)}{\partial t_1}\Big| + \int_0^{t_L} dt_2 \Big|\frac{\partial A(t_R,t_2)}{\partial t_2}\Big|. \tag{9}$$

We further have

$$\Big|\frac{dA(t_1,0)}{\partial t_1}\Big| = \Big|\mathrm{Tr}\{[H_R,\rho_0(t_1)][H^{(m)},O]\}\Big| \le 2\|H_R\|\|[H^{(m)},O]\| \le 2KC, \tag{10}$$

and

$$\Big|\frac{dA(t_R,t_2)}{\partial t_2}\Big| = \Big|\mathrm{Tr}\{\rho_0(t_R)[H^{(m)},[H_L,e^{iH_Lt}Oe^{-iH_Lt}]]\}\Big| \le 2\|H^{(m)}\|\|[H_L,O]\| \le 2KC, \tag{11}$$

where $\rho_0(t) = e^{-iH_Rt}\rho_0 e^{iH_Rt}$.

Integrating both sides, we have

$$|A(t_R,t_L)-A(0,0)| \le 2KC(t_R+t_L). \tag{12}$$

By entering $t_R+t_L \le t_c = |A(0,0)|/(4KC)$, we obtain $|A(t_R,t_L)-A(0,0)| \le |A(0,0)|/2$, i.e.,

$$A(0,0)-|A(0,0)|/2 \le A(t_R,t_L) \le A(0,0)+|A(0,0)|/2. \tag{13}$$

We obtain the desired inequality as $A(t_R,t_L) \ge A(0,0)/2 > 0$, if $A(0,0) > 0$, and $A(t_R,t_L) \le A(0,0)/2 < 0$, otherwise. □

Let us assume that all of $H^{(m)}$, H_L, and H_R are k-local Hamiltonians, where each term acts at most k nearby sites in a given lattice for a constant k, and O is a local operator acting on at most a constant number of sites. Under this assumption, which is the case we consider in this paper, we have $t_c = \Theta(1/N)$ when $|\mathrm{Tr}[\rho_0[H^{(m)},O]]| = \Theta(1)$. We prove this fact in the rest of the subsection.

For any k-local Hamiltonian H, we can write

$$H = \sum_{i\in\Lambda} h_i \tag{14}$$

where $\Lambda = \{1,\cdots,N\}$ is the collection of all sites, and h_i is an operator supported by k sites centered at i. Formally, we write the support of h_i (a set of sites h_i acts on) as

$$\mathrm{supp}(h_i) = \{j\in\Lambda : \mathrm{dist}(i,j) \le k\} \tag{15}$$

where $\mathrm{dist}(i,j)$ is the distance between two sites in the given lattice. We thus have

$$\|H\| \le N\max_i \|h_i\|, \quad \|[H,O]\| \le 2s\|O\|\max_i\|h_i\| \tag{16}$$

for a local operator O. Here,

$$s = |\{i\in\Lambda : \mathrm{dist}(i,O) \le k\}| \tag{17}$$

is a constant for a given lattice, where $\mathrm{dist}(i,O) = \min_{j\in\mathrm{supp}(O)}\mathrm{dist}(i,j)$ and $\mathrm{supp}(O) \subset \Lambda$ is the support of O. Given that $\|h_i\|$, $\|O\|$ are bounded by a constant for a spin system, and s is a constant for a finite-dimensional lattice, we have

$$\|H\| = \mathcal{O}(N), \quad \|[H,O]\| = \mathcal{O}(1) \tag{18}$$

for any k-local Hamiltonain H. We also note that physical Hamiltonians must have $\|H\| = \Theta(N)$, which is a necessary condition to be thermodynamically well-defined. Therefore, we obtain $t_c = \Theta(1/N)$ if the circuit has a large initial gradient component, i.e., if there exists m such that $|\mathrm{Tr}[\rho_0[H^{(m)},O]]| = \Theta(1)$.

For example, when $H^{(m)} = \sum_i Y_i$, $O = Z_1$, we have $i\,\mathrm{Tr}[\rho_0[H^{(m)},O]] = -2$ for $|\psi_0\rangle = |+\rangle^{\otimes N}$. Moreover, we have $K = \Theta(N)$ and $C = \Theta(1)$ for Proposition 1, when H_R and H_L are also k-local, which implies $t_c = \Theta(1/N)$. Therefore, the gradient component is $\Theta(1)$ for all $t \le t_c$.

While we mostly consider geometrically local Hamiltonians in this paper (i.e., local in a finite-dimensional lattice), our result can be extended to Hamiltonians defined on a general (hyper)graph. Such a complex Hamiltonian appears when a fermionic Hamiltonian is translated to a spin Hamiltonian using, e.g., the Jordan–Wigner transformation (see, e.g., Ref. [7]). These Hamiltonians can have smaller t_c because $C = \max\{\|[H^{(m)},O]\|, \|[H_L,O]\|\}$ in Proposition 1 may scale linearly with N. For example, consider a Hamiltonian H, each term of which acts on all sites. Namely, we have H given as

$$H = \sum_{i=1}^{N} h_i, \tag{19}$$

where each h_i is a Pauli string acting non-trivially on all sites, i.e., $h_i \in \{X, Y, Z\}^{\otimes N}$. We still have $\|H\| = \Theta(N)$ for this Hamiltonian. However, for a local operator, O, acting on site i, we can have $\|[O, H]\| = \Theta(N)$. Thus, applying Proposition 1 to this Hamiltonian yields $t_c = \Theta(1/N^2)$ instead of $\Theta(1/N)$.

3.2 Gradient scaling for long-time evolution

Next, we consider long-time evolution. Following usual arguments for equilibration [41, 40, 42, 43, 44], we assume that a Hamiltonian H follows the non-degenerate energy-gap condition: $E_i - E_j = E_k - E_l$ iff $i = k$ and $j = l$; or $i = j$ and $k = l$, where E_i is the i-th eigenvalue of H with the corresponding eigenvector $|E_i\rangle$. With an additional assumption that the Hamiltonian thermalizes [38, 39, 40], in the sense that the observable after equilibration gives a similar value to the thermal average, it is known that the second moment of the Hamiltonian evolution behaves differently from a unitary 2-design [45].

Precisely, Huang et al. [45] considered the saturated value of the out-of-time correlator (OTOC), a widely used measure for detecting quantum chaos. For local Hermitian operators, O_i and O_j acting on sites i and j, respectively, the OTOC is defined by

$$\mathrm{OTOC}(O_i, O_j) := \mathrm{Tr}\left[\rho_0 (U O_i U^\dagger O_j)^2\right]. \tag{20}$$

Here, ρ_0 is the initial state, and U is a unitary operator determining the time evolution of the system.

One often considers the infinite temperature initial state given by $\rho_0 = \mathbb{1}/2^N$ for a system with N qubits. When our unitary operator U forms a 2-design, we obtain

$$\mathrm{OTOC}(O_i, O_j) = \frac{1}{2^N} \mathrm{Tr}\left[U O_i U^\dagger O_j U O_i U^\dagger O_j\right] \tag{21}$$

$$\xrightarrow{\text{Haar } U} -\frac{2^N}{2^{2N} - 1}, \tag{22}$$

for traceless O_i and O_j (e.g., local Pauli operators; see, e.g., Ref. [46] for a proof). Namely, $\mathrm{OTOC}(O_i, O_j)$ scales inverse *exponentially* with N in this case. On the other hand, $\mathrm{OTOC}(O_i, O_j)$ scales only inverse *polynomially* with the system size for local Hamiltonian evolution, i.e., when $U = e^{-iHt}$ for a local Hamiltonian

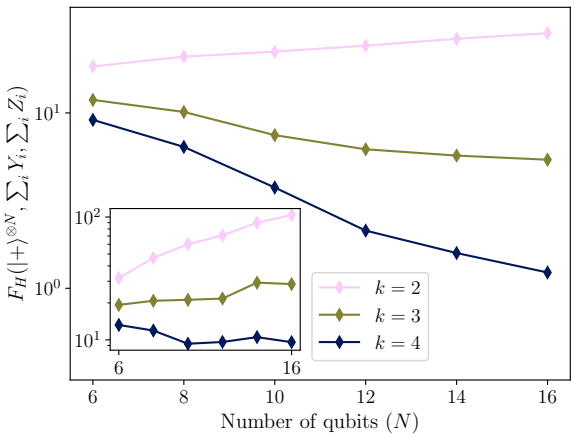

Figure 2: The lower bound $F_H(|\psi\rangle, H^{(1)}, O)$ for $|\psi\rangle = |+\rangle^{\otimes N}$, $H^{(1)} = \sum_i Y_i$, and $O = \sum_i Z_i$ averaged over 2^{10} randomly generated k-local Hamiltonians. Inset: The same results but for Hamiltonians with time-reversal symmetry.

H [45]. Such a huge difference mainly comes from the fact that $U = e^{-iHt}$ conserves the energy, i.e., $[H, U] = 0$.

Similarly, one might expect that the variance of gradients saturates to a value that scales only inverse polynomially with N for local Hamiltonian evolution. To see that this is the case, we compute the square of the first element of the gradient (for $\theta_{1,1}$) when $U_L = e^{-iH_L t}$ and the initial state is given as a pure state $\rho_0 = |\psi_0\rangle\langle\psi_0|$. Then we obtain a lower bound of $(\partial_{\theta_{1,1}} C)^2 = -\langle\psi_0|[H^{(1)}, O(t)]|\psi_0\rangle^2$ where $O(t) = e^{iH_L t} O e^{-iH_L t}$ as follows:

Proposition 2. *Assume that H_L satisfies the non-degenerate energy-gap condition. Then the long-time average of $-\langle\psi_0|[H^{(1)}, O(t)]|\psi_0\rangle^2$ is lower bounded by $F_{H_L}(|\psi_0\rangle, H^{(1)}, O)$. Here, $F_H(|\psi\rangle, G, O)$ is a function given by*

$$\begin{aligned} &F_H(|\psi\rangle, G, O) \\ &\quad = 2\sum_{ijkl} C_i^* G_{ij} O_{jk} |C_k|^2 O_{kj} G_{jl} C_l \\ &\quad - \sum_{ijkl} C_i^* O_{ij} G_{jk} C_k C_j^* O_{ji} G_{il} C_l \\ &\quad - \sum_{ijkl} C_i^* G_{ij} O_{jk} C_k C_l^* G_{lk} O_{kj} C_j, \end{aligned} \tag{23}$$

where $C_i = \langle E_i|\psi_0\rangle$, $G_{ij} = \langle E_i|G|E_j\rangle$, and $O_{jk} = \langle E_j|O|E_k\rangle$. Here, $|E_i\rangle$ is the i-th eigenstate of H.

A proof can be found in Appendix B. Unfortunately, we cannot use techniques to analytically

compute the OTOC for a maximally mixed initial state $\rho_0 = \mathbb{1}/2^N$ [45], as we here consider a pure initial state. Instead, we provide numerical evidence that F_H does not decay exponentially for translationally invariant local Hamiltonians with and without the time-reversal symmetry. We especially consider time-reversal symmetric Hamiltonians as they are an important subclass of Hamiltonians widely considered for thermalization [47].

After creating a random k-local Hamiltonian H (see Appendix C for numerical details), we diagonalize H and compute F_H using the obtained eigenstates for the initial state $|\psi\rangle = |+\rangle^{\otimes N}$, $G = \sum_i Y_i$, and $O = \sum_i Z_i$. As all observables, the Hamiltonian, and the initial state are translationally-invariant, we can compute F_H within the translationally-invariant subspace.

We plot the result in Fig. 2 up to $N = 16$ for random Hamiltonians without (main figure) and with (inset) the time-reversal symmetry. For $k = 2$, we observe that the lower bound F_H does not decay at all in both cases. For $k = 3, 4$, F_H decreases with N for the Hamiltonians without the time-reversal symmetry. Even though it is not conclusive to tell the exact decaying rate of F_H from this plot, we strongly believe that it is not exponential from the universality of thermalization dynamics for non-integrable models [38, 39, 40], i.e., if F_H for Hamiltonians with the time-reversal symmetry does not decay exponentially, F_H for any thermalizing Hamiltonians also does not decay exponentially.

We also recall Ref. [32], which conjectured that the variance of gradients only scales inverse polynomially with the dimension of the dynamical Lie algebra $\mathcal{G}$ spanned by the gate generators, i.e., $\mathcal{G} = \langle iG_1, \cdots, iG_l \rangle_{\text{Lie}}$ where $\langle \cdot \rangle_{\text{Lie}}$ is the Lie closure containing all nested commutators of the listed elements. A reason behind using dynamical Lie algebra is that the algebra generates any arbitrary circuit that the given PQC can express, i.e., there is a $g \in \mathcal{G}$ such that $U(\boldsymbol{\theta}) = e^g$ for any $\boldsymbol{\theta}$. However, for random Hamiltonians, as in our case, it is more natural to use a vector space of the Hamiltonians (which also generates a unitary operation but not a Lie algebra) instead. As the dimension of k-local random Hamiltonians is $\Theta(N)$, the conjecture with a slightly relaxed condition would also indicate that the gradient does not decay exponentially.

We explicitly write down our version of the conjecture, which is also supported by our numerical results, as follows:

Conjecture 1. *Let V be a vector space of local Hamiltonians. Then for any given initial state $|\psi_0\rangle$, $G \in V$, and a local operator O, we have*

$$-\int_{\nu \in V} d\nu \, \langle \psi_0 | [G, e^{i\nu} O e^{-i\nu}] | \psi_0 \rangle^2 = \frac{1}{\text{poly}(N)} \tag{24}$$

where $d\nu$ is a proper measure for the vector space.

4 Approximating HVA to local Hamiltonian evolution

In the previous section, we argued that a circuit given by the time evolution under a local Hamiltonian does not have barren plateaus. In this section, we find a parameter condition for which the HVA given by Eq. (5) is well approximated by local Hamiltonian evolution. We first interpret the HVA as a unitary operator generated by a time-dependent Hamiltonian. Then, we utilize the Floquet-Magnus (FM) expansion to obtain an effective *time-independent* Hamiltonian that describes the HVA within a small error.

We here consider each Hamiltonian $H^{(j)}$ satisfying the following conditions: (C1) $H^{(j)}$ is a sum of commuting Pauli strings, (C2) $H^{(j)}$ is k-local (each term acts on at most k nearby qubits in a given lattice), and (C3) each Pauli string of $H^{(j)}$ uniquely supports a subset $X \subset \Lambda$ (e.g., $H^{(j)}$ cannot have terms X_1X_2 and Z_1Z_2 simultaneously). In other words, we consider $H^{(j)}$ defined as

$$H^{(j)} = \sum_{|X| \leq k} h_X^{(j)}, \tag{25}$$

where the summation is over all subsets of sites $X \subseteq \Lambda$ whose length is $\leq k$, and $\Lambda = \{1, \cdots, N\}$ is a set of all sites. In addition, $h_X^{(j)}$ is a single Pauli string (if there is a term whose support is X) or 0 (otherwise), and $[h_X^{(j)}, h_Y^{(j)}] = 0$ for all $X, Y \subseteq \Lambda$. As we assume that $H^{(j)}$ is k-local, $h_X^{(j)} = 0$ if X contains any non-nearby sites (i.e., if there are $a, b \in X$ such that the distance between a and b is larger than k).

We also define parameters for the FM expan-

sion. First, we define

$$H_{\max} = \max_j \sum_{|X|\le k} \|h_X^{(j)}\| = \max_j |\{X \subseteq \Lambda : h_X^{(j)} \neq 0\}|, \quad (26)$$

where we obtained the last equality using the fact that $\{h_X^{(j)}\}$ are commuting Pauli strings. Thus, $H_{\max}$ is the maximum number of terms in $H^{(j)}$. We also introduce a parameter J that upper bounds the local interaction strength

$$\max_j \sum_{X:X\ni a} \|h_X^{(j)}\| \le J, \qquad \forall a \in \Lambda. \quad (27)$$

As $\{h_X^{(m)}\}$ are Pauli strings, we can use

$$J = \max_{a\in\Lambda} \max_j |\{X : X \ni a \text{ and } h_X^{(j)} \neq 0\}|. \quad (28)$$

From the locality of Hamiltonians $\{H^{(j)}\}$, we have $H_{\max} = \Theta(N)$ (see discussion below Proposition 1), and J is upper bounded by the number of vertices whose L^1 distance to the origin is $\le k$, which is a constant for a finite-dimensional lattice.

We further assume that all parameters $\{\theta_{i,j}\}$ in the HVA are larger than 0. Under this setup, the following Proposition shows that the subcircuits U_R and U_L of the HVA can be approximated by the time evolution under a few-body Hamiltonian when the sum of parameters is small.

Proposition 3. *For the HVA composed of $H^{(j)}$, given in Eq. 5, we consider a subcircuit $U_R = e^{-iH^{(j-1)}\theta_{i,j-1}} \cdots e^{-iH^{(1)}\theta_{1,1}}$. We additionally assume that all $H^{(j)}$ satisfy the conditions (C1-C3) defined above. Then, there is a Hamiltonian $H_R^{(n)}$ acting on at most $(n+1)k$-local sites such that*

$$\left\| U_R - e^{-iH_R^{(n)} t_R} \right\| \le 6H_{\max} 2^{-n_0} t_R + \frac{2H_{\max}(2kJ)^{n+1}}{(n+2)^2}(n+1)! t_R^{n+2} \quad (29)$$

with $t_R = \theta_{1,1} + \cdots + \theta_{i,j-1}$ for all $n \le n_0 = \lfloor 1/(32kJt_R) \rfloor$. Likewise, there is a Hamiltonian $H_L^{(n)}$ that approximates $U_L = e^{-iH^{(q)}\theta_{p,q}} \cdots e^{-iH^{(j)}\theta_{i,j}}$ with the same error but for $t_L = \theta_{i,j} + \cdots + \theta_{p,q}$.

We derive the bound and properties of $H_{R,L}^{(n)}$ in Appendix D. Our derivation is based on the truncated FM expansion rigorously proven in Ref. [48].

The above bound tells us that $U_{R,L}$ can be approximated by local Hamiltonian evolution when $t_{R,L}$ are small. For example, when $t_R = \mathcal{O}(1/N)$ and for a constant n, the first term in the bound is exponentially small in N and the second term is $\mathcal{O}(1/N^{n+1})$. Then, one may further employ Proposition 1 to get a large gradient, which we summarize as the following theorem:

Theorem 1. *For the HVA [Eq. (5)] and for a local observable O acting on at most $\mathcal{O}(1)$ sites, assume that there exists an initial state $\rho_0 = |\psi_0\rangle\langle\psi_0|$ which gives $g := |\operatorname{Tr}[\rho_0[H^{(m)}, O]]| = \Theta(1)$ regardless of N. Then, there is $\tau_0 = \Theta(1/N)$ such that*

$$\left|\frac{\partial C}{\partial \theta_{n,m}}\right| \ge \frac{g}{4} \quad (30)$$

for all n, if $\sum_{ij}\theta_{i,j} = t_L + t_R \le \tau_0$.

The proof can be found in Appendix E.

We provide two remarks on Theorem 1. First, if there exists any constraint with $\tilde{\tau}_0$ such that a gradient component is bounded below by a constant for all $\sum_{ij}\theta_{ij} \le \tilde{\tau}_0$, then $\tilde{\tau}_0 \le \pi/4$. This is because one can easily find the HVA with suitable ρ_0 and O which satisfies $|\operatorname{Tr}[\rho_0[H^{(m)}, O]]| = \Theta(1)$, but $\partial_{n,m}C = i\operatorname{Tr}[U(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger [H^{(m)}, O]] = 0$ for $\sum_{i,j}\theta_{i,j} = \pi/4$. We provide such an example in Appendix F. This implies that Theorem 1 can be improved at most $\tau_0 = \Theta(1)$ in our current set-up.

Second, the theorem implies that for *any* probability distribution $p(\boldsymbol{\theta})$ defined for $\theta_{i,j} \ge 0$ and $\sum_{i,j}\theta_{i,j} \le \tau_0$,

$$\int d\boldsymbol{\theta} p(\boldsymbol{\theta}) \Big(\frac{\partial C}{\partial \theta_{n,m}}\Big)^2 = \Theta(1), \quad (31)$$

which is much stronger than the non-exponential decay of gradients. If one considers a different condition, e.g., a polynomially decaying gradients under uniformly generated initial parameters, a parameter constraint may be further relaxed. Namely, we open up the possibility that a constraint with $\tau_1 = \Omega(1)$ exists such that

$$\int_{\theta_{i,j}\ge 0, \sum_{i,j}\theta_{i,j}\le\tau_1} \prod_{i,j} d\theta_{i,j} \Big(\frac{\partial C}{\partial \theta_{n,m}}\Big)^2 = \frac{1}{\operatorname{poly}(N)} \quad (32)$$

is satisfied for all $\theta_{i,j} \ge 0$ and $\sum_{i,j}\theta_{i,j} \le \tau_1$. We numerically investigate related scenarios in the following section.

5 Numerical comparison between initialization methods

Theorem 1 tells us that there is $\tau_0 = \Theta(1/N)$ such that the HVA does not have barren plateaus when the sum of all parameters is less than τ_0. Still, the exact value of τ_0 for the Theorem is difficult to obtain or can be unrealistically small. Thus, in this subsection, we introduce an initialization method based on Theorem 1 and compare it to the small constant initialization considered in Refs. [49, 50, 51, 52].

We numerically test the following three different initialization methods. (1) **Random**: complete uniformly random initialization, such that all parameters are from $\mathcal{U}_{[0,2\pi]}$, (2) **Constrained**: the sum of parameters in each layer is constrained to be $T = c/N$ with a constant c, i.e., $\sum_j \theta_{i,j} = T$ for all i, and (3) **Small**: $\theta_{i,j} \sim \mathcal{U}_{[0,\epsilon]}$ for a small ϵ independent to N [49, 50, 51]. For method (2), we show that a relatively large value of $c = \pi/2$ already gives $\Theta(1)$ gradient magnitudes. On the other hand, we observe two different scaling behaviors for the constant small initialization [method (3)]. There is a value N_0 depending on p and ϵ such that the gradient magnitudes decay exponentially for $N < N_0$ whereas they only decay polynomially for $N > N_0$. This observation suggests that there can be another parameter regime in the HVA that does not have barren plateaus.

We use the HVA for the one-dimensional (1D) and two-dimensional (2D) spin-1/2 Heisenberg-XYZ models $\mathcal{H} = \sum_{\langle a,b\rangle} J_x X_a X_b + J_y Y_a Y_b + J_z Z_a Z_b$ to test these methods, which is given by

$$|\psi(\boldsymbol{\theta})\rangle = \prod_{i=p}^{1} e^{-i\theta_{i,3}\sum_{\langle a,b\rangle} Z_a Z_b} e^{-i\theta_{i,2}\sum_{\langle a,b\rangle} Y_a Y_b} \times e^{-i\theta_{i,1}\sum_{\langle a,b\rangle} X_a X_b} |\psi_0\rangle \tag{33}$$

where $\langle a, b\rangle$ are two nearest neighbors and $|\psi_0\rangle$ is the Néel state in the given lattice. We use N spins in the periodic boundary condition (a ring) for the one-dimensional model. For the two-dimensional model, we consider a rectangular lattice with the periodic boundary condition (a torus) size of $L_x \times L_y$. Thus the Néel state in these lattices are given by $|\psi_0\rangle = (|{\downarrow\uparrow}\rangle^{\otimes N/2} + |{\uparrow\downarrow}\rangle^{\otimes N/2})/\sqrt{2}$ and $[(|{\downarrow\uparrow}\rangle^{\otimes L_x/2} |{\uparrow\downarrow}\rangle^{\otimes L_x/2})^{\otimes L_y/2} + (|{\uparrow\downarrow}\rangle^{\otimes L_x/2} |{\downarrow\uparrow}\rangle^{\otimes L_x/2})^{\otimes L_y/2}]/\sqrt{2}$ for the 1D and 2D models, respectively.

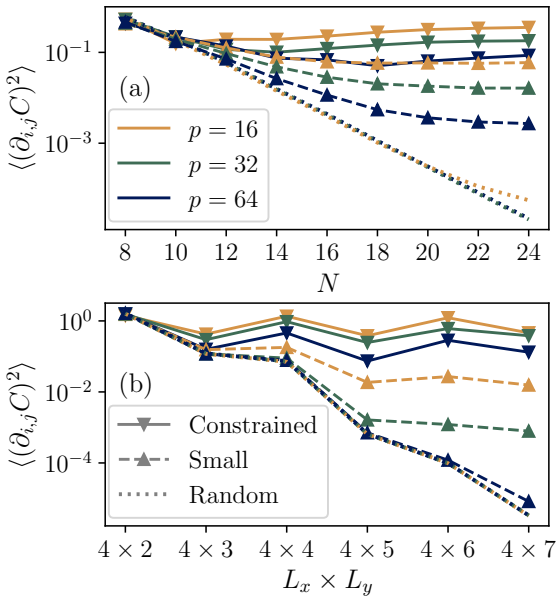

Figure 3: (a) Scaling of the gradient square $(\partial_{i,j}C)^2$ from the HVA for the 1D Heisenberg-XYZ model with different depths p (see main text for details). Plots show results from the constrained (solid), small (dashed), and completely random (dotted) parameter initializations. We compute $(\partial_{i,j}C)^2$ for 2^{10} parameters samples from each distribution and plot the averaged results over all samples and gradient components (i, j). (b) The same plot as (a) but for the 2D Heisenberg-XYZ model with the lattice size $L_x \times L_y$. We also see that the results fluctuate for odd and even L_y because the 2D Néel state (our initial state) violates the symmetry of the Hamiltonian for odd L_y. We also compute the relative standard deviation, $r = \sigma(X)/\mathbb{E}[X]$, where $X := \sum_{ij}(\partial_{i,j}C)^2/(3p)$ is the squared partial derivatives averaged over all parameters. Here, the standard deviation and the expectation value are taken over the circuit instances. The values of r for the 1D model with $p = 64$ and $N = 24$ are given by 0.20, 0.53, and 0.13 for the contained, small, and random initialization, respectively. For the 2D model with $p = 64$ and $L_x \times L_y = 4 \times 7$, we obtained $r \approx 0.39$ (constrained), 1.02 (small), 0.11 (random), respectively.

5.1 Scaling of gradients

We compute gradients of the cost function with $O = Y_0 Y_1$ (thus $C = \langle\psi(\boldsymbol{\theta})|Y_0 Y_1|\psi(\boldsymbol{\theta})\rangle$) obtained from different initialization methods. For constrained initialization, random values $\tilde{\theta}_{i,j}$ are first sampled from the uniform distribution, i.e., $\tilde{\theta}_{i,j} \sim \mathcal{U}_{[0,2\pi]}$, and then parameters are assigned by normalizing them: $\theta_{i,j} = \tilde{\theta}_{i,j} \times T/(\sum_{j=1}^{3} \tilde{\theta}_{i,j})$ for all i, j. This method ensures that $\sum_{j=1}^{3} \theta_{i,j} = T$. We here use $T = \pi/(2N)$. The results are compared to small parameters $\theta_{i,j} \sim \mathcal{U}_{[0,\epsilon]}$ with $\epsilon = 0.2$ as well as complete random parameters $\theta_{i,j} \sim \mathcal{U}_{[0,2\pi]}$. For each set of system parameters (size and depth p) and the initialization method, we compute all gradient components and plot the averaged squared magnitudes (i.e., we averaged $\langle(\partial_{i,j}C)^2\rangle = \sum_{i,j}(\partial_{i,j}C)^2/(3p)$ over 2^{10} random circuit instances).

The results for the 1D and 2D models are shown in Fig. 3(a) and (b), respectively. The 1D model clearly shows that the magnitudes of gradients do not decay with N, i.e., $(\partial_{i,j}C)^2 \approx \Theta(1)$, for the constrained initialization, which is consistent with Theorem 1. On the other hand, the gradient magnitudes decay exponentially with N when the complete random initialization is used with $p \in [16, 32]$. However, we could observe an interesting behavior when parameters are initialized to be small ($\theta_{i,j} \sim \mathcal{U}_{[0,\epsilon]}$). In this case, the gradient decays exponentially up to some N_0, i.e., for $N \leq N_0 \approx 18$, but it decays slower after that. One can already see this signature even for the complete random initialization when $p = 16$ and $N = 24$, where the averaged gradient magnitudes are larger than that from $p \in [32, 64]$. We also see similar behaviors for the 2D model, besides the results from each initialization oscillate for odd and even L_y, since our initial state (2D Néel state) is not fully symmetric for odd L_y.

As non-exponential decaying gradients from the small constant initialization have not been clearly reported in previous studies, we explore these phenomena more closely here. We compute the averaged squared gradients using the HVA for the 1D XYZ model with $p = 16$ when the circuit parameters are sampled from $\mathcal{U}_{[0,\epsilon]}$ for different values of N and ϵ. We plot the result as a function of ϵ in Fig. 4. The results show that the magnitudes of gradients saturate to an exponentially small value as ϵ increases, but the point it saturates, $\epsilon_0(N)$, also increases with N.

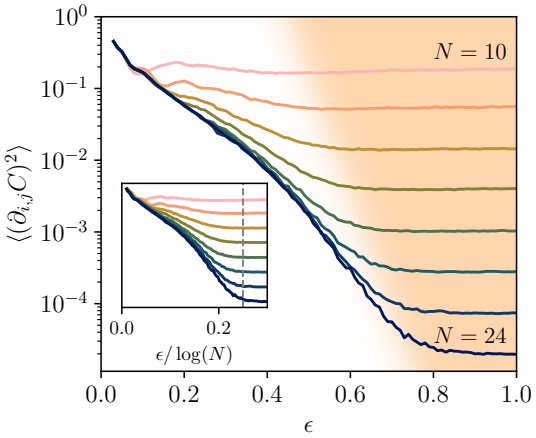

Figure 4: Averaged magnitudes of gradients $\langle(\partial_{i,j}C)^2\rangle$ as a function of ϵ. The HVA for the 1D XYZ model with $p = 16$ is used. All parameters are samples from the uniform distribution, i.e., $\theta_{i,j} \sim \mathcal{U}_{[0,\epsilon]}$. We observe that there is a value $\epsilon_0(N)$ such that the gradient decays exponentially with N only for $\epsilon > \epsilon_0(N)$ (colored region). Inset shows the same data but as a function of $\epsilon/\log(N)$. We see the gradient magnitudes saturate after the dashed vertical line, which suggests that $\epsilon_0(N) \propto \log N$.

For $\epsilon < \epsilon_0(N)$, we observe that the gradient does not decay exponentially. This fact also confirms that there can be another parameter regime beyond the one we mainly considered in this paper, which is also free from barren plateaus.

To see how $\epsilon_0(N)$ scales with N, we plot the same data but as a function of $\epsilon/\log(N)$ (inset). The plot shows that the gradient magnitudes saturate when $\epsilon/\log(N)$ is larger than a constant, which suggests that $\epsilon_0(N) \propto \log N$. In general, we expect that there is a relation between $\Upsilon := \sum_{i,j} \theta_{i,j}$ (which is $\propto p\epsilon$ in this case) in the HVA for a 2-local Hamiltonian and a random local circuit with depth $\propto \Upsilon$. Such a connection explains the observed behavior as a 1D random local circuit requires its depth larger than $\Theta(\log(N))$ to show exponential decay of gradient magnitudes when the cost function is given by the expectation value of a local observable [16].

We still note that it is less clear whether a small constant parameter initialization [method (3)] gives the same quantitative behavior for the HVAs with more complex Hamiltonians (e.g., 1D k-local with $k \geq 3$ or defined in a high-dimensional lattice). In contrast, we expect to have $\Theta(1)$ gradient magnitudes regardless of the dimension with our initialization method [method (2)]. As a detailed investigation of the

relation between the HVA and a local random circuit is out of the scope of the current work, we leave it to future work.

5.2 Full simulation of variation quantum eigensolver

We now explore whether our initialization improves learning procedures by fully simulating a variational quantum eigensolver (VQE) using the Heisenberg model ($J_x = J_y = J_z = 1$). We define the Hamiltonian expectation values as the cost function ($C = \langle\psi(\boldsymbol{\theta})|\mathcal{H}|\psi(\boldsymbol{\theta})\rangle$) and train the circuit using the Adam optimizer [53].

We first simulate the VQEs using exact gradients. Quantum hardware cannot compute exact gradients, as each gradient component should be estimated from the measurement outcomes from shots. However, classical quantum circuit simulators support multiple algorithms to obtain exact gradients. For our simulation, we use the adjoint method [54] implemented in PennyLane [55]. We present learning curves from different parameter initialization methods for the one-dimensional lattices (with learning rate $\alpha = 0.025$ and the default values for hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$) in Fig. 5 (a), which shows that our initialization scheme outperforms other initialization schemes. The completely random parameter initialization fails to find the ground state. This is an expected behavior from the presence of the barren plateaus. On the other hand, initializing all parameters to π ($\theta_{i,j} = \pi$ for all i, j), which is used in Ref. [31], works but is subject to large initial fluctuations. Generally speaking, such an initialization without randomness is prone to local minima [56]. We also found a similar behavior for the 2D Heisenberg model, shown in Fig. 5(b).

We next compare results from different initialization methods when a finite number of shots is used. We solve the 1D Heisenberg model, but gradients are now estimated using n_{shot} shots.

The gradient with a finite number of shots can be obtained as follows. First, we introduce another PQC with the same shape as the HVA, Eq. (33), but all gates have different parameters.

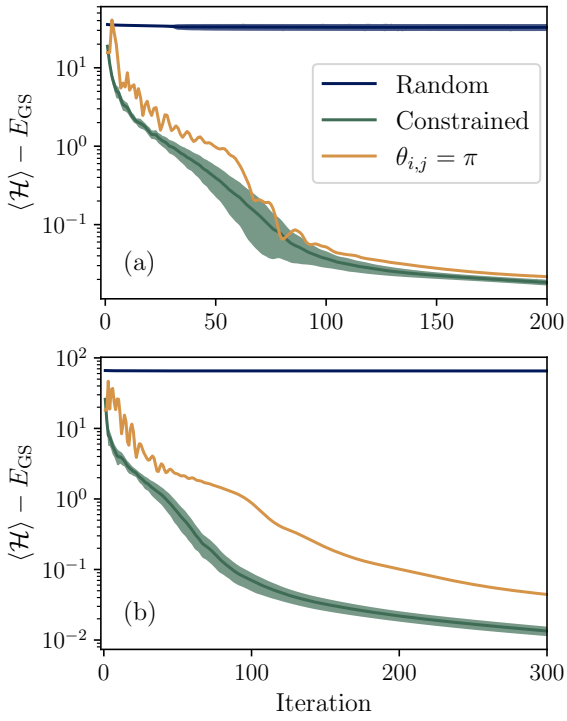

Figure 5: (a) Learning curves from different initialization schemes for the 1D Heisenberg model with $N = 20$. We use the circuit ansatz with $p = 20$ and the Adam optimizer with the learning rate $\alpha = 0.025$. For each iteration, the curves for random and constrained initializations show the averaged results over 32 different initial parameters. The shaded regions indicate one standard deviation ($[m - \sigma/2, m + \sigma/2]$). Note that results from the constant initialization ($\theta_{i,j} = \pi$) do not vary between instances as there is no randomness in the simulation. The ground state energy E_{GS} is obtained using the exact diagonalization. (b) The same result for the 2D Heisenberg model with $L_x \times L_y = 4 \times 6$ lattice. The circuit ansatz with $p = 24$ and the Adam optimizer with learning rate $\alpha = 0.005$ are used. Results for random and constrained initializations are averaged over 16 different initial parameters. For optimization, we compute the gradient exactly (without shot noise). Shaded regions are barely visible for the random initialization. This is because the loss function, $\langle\mathcal{H}\rangle$, is not trained at all for most instances. Thus, the loss function preserves its initial value $\langle\mathcal{H}\rangle \approx 0$, which is from the fact that the circuit forms a 2-design for the random initialization.

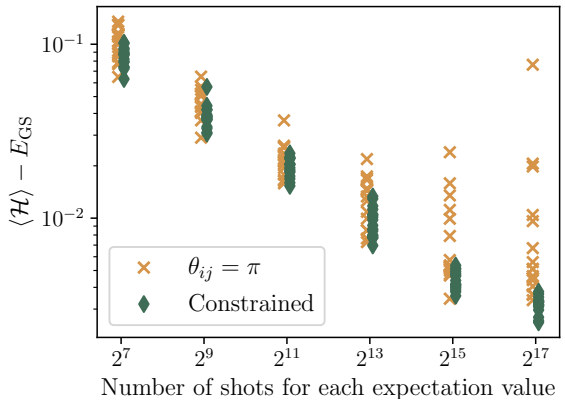

Figure 6: Converged energies from the VQE for the one-dimensional Heisenberg model with $N = p = 16$ as a function of the number of shots. For each number of shots, $n_{\mathrm{shot}} \in [2^7, 2^9, 2^{11}, 2^{13}, 2^{15}, 2^{17}]$, we fully simulate the VQE 16 times. The converged energies $\langle\mathcal{H}\rangle$ for each independent VQE instance are presented. For the initialization $\theta_{ij} = \pi$, the shot noise is the only source of the randomness. On the other hand, the initial parameters are also random when the constrained parameter initialization is used.

Such a PQC is written as

$$\begin{aligned}&|\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})\rangle\\ &= \prod_{i=p}^{1} e^{-i\sum_{\langle a,b\rangle}\gamma_{i,a,b}Z_aZ_b} e^{-i\sum_{\langle a,b\rangle}\beta_{i,a,b}Y_aY_b}\\ &\qquad \times e^{-i\sum_{\langle a,b\rangle}\alpha_{i,a,b}X_aX_b}|\psi_0\rangle\,. \end{aligned} \tag{34}$$

Compared to Eq. (33), all gates now have independent parameters. Next, we obtain each gradient component using the two-term parameter-shift rule [57, 58]. By defining $f(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \langle\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})|\mathcal{H}|\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})\rangle$, we obtain its gradient for $\alpha_{i,a,b}$ as follows:

$$\begin{aligned}\frac{\partial f}{\partial\alpha_{i,a,b}} = \frac{1}{2}\Big[&f(\boldsymbol{\alpha}+\frac{\pi}{2}\boldsymbol{\delta}_{i,a,b},\boldsymbol{\beta},\boldsymbol{\gamma})\\ &- f(\boldsymbol{\alpha}-\frac{\pi}{2}\boldsymbol{\delta}_{i,a,b},\boldsymbol{\beta},\boldsymbol{\gamma})\Big],\end{aligned} \tag{35}$$

where $\boldsymbol{\delta}_{i,a,b}$ is a vector components of which is 1, if the index is (i,a,b), or 0, otherwise. Gradient for β and γ also can be obtained similarly.

Shot noise is introduced when we estimate $f(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$. From

$$\langle\mathcal{H}\rangle = \sum_{\langle a,b\rangle}\langle X_aX_b\rangle + \langle Y_aY_b\rangle + \langle Z_aZ_b\rangle\,, \tag{36}$$

we can estimate $f = \langle\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})|\mathcal{H}|\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})\rangle$ using the samples of $\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ in the X, Y, and Z bases. For example, let us estimate $\langle X_aX_b\rangle$ using n_{shot} samples. We first obtain bitstrings $\{x^{(1)},\cdots,x^{(n_{\mathrm{shot}})}\}$ from the probability distribution $p(x) = |\langle x|H^{\otimes N}|\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})\rangle|^2$, where each $x^{(i)} \in \{0,1\}^N$ is a bitstring with length N. Then, each $\langle X_aX_b\rangle$ can be estimated using these samples by

$$\langle X_aX_b\rangle \approx \frac{1}{n_{\mathrm{shot}}}\sum_{i=1}^{n_{\mathrm{shot}}}(1-2x_a^{(i)})(1-x_b^{(i)}), \tag{37}$$

where $1-2x_a^{(i)}$ is from the fact that X_a has the value 1 if $x_a^{(i)} = 0$, and -1 if $x_a^{(i)} = 1$. Note that we can use the same set of samples, $\{x^{(1)},\cdots,x^{(n_{\mathrm{shot}})}\}$, for all $\langle X_aX_b\rangle$. Thus, $f = \langle\mathcal{H}\rangle$ for each set of parameters is estimated using $3n_{\mathrm{shot}}$ samples, and each gradient component is estimated using $6n_{\mathrm{shot}}$ samples.

Finally, the gradient of the original HVA, Eq. (33), which shares the parameters between the gates, can be obtained by summing over the gradient components. Namely, we have

$$\frac{\partial\,\langle\psi(\boldsymbol{\theta})|\mathcal{H}|\psi(\boldsymbol{\theta})\rangle}{\theta_{i,1}} = \sum_{\langle a,b\rangle}\frac{\partial f}{\partial\alpha_{i,a,b}}, \tag{38}$$

$$\frac{\partial\,\langle\psi(\boldsymbol{\theta})|\mathcal{H}|\psi(\boldsymbol{\theta})\rangle}{\theta_{i,2}} = \sum_{\langle a,b\rangle}\frac{\partial f}{\partial\beta_{i,a,b}}, \tag{39}$$

$$\frac{\partial\,\langle\psi(\boldsymbol{\theta})|\mathcal{H}|\psi(\boldsymbol{\theta})\rangle}{\theta_{i,3}} = \sum_{\langle a,b\rangle}\frac{\partial f}{\partial\alpha_{i,a,b}}. \tag{40}$$

Given that the circuit has $3Np$ gates in total, and each gradient component is estimated from $6n_{\mathrm{shot}}$ samples, $18Npn_{\mathrm{shot}}$ samples are used for each iteration to estimate all gradient components.

For the HVA with $N = p = 16$ [see Eq. (33)], we plot the converged energies after 10^3 iterations from the VQE simulations as a function of $n_{\mathrm{shot}} \in [2^7, 2^9, 2^{11}, 2^{13}, 2^{15}, 2^{17}]$ in Fig. 6. While the results show that the best-converged energies from our constrained initialization scheme and $\theta_{ij} = \pi$ are similar, the deviations between instances from our initialization scheme are much smaller. For example, the worst-performing instance for $n_{\mathrm{shot}} = 2^{15}$ gives $\langle\mathcal{H}\rangle - E_{\mathrm{GS}} \approx 3\times10^{-2}$ when all parameters are initialized with π, but that from our initialization is $\approx 6\times10^{-3}$.

Our parameter constraint can also be imposed throughout the optimization steps (the ansatz itself). When imposed on the ansatz, we can slightly change the cost function to ensure the parameters always follow the constraints. This can

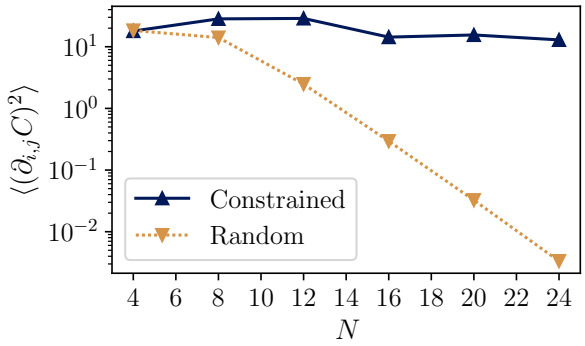

Figure 7: Scaling of the averaged squared gradients $(\partial_{i,j}C)^2$ from the repeated ansatz Eq. (41) with (solid) and without (dotted) a parameter constraint. For the parameter-constrained ansatz, we sample parameters under the constraint $\theta_{i,1}+\theta_{i,2}+\theta_{i,3}=\pi/(2N)$. In contrast, $\theta_{i,j}\sim\mathcal{U}_{[0,2\pi]}$ is used for the ansatz without the constraint. The HVA is for the 1D XYZ model with $O=Y_0Y_1$ with $\tilde{p}=16$ and $r=N^2/4$. The results are averaged over 2^{10} random parameters.

be simply done by assigning $\theta_{j,q}=T-\sum_{i=1}^{q-1}\theta_{j,i}$ and replacing the cost function C with $\tilde{C}=[\prod_{i,j}\mathbf{1}(\theta_{j,i})][\prod_j\mathbf{1}(T-\sum_{i=1}^{q-1}\theta_{j,i})]C$ where $\mathbf{1}(x)$ is the Heaviside step function. One can see this (piecewise differentiable) cost function (1) restricts all parameters to be larger than 0, and (2) the sum of the parameters in each block is given by T throughout the training.

Still, the constraint-imposed ansatz may not be useful for solving complex problems, as we require pT to be small. Even though the ansatz itself allows a large-depth circuit (e.g., $p=\Theta(N^2)$ and $T=1/N^3$), the Suzuki-Trotter decomposition [59] tells us that such a circuit always can be approximated by a short-depth circuit, i.e., there is another circuit with depth $d=\mathrm{poly}(pT)=\mathrm{poly}(1/N)$ that can express our constrained HVA with a small error. Using the notion of quantum circuit complexity [60, 61, 62, 63], defined by the minimum number of two-qubit gates in any circuit that implements the given unitary, we can say that this ansatz has a small approximate circuit complexity.

5.3 Long-time evolution with repeated parameters

To overcome the problem that a simple parameter-constrained ansatz introduced in the previous subsection is not expressive enough, we propose another ansatz with better expressivity. Our solution is to repeat the circuit multiple times instead of adding free parameters. In this case, the circuit is given as

$$U(\boldsymbol{\theta})=\Big[\prod_{i=\tilde{p}}^{1}e^{-iH^{(q)}\theta_{i,q}}\cdots e^{-iH^{(1)}\theta_{i,1}}\Big]^r \tag{41}$$

with the constraint $\sum_j\theta_{i,j}=T$. Thus, the circuit has a total of $\tilde{p}qr$ layers but only has $\tilde{p}q$ parameters. This ansatz can be approximated by $e^{-iK(\tilde{p}rT)}$ for a local Hamiltonian K with an error $\mathcal{O}(r(\tilde{p}T)^{n+2})$ when $\tilde{p}T$ is inverse polynomial with N (i.e., $\tilde{p}T=\mathcal{O}(N^{-\gamma})$ for $\gamma>0$). This fact follows from Proposition 3 and $\|U_1^r-U_2^r\|\leq r\|U_1-U_2\|$ which holds for arbitrary U_1 and U_2 [2].

We then further expect that the gradients scale polynomially with N from Conjecture 1 when $\tilde{p}rT$ is sufficiently large enough to equilibrate the system [3]. We also numerically test the gradient scaling of this ansatz for $r=N^2/4$, $\tilde{p}=16$, and $T=\pi/(2N)$ using the HVA for the one-dimensional XYZ model in Fig. 7. The plot shows that the gradient does not decay when the parameters are constrained, whereas it decays exponentially otherwise.

We now argue that the repeated ansatz of Eq. (41) (1) can generate sufficiently complex unitary operators and (2) is useful for variational time evolution [33, 34, 35, 36]. The complexity of the circuit directly follows from the observation that the circuit approximates to $e^{-iK(\tilde{p}rT)}$ for a Hamiltonian K with a large $\tilde{p}rT$. It is commonly believed that simulating the long-time evolution of a general local Hamiltonian requires a large depth circuit (which is also formally conjectured in Refs. [64, 65] in terms of quantum circuit complexity). Next, the given ansatz can express the time evolution of a given Hamiltonian $H=\sum_{j=1}^{\tilde{p}}\alpha_jH_j$. Using the first-order Suzuki-Trotter decomposition, we can write

$$e^{-iHt}=(e^{-iHt/r})^r\approx\Big[\prod_{j=\tilde{p}}^{1}e^{-i\alpha_jH_jt_0}\Big]^r \tag{42}$$

with $t_0=t/r$ and an error $\mathcal{O}(Nrt_0^2)$. Thus the approximation has an error $\mathcal{O}(1/N)$ if we use $t_0=$

[2] This is from $\|U_1^r-U_2^r\|=\|U_1U_1^{r-1}-U_1U_2^{r-1}+U_1U_2^{r-1}-U_2U_2^{r-1}\|\leq\|U_1^{r-1}-U_2^{r-1}\|+\|U_1-U_2\|$, where we used $\|U_1\|=\|U_2^{r-1}\|=1$.

[3] Precisely, one also require an ergodicity assumption that resulting Hamiltonians (Ks) are uniformly distributed over the vector space for random parameters $\theta_{i,j}$.

$\Theta(1/N)$. As the right-hand side is nothing but Eq. (41) with $T = \sum_j \alpha_j t_0$, the ansatz with $T = \Theta(1/N)$ can approximate e^{-iHt}.

6 Conclusion

We studied the scaling behaviors of the gradients in the hamiltonian variational ansatz (HVA) and showed that adding a simple parameter constraint to the ansatz results in large gradients. We demonstrated that the gradient magnitudes scale as $\Theta(1)$ when the circuit is given by short-time evolution and $1/\text{poly}(N)$ when it is given by long-time evolution. For the short-time regime, we provided a rigorous proof based on the rate of the gradient evolution, while we showed numerical evidence based on quantum thermalization [38, 39, 40, 45] for long-time evolution. We then found the parameter constraints for which the HVA can be approximated by short-time as well as long-time evolution under a local Hamiltonian. We further supported our arguments with extensive numerics for up to 28 qubits, which also consistently showed the correctness of our arguments.

For long-time evolution, our argument is based on the fact that the dynamics generated by thermalizing Hamiltonians are more restricted than unitary 2-designs. Albeit typical Hamiltonians thermalize [44], there are two other important classes of Hamiltonians with different dynamic properties: integrable and many-body localized systems. In contrast to thermalizing systems where information of initial states spreads out through the Hilbert space (but within a subspace preserving the energy), initial information on integrable and many-body localized systems can be easily accessed by simple operators at any time, i.e., their dynamics are even more restrictive than thermalizing Hamiltonians. Given this interesting property, we expect that there could be a different parameter condition that parameterized quantum circuits approximate to integrable or many-body localized systems, which are also free from barren plateaus.

For example, it is known that the out-of-time correlator of many-body localized systems does not decay exponentially with the system size for particular choices of observables and initial states (see, e.g., Refs. [66, 67, 68]). Following the arguments in Sec. 3.2, we hope to find a class of parameterized quantum circuits that approximate to a many-body localized system and have large gradients. On the other hand, Ref. [32] showed that the dynamic Lie algebra $\mathcal{G}$ generated by the HVA for the XXZ model ($J_x = J_y$) can have a small dimension, i.e., the Lie algebra $\langle i\sum_i(X_iX_{i+1} + Y_iY_{i+1}), i\sum_i Z_iZ_{i+1}\rangle_{\text{Lie}}$, has a small dimension. As the XXZ model is solvable by the Bethe ansatz (thus integrable), we believe that a fundamental connection exists between the low-dimensional dynamical Lie algebra, the integrability of the system, and large gradients. Such a connection might be studied in future work.

This paper did not consider incoherent noise, which prevails in noisy quantum devices. This type of noise is known to be another source of barren plateaus [69]. When the circuit is short enough, we believe that our initialization schemes can help compensate for the vanishing gradients from the incoherent noises. Still, how the strength of the noise, the circuit depth, and the initialization schemes interplay in general is another big question that requires a separate study. Since this is important for practical applications of variational quantum algorithms, further research on the effect of incoherent noises is necessary.

Acknowledgements

The authors thank Modjtaba Shokrian Zini for helpful discussions and David Wierichs, Maria Schuld, and Joseph Bowles for valuable comments. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award NERSC DDR-ERCAP0025705. Numerical simulations were performed using PennyLane [55] software package with Lightning [70] and Lightning-GPU [71] plugins. The source code used for simulations is available in Ref. [72].

References

[1] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo,

Fernando GSL Brandao, David A Buell, et al. “Quantum supremacy using a programmable superconducting processor”. Nature **574**, 505–510 (2019).

[2] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. “Quantum computational advantage using photons”. Science **370**, 1460–1463 (2020).

[3] Lars S Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob FF Bulmer, Filippo M Miatto, Leonhard Neuhaus, Lukas G Helt, Matthew J Collins, et al. “Quantum computational advantage with a programmable photonic processor”. Nature **606**, 75–81 (2022).

[4] John Preskill. “Quantum computing in the NISQ era and beyond”. Quantum **2**, 79 (2018).

[5] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. “A quantum approximate optimization algorithm” (2014). arXiv:1411.4028.

[6] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O’Brien. “A variational eigenvalue solver on a photonic quantum processor”. Nat. Comm. **5**, 1–7 (2014).

[7] Dave Wecker, Matthew B Hastings, and Matthias Troyer. “Progress towards practical quantum variational algorithms”. Phys. Rev. A **92**, 042303 (2015).

[8] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. “Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets”. Nature **549**, 242–246 (2017).

[9] Stuart Hadfield, Zhihui Wang, Bryan O’Gorman, Eleanor G Rieffel, Davide Venturelli, and Rupak Biswas. “From the quantum approximate optimization algorithm to a quantum alternating operator ansatz”. Algorithms **12**, 34 (2019).

[10] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. “An introduction to quantum machine learning”. Contemporary Physics **56**, 172–185 (2015).

[11] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. “Quantum machine learning”. Nature **549**, 195–202 (2017).

[12] Maria Schuld and Nathan Killoran. “Quantum machine learning in feature Hilbert spaces”. Phys. Rev. Lett. **122**, 040504 (2019).

[13] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. “A rigorous and robust quantum speed-up in supervised machine learning”. Nat. Phys. **17**, 1013–1017 (2021).

[14] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. “Variational quantum algorithms”. Nat. Rev. Phys. **3**, 625–644 (2021).

[15] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. “Barren plateaus in quantum neural network training landscapes”. Nat. Comm. **9**, 1–6 (2018).

[16] Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. “Cost function dependent barren plateaus in shallow parametrized quantum circuits”. Nat. Comm. **12**, 1–12 (2021).

[17] Zoë Holmes, Kunal Sharma, Marco Cerezo, and Patrick J Coles. “Connecting ansatz expressibility to gradient magnitudes and barren plateaus”. PRX Quantum **3**, 010313 (2022).

[18] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. Neural computation **9**, 1735–1780 (1997).

[19] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. In Proceedings of the fourteenth international conference on artificial intelligence and statistics. Pages 315–323. JMLR Workshop and Conference Proceedings (2011). url: https://proceedings.mlr.press/v15/glorot11a.html.

[20] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In Proceedings of the thirteenth international conference on artificial intelligence and statistics. Pages 249–256. JMLR Workshop and Conference Proceedings (2010). url: https://proceedings.mlr.press/v9/glorot10a.html.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In Proceedings of the IEEE international conference on computer vision. Pages 1026–1034. (2015).

[22] Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. "Toward trainability of quantum neural networks" (2020). arXiv:2011.06258.

[23] Tyler Volkoff and Patrick J Coles. "Large gradients via correlation in random parameterized quantum circuits". Quantum Science and Technology **6**, 025008 (2021).

[24] Arthur Pesah, Marco Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. "Absence of barren plateaus in quantum convolutional neural networks". Phys. Rev. X **11**, 041011 (2021).

[25] Xia Liu, Geng Liu, Jiaxin Huang, Hao-Kai Zhang, and Xin Wang. "Mitigating barren plateaus of variational quantum eigensolvers" (2022). arXiv:2205.13539.

[26] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. "An initialization strategy for addressing barren plateaus in parametrized quantum circuits". Quantum **3**, 214 (2019).

[27] Nishant Jain, Brian Coyle, Elham Kashefi, and Niraj Kumar. "Graph neural network initialisation of quantum approximate optimisation". Quantum **6**, 861 (2022).

[28] Kaining Zhang, Liu Liu, Min-Hsiu Hsieh, and Dacheng Tao. "Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits". In Advances in Neural Information Processing Systems. Volume 35, pages 18612–18627. (2022). url: https://doi.org/10.48550/arXiv.2203.09376.

[29] Antonio A. Mele, Glen B. Mbeng, Giuseppe E. Santoro, Mario Collura, and Pietro Torta. "Avoiding barren plateaus via transferability of smooth solutions in a Hamiltonian variational ansatz". Phys. Rev. A **106**, L060401 (2022).

[30] Manuel S Rudolph, Jacob Miller, Danial Motlagh, Jing Chen, Atithi Acharya, and Alejandro Perdomo-Ortiz. "Synergistic pretraining of parametrized quantum circuits via tensor networks". Nature Communications **14**, 8367 (2023).

[31] Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. "Exploring entanglement and optimization within the Hamiltonian variational ansatz". PRX Quantum **1**, 020319 (2020).

[32] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J Coles, and M Cerezo. "Diagnosing barren plateaus with tools from quantum optimal control". Quantum **6**, 824 (2022).

[33] Ying Li and Simon C Benjamin. "Efficient variational quantum simulator incorporating active error minimization". Phys. Rev. X **7**, 021050 (2017).

[34] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C Benjamin. "Theory of variational quantum simulation". Quantum **3**, 191 (2019).

[35] Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J Coles, and Andrew Sornborger. "Variational fast forwarding for quantum simulation beyond the coherence time". npj Quantum Information **6**, 1–10 (2020).

[36] Sheng-Hsuan Lin, Rohit Dilip, Andrew G Green, Adam Smith, and Frank Pollmann. "Real-and imaginary-time evolution with compressed quantum circuits". PRX Quantum **2**, 010342 (2021).

[37] Conor Mc Keever and Michael Lubasch. "Classically optimized hamiltonian simulation". Phys. Rev. Res. **5**, 023146 (2023).

[38] Josh M Deutsch. "Quantum statistical mechanics in a closed system". Phys. Rev. A **43**, 2046 (1991).

[39] Mark Srednicki. "Chaos and quantum thermalization". Phys. Rev. E **50**, 888 (1994).

[40] Marcos Rigol, Vanja Dunjko, and Maxim Olshanii. "Thermalization and its mechanism for generic isolated quantum systems". Nature **452**, 854–858 (2008).

[41] Peter Reimann. "Foundation of statistical mechanics under experimentally realistic conditions". Phys. Rev. Lett. **101**, 190403 (2008).

[42] Noah Linden, Sandu Popescu, Anthony J Short, and Andreas Winter. "Quantum mechanical evolution towards thermal equilibrium". Phys. Rev. E **79**, 061103 (2009).

[43] Anthony J Short. "Equilibration of quantum systems and subsystems". New Journal of Physics **13**, 053009 (2011).

[44] Christian Gogolin and Jens Eisert. "Equilibration, thermalisation, and the emergence of statistical mechanics in closed quantum systems". Reports on Progress in Physics **79**, 056001 (2016).

[45] Yichen Huang, Fernando GSL Brandão, Yong-Liang Zhang, et al. "Finite-size scaling of out-of-time-ordered correlators at late times". Phys. Rev. Lett. **123**, 010601 (2019).

[46] Daniel A Roberts and Beni Yoshida. "Chaos and complexity by design". Journal of High Energy Physics **2017**, 1–64 (2017).

[47] Hyungwon Kim, Tatsuhiko N Ikeda, and David A Huse. "Testing whether all eigenstates obey the eigenstate thermalization hypothesis". Phys. Rev. E **90**, 052105 (2014).

[48] Tomotaka Kuwahara, Takashi Mori, and Keiji Saito. "Floquet–Magnus theory and generic transient dynamics in periodically driven many-body quantum systems". Annals of Physics **367**, 96–124 (2016).

[49] David Wierichs, Christian Gogolin, and Michael Kastoryano. "Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer". Phys. Rev. Research **2**, 043246 (2020).

[50] Chae-Yeun Park. "Efficient ground state preparation in variational quantum eigensolver with symmetry breaking layers" (2021). arXiv:2106.02509.

[51] Jan Lukas Bosse and Ashley Montanaro. "Probing ground-state properties of the kagome antiferromagnetic heisenberg model using the variational quantum eigensolver". Phys. Rev. B **105**, 094409 (2022).

[52] Joris Kattemölle and Jasper van Wezel. "Variational quantum eigensolver for the heisenberg antiferromagnet on the kagome lattice". Phys. Rev. B **106**, 214429 (2022).

[53] Diederik P. Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. (2015). url: https://doi.org/10.48550/arXiv.1412.6980.

[54] Tyson Jones and Julien Gacon. "Efficient calculation of gradients in classical simulations of variational quantum algorithms" (2020). arXiv:2009.02823.

[55] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M. Sohaib Alam, Guillermo Alonso-Linaje, et al. "Pennylane: Automatic differentiation of hybrid quantum-classical computations" (2018). arXiv:1811.04968.

[56] Lodewyk FA Wessels and Etienne Barnard. "Avoiding false local minima by proper initialization of connections". IEEE Transactions on Neural Networks **3**, 899–905 (1992).

[57] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. "Quantum circuit learning". Phys. Rev. A **98**, 032309 (2018).

[58] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. "Evaluating analytic gradients on quantum hardware". Phys. Rev. A **99**, 032331 (2019).

[59] Masuo Suzuki. "General theory of fractal path integrals with applications to many-body theories and statistical physics". Journal of Mathematical Physics **32**, 400–407 (1991).

[60] Michael A. Nielsen. "A geometric approach to quantum circuit lower bounds" (2005). arXiv:quant-ph/0502070.

[61] Michael A Nielsen, Mark R Dowling, Mile Gu, and Andrew C Doherty. "Quantum computation as geometry". Science **311**, 1133–1135 (2006).

[62] Douglas Stanford and Leonard Susskind. "Complexity and shock wave geometries". Phys. Rev. D **90**, 126007 (2014).

[63] Jonas Haferkamp, Philippe Faist, Naga BT Kothakonda, Jens Eisert, and Nicole Yunger Halpern. "Linear growth of quantum circuit complexity". Nat. Phys. **18**, 528–532 (2022).

[64] Adam R Brown, Leonard Susskind, and Ying Zhao. "Quantum complexity and negative curvature". Phys. Rev. D **95**, 045010 (2017).

[65] Adam R Brown and Leonard Susskind. "Second law of quantum complexity". Phys. Rev. D **97**, 086015 (2018).

[66] Yu Chen. "Universal logarithmic scrambling in many body localization" (2016). arXiv:1608.02765.

[67] Ruihua Fan, Pengfei Zhang, Huitao Shen, and Hui Zhai. "Out-of-time-order correlation for many-body localization". Science Bulletin **62**, 707–711 (2017).

[68] Juhee Lee, Dongkyu Kim, and Dong-Hee Kim. "Typical growth behavior of the out-of-time-ordered commutator in many-body localized systems". Phys. Rev. B **99**, 184202 (2019).

[69] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. "Noise-induced barren plateaus in variational quantum algorithms". Nat. Comm. **12**, 6961 (2021).

[70] "PennyLane–Lightning plugin `https://github.com/PennyLaneAI/pennylane-lightning`" (2023).

[71] "PennyLane–Lightning-GPU plugin `https://github.com/PennyLaneAI/pennylane-lightning-gpu`" (2023).

[72] "GitHub repository `https://github.com/XanaduAI/hva-without-barren-plateaus`" (2023).

[73] Wilhelm Magnus. "On the exponential solution of differential equations for a linear operator". Commun. Pure. Appl. Math. **7**, 649–673 (1954).

[74] Dmitry Abanin, Wojciech De Roeck, Wen Wei Ho, and François Huveneers. "A rigorous theory of many-body prethermalization for periodically driven and closed quantum systems". Commun. Math. Phys. **354**, 809–827 (2017).

A Big-O and related notations

In the main text, we have used big-O and related notations. This appendix formally defines these notations as follows:

- $f(n) = \mathcal{O}(g(n))$ if there exist $n_0 \in \mathbb{N}^+$ and $c \in \mathbb{R}^+$ such that $f(n) \leq cg(n)$ for all $n \geq n_0$.
- $f(n) = \Omega(g(n))$ if there exist $N_0 \in \mathbb{N}^+$ and $c \in \mathbb{R}^+$ such that $f(n) \geq cg(n)$ for all $n \geq n_0$.
- $f(n) = \Theta(g(n))$ if $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$.

B Long-time average of the variance of gradients in the Hamiltonian dynamics

In this appendix, we prove Proposition 2, which gives the lower bound of a long-time average of the squared gradient for a Hamiltonian evolution. Direct computation of $\langle\psi_0|i[G, O(t)]|\psi_0\rangle^2$ gives

$$\langle\psi_0|i[G,O(t)]|\psi_0\rangle^2 = -\Big[\langle\psi_0|Ge^{iHt}Oe^{-iHt}|\psi_0\rangle - \langle\psi_0|e^{iHt}Oe^{-iHt}G|\psi_0\rangle\Big]^2 \tag{B.1}$$

$$= -\Big[\sum_{ijk} C_i^* G_{ij} e^{i(E_j-E_k)t} O_{jk} C_k - \sum_{lmn} C_l^* e^{i(E_l-E_m)t} O_{lm} G_{mn} C_n\Big]^2 \tag{B.2}$$

$$\begin{aligned} = &- \sum_{ijki'j'k'} C_i^* G_{ij} e^{i(E_j-E_k)t} O_{jk} C_k C_{i'}^* G_{i'j'} e^{i(E_{j'}-E_{k'})t} O_{j'k'} C_{k'} \\ &- \sum_{lmnl'm'n'} C_l^* e^{i(E_l-E_m)t} O_{lm} G_{mn} C_n C_{l'}^* e^{i(E_{l'}-E_{m'})t} O_{l'm'} G_{m'n'} C_{n'} \\ &+ 2\sum_{ijklmn} C_i^* G_{ij} e^{i(E_j-E_k)t} O_{jk} C_k C_l^* e^{i(E_l-E_m)t} O_{lm} G_{mn} C_n \end{aligned} \tag{B.3}$$

where $C_i = \langle E_i|\psi_0\rangle$, $G_{ij} = \langle E_i|G|E_j\rangle$, and $O_{jk} = \langle E_j|O|E_k\rangle$.

We assume that the Hamiltonian H satisfies the non-degenerate energy-gap condition, i.e.,

$$E_i - E_j = E_k - E_l \quad \text{iff} \quad \begin{cases} i = k \text{ and } j = l \\ i = j \text{ and } k = l \end{cases}. \tag{B.4}$$

Under this condition, averaging Eq. (B.3) over time yields

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T dt\Big\{-\langle[G,O(t)]\rangle^2\Big\} = 2\sum_{ijkn} C_i^* G_{ij} O_{jk} |C_k|^2 O_{kj} G_{jn} C_n - \sum_{lmnn'} C_l^* O_{lm} G_{mn} C_n C_m^* O_{ml} G_{ln'} C_{n'}$$
$$- \sum_{ijkl} C_i^* G_{ij} O_{jk} C_k C_l^* G_{lk} O_{kj} C_j - \langle\psi|[G,\tilde{O}]|\psi\rangle^2 \tag{B.5}$$

$$= F_H(\psi,G,O) - \langle\psi|[G,\tilde{O}]|\psi\rangle^2 \geq F_H(\psi,G,O) \tag{B.6}$$

where $\tilde{O} = \sum_j O_{jj}\,|E_j\rangle\langle E_j|$. As $\langle\psi|[G,\tilde{O}]|\psi\rangle$ is purely imaginary, we obtain the last inequality. The inequality in the main text is then obtained by changing the summation indices.

We also note that the eigenstate thermalization hypothesis [38, 39, 40] suggests that the last term, $\langle\psi|[G,\tilde{O}]|\psi\rangle^2$, is exponentially small in N.

C Generating random Hamiltonians

In the main text, we numerically observed the scaling behaviors of gradient magnitudes using random Hamiltonians. Here, we describe detailed steps to generate such random k-local Hamiltonians. For a given k, we first create a set of terms $S = \{\sigma_{a_1}^1\sigma_{a_2}^2\cdots\sigma_{a_k}^k\}$ where each $\sigma_{a_i}^i$ is one of the the Pauli matrices at site i ($\{I_i, X_i, Y_i, Z_i\}$) where $a_k \in \{0,1,2,3\}$ for all k. We then remove terms duplicated under translation. For example, as $X\otimes I\otimes I$, $I\otimes X\otimes I$, and $I\otimes I\otimes X$ generate the same terms under translation, we only keep one of them. We then construct a random Hamiltonian $H = \sum_{s\in S} c_s \sum_{n=1}^N \mathsf{T}^n s$, where the coefficients c_s are samples from the normal distribution $\mathcal{N}(0,1)$ and T is the translation operator ($\mathsf{T}\sigma_a^i = \sigma_a^{i+1}$). We also generate random time-reversal symmetric Hamiltonians ($H^* = H$) using the same method but removing purely imaginary operators (that contain odd numbers of Pauli-Ys) from S.

D Approximation of the HVA from the truncated Floquet-Magnus expansion

In this appendix, we prove Proposition 3 using the truncated Floquet-Magnus (FM) expansion. The FM expansion [73] provides a time-independent effective Hamiltonian for a unitary evolution from a time-dependent Hamiltonian. While this expansion diverges for a general many-body Hamiltonian, recent works [48, 74] have shown that we can still use the expansion after truncating high-order terms.

D.1 Truncated Floquet-Magnus expansion

Let us introduce the truncated FM expansion following the notation in Ref. [48]. We consider a system defined on a lattice with N spins where each spin is labeled by $i = 1$ to N. The set of all spins is denoted by $\Lambda = \{1, \cdots, N\}$. We consider a time-dependent Hamiltonian $H(t)$ defined for $0 \le t \le \tau$. We decompose the Hamiltonian into $H(t) = H_0 + V(t)$ where H_0 is the time-independent part and $V(t)$ is the remaining time-dependent part. Both parts have at most k-body interactions (we do not impose geometric locality yet). We then write the Hamiltonian terms as

$$H_0 = \sum_{|X|\le k} h_X, \qquad V(t) = \sum_{|X|\le k} v_X(t), \tag{D.7}$$

where X is all possible subsets of Λ and $|X|$ is the number of elements in the set.

We introduce a parameter J that upper bounds local interaction strength and some additional parameters for the expansion:

$$\sum_{X:X\ni i} (\|h_X\| + \|v_X(t)\|) \le J \quad \forall i \in \Lambda, \qquad V_0 := \sum_{|X|\le k} \frac{1}{\tau}\int_0^\tau \|v_X(t)\|dt, \qquad \lambda := 2kJ. \tag{D.8}$$

Under this setting, we are interested in the Floquet Hamiltonian H_F defined as

$$e^{-iH_F\tau} := \mathcal{T}[e^{-i\int_0^\tau H(t)dt}], \tag{D.9}$$

where $\mathcal{T}[\cdot]$ is the time-ordering operator. One can expand the Floquet Hamiltonian as $H_F = \sum_{n=0}^\infty \tau^n \Omega_n$ where the terms $\{\Omega_n\}_{n=0}^\infty$ are given by the FM expansion as follows:

$$\begin{aligned}\Omega_n = &\frac{1}{(n+1)^2}\sum_{\sigma\in S_{n+1}} (-1)^{n-\omega(\sigma)}\frac{\omega(\sigma)!(n-\omega(\sigma))!}{n!} \\ &\times \frac{1}{i^n\tau^{n+1}}\int_0^\tau dt_{n+1}\cdots\int_0^{t_3} dt_2\int_0^{t_2} dt_1[H(t_{\sigma(n+1)}),[H(t_{\sigma(n)}),\cdots,[H(t_{\sigma(2)}),H(t_{\sigma(1)})]\cdots]],\end{aligned} \tag{D.10}$$

where S_{n+1} is the permutation group on $n+1$ letters, $\omega(\sigma) = \sum_{i=1}^n \mathbf{1}[\sigma(i+1) - \sigma(i)]$, and $\mathbf{1}(x)$ is the Heaviside step function. In our setup, where the Hamiltonian terms have at most k-body interactions, Ω_n has at most $(n+1)k$-body interactions. We further have the following upper bound for Ω_n (Lemma 1 in Ref. [48]):

$$\|\Omega_n\| \le \frac{2V_0\lambda^n}{(n+1)^2}n! =: \overline{\Omega}_n. \tag{D.11}$$

One can see that the convergence condition $\|\Omega_{n+1}\|\tau^{n+1} < \|\Omega_n\|\tau^n$ only holds up to $n \approx (\lambda\tau)^{-1}$. Indeed, the FM expansion diverges for a many-body Hamiltonian unless τ also scales with N. Even though this fact suggests that the FM expansion might not be useful, it turned out that one can still use a truncated series $H_F^{(n)} := \sum_{m=0}^n \tau^m\Omega_m$ to describe long-time dynamics accurately:

Theorem D.1 (Theorem 1 in Ref. [48])**.** *With our parameters in Eq. (D.8) and additional condition $\tau \le 1/(4\lambda)$, the time evolution under the time-dependent Hamiltonian $H(t) = H_0 + V(t)$ is close to that generated by the truncated Floquet Hamiltonian $H_F^{(n_0)} = \sum_{m=0}^{n_0}\Omega_m\tau^m$ with*

$$n_0 := \left\lfloor \frac{1}{16\lambda\tau} \right\rfloor, \tag{D.12}$$

in the sense that

$$\|e^{-iH_F\tau} - e^{-iH_F^{(n_0)}\tau}\| \le 6V_0\tau 2^{-n_0}. \tag{D.13}$$

However, as n_0 increases with N, if $\tau \sim N^{-\alpha}$ for a positive α, the interaction range of $H_F^{(n_0)}$ increases with N. As we want strict locality in our Hamiltonian (it must be k'-local for a constant k'), a bound for $H_F^{(n)}$ for a fixed n should be useful, which is provided by the following Corollary.

Corollary D.1 (Corollary 1 in Ref. [48])**.** *Under the same condition, we have*

$$\|e^{-iH_F\tau} - e^{-iH_F^{(n)}\tau}\| \leq 6V_0\tau 2^{-n_0} + \overline{\Omega}_{n+1}\tau^{n+2}, \tag{D.14}$$

where $H_F^{(n)} = \sum_{m=0}^{n} \Omega_m \tau^m$ *for arbitrary* $n \leq n_0$.

D.2 Approximating the HVA

We now consider the HVA given by

$$U = \prod_{i=p}^{1} e^{-iH^{(q)}\theta_{i,q}} \cdots e^{-iH^{(1)}\theta_{i,1}} \tag{D.15}$$

where we assume that each $H^{(j)}$ is the sum of commuting Pauli strings (products of Pauli operators) acting on at most k geometrically local sites. For example, $\sum_{i=1}^{N} X_i X_{i+1}$ from the HVA for the transverse-field Ising model satisfies this (both for the periodic and open boundary conditions) with $k = 2$.

We now interpret the HVA as a time-dependent Hamiltonian given as

$$\tilde{H}(t) := \begin{cases} H^{(1)} & \text{for } 0 \leq t \leq \theta_{1,1} \\ H^{(2)} & \text{for } \theta_{1,1} \leq t \leq \theta_{1,1} + \theta_{1,2} \\ \cdots & \\ H^{(q)} & \text{for } \sum_{j=1}^{q-1} \theta_{1,j} \leq t \leq \sum_{j=1}^{q} \theta_{1,j} \\ H^{(1)} & \text{for } \sum_{j=1}^{q} \theta_{1,j} \leq t \leq \sum_{j=1}^{q} \theta_{1,j} + \theta_{2,1} \\ \cdots & \\ H^{(q)} & \text{for } \sum_{i=1}^{p} \sum_{j=1}^{q-1} \theta_{i,j} \leq t \leq \sum_{i=1}^{p} \sum_{j=1}^{q} \theta_{i,j} \end{cases}. \tag{D.16}$$

For convenience, we define $\Upsilon_{n,m} := \sum_{i=1}^{n-1} \sum_{j=1}^{q} \theta_{i,j} + \sum_{j=1}^{m} \theta_{n,j}$ which is the cumulative sum of $\{\theta\}$. For any subcircuit of the HVA $U_b \cdots U_a$, where $a = (i, j)$ and $b = (i', j')$ are indices for the layers, we consider $H(t) = H_0 + V(t)$ with $H_0 = 0$ and $V(t) = \tilde{H}(t + \Upsilon_{a-1})$ [Eq. (D.16)] defined for $0 \leq t \leq \Upsilon_b - \Upsilon_{a-1} := \tau$ (where we use Υ_{a-1} to denote the sum of the parameters before layer $a = (i, j)$ and $\Upsilon_b = \Upsilon_{i',j'}$ for $b = (i', j')$).

Parameters for the FM expansion can be obtained by writing $V(t)$ as

$$V(t) = \tilde{H}(t + \Upsilon_{a-1}) = \sum_{|X| \leq k} h_X(t). \tag{D.17}$$

Following the notation in the main text, we have

$$V_0 = \frac{1}{\tau} \int_0^\tau \sum_{|X| \leq k} \|h_X(t)\| dt \leq \sup_{0 \leq t \leq \tau} \sum_{|X| \leq k} \|h_X(t)\| = \max_m \sum_{|X| \leq k} \|h_X^{(m)}\| = H_{\max} \tag{D.18}$$

for V_0 defined in Eq. (D.8) and $H_{\max}$ defined in Eq. (26). Under this setup, applying Corollary D.1 to U_R and U_L defined in the main text yields Proposition 3. Precisely, for a given $n \le n_0 = \lfloor 1/(32kJt_{R,L})\rfloor$, there are $(n+1)k$-local Hamiltonians H_R and H_L such that

$$\left\|U_R - e^{-iH_R t_R}\right\| \le 6H_{\max}2^{-\lfloor 1/(32kJt_R)\rfloor}t_R + \frac{2H_{\max}(2kJ)^{n+1}}{(n+2)^2}(n+1)!t_R^{n+2}, \tag{D.19}$$

$$\left\|U_L - e^{-iH_L t_L}\right\| \le 6H_{\max}2^{-\lfloor 1/(32kJt_L)\rfloor}t_L + \frac{2H_{\max}(2kJ)^{n+1}}{(n+2)^2}(n+1)!t_L^{n+2} \tag{D.20}$$

are satisfied (where we put $\lambda = 2kJ$ from Eq. (D.8)).

Furthermore, H_R and H_L share any symmetries that $\{H^{(j)}\}$ have, which follows from the property of the commutator, i.e., $W[H_1,H_2]W^{-1} = W(H_1H_2 - H_2H_1)W^{-1} = (WH_1W^{-1})(WH_2W^{-1}) - (WH_2W^{-1})(WH_1W^{-1}) = [WH_1W^{-1}, WH_2W^{-1}]$. Thus, for example, if all $\{H^{(j)}\}$ are translationally invariant, the resulting Hamiltonians H_R and H_L are also translationally invariant.

Obtaining the norm of each term of H_R (H_L) is also possible. For convenience, let K be one of H_R or H_L defined by $K := H_F^{(n)} = \sum_{m=0}^{n}\Omega_m\tau^m$ for $\tau = t_R$ or $\tau = t_L$. For each time t, let us define $j[t] \in \{1,\cdots,q\}$ to be the index such that $V(t) = H^{(j[t])}$. Then $V(t_{\sigma(1)}) = H^{(j[t_{\sigma(1)}])} = \sum_X h_X^{(j[t_{\sigma(1)}])}$ where $h_X^{(j[t_{\sigma(1)}])}$ acts on at most k sites. Inserting this expression in Eq. (D.10) gives

$$\begin{aligned}\Omega_n = &\sum_X \frac{1}{(n+1)^2}\sum_{\sigma\in S_{n+1}}(-1)^{n-\omega(\sigma)}\frac{\omega(\sigma)!(n-\omega(\sigma))!}{n!}\\ &\times\frac{1}{i^n\tau^{n+1}}\int_0^\tau dt_{n+1}\cdots\int_0^{t_3}dt_2\int_0^{t_2}dt_1[H(t_{\sigma(n+1)}),[H(t_{\sigma(n)}),\cdots,[H(t_{\sigma(2)}),h_X^{(j[t_{\sigma(1)}])}]\cdots]].\end{aligned} \tag{D.21}$$

So we write $K = \sum_X k_{\tilde{X}}$ with

$$\begin{aligned}k_{\tilde{X}} = &\sum_{m=0}^{n}\frac{\tau^m}{(m+1)^2}\sum_{\sigma\in S_{m+1}}(-1)^{m-\omega(\sigma)}\frac{\omega(\sigma)!(m-\omega(\sigma))!}{m!}\\ &\times\frac{1}{i^m\tau^{m+1}}\int_0^\tau dt_{m+1}\cdots\int_0^{t_3}dt_2\int_0^{t_2}dt_1[H(t_{\sigma(m+1)}),[H(t_{\sigma(m)}),\cdots,[H(t_{\sigma(2)}),h_{X_i}^{(j[t_{\sigma(1)}])}]\cdots]].\end{aligned} \tag{D.22}$$

Locality of $k_{\tilde{X}}$ follows from the fact that the multicommutator $[H^{(i_n)},[H^{(i_{n-1})},\cdots,[H^{(1)},O]\cdots]]$ acts on at most $(n+1)k$ nearby sites for any operator O acting on at most k local sites, and the Hamiltonians $H^{(1)},\cdots,H^{(n)}$ are k-local. Precisely, each $k_{\tilde{X}}$ is supported by augmented sites $\tilde{X} = \{i\in\Lambda \mid \mathrm{dist}(i,X)\le nk\}$ where $\mathrm{dist}(i,X) = \min_{j\in X}\mathrm{dist}(i,j)$.

Finally, we obtain a bound of the norm of $k_{\tilde{X}}$ using the inequality

$$\begin{aligned}&\left\|\int_0^\tau dt_{m+1}\cdots\int_0^{t_3}dt_2\int_0^{t_2}dt_1[H(t_{\sigma(m+1)}),[H(t_{\sigma(m)}),\cdots,[H(t_{\sigma(2)}),h_X^{(j[t_{\sigma(1)}])}]\cdots]]\right\|\\ &\le\frac{\tau^{n+1}}{(n+1)!}\max_{i_1,\cdots,i_n}\|[H^{(i_n)},\cdots,[H^{(i_2)},h_X^{(i_1)}]\cdots]]\|,\end{aligned} \tag{D.23}$$

and the following lemma.

Lemma 1 (Consequence of Lemma 3 in Ref. [48])**.** *Let $\{H^{(j)}\}$ be k-local and $\sum_{X:X\ni i}\|h_X^{(j)}\| \le J$ for all j. Then for an arbitrary operator O supported on k local sites, we have*

$$\|[H^{(i_n)},[H^{(i_{n-1})},\cdots,[H^{(1)},O]\cdots]]\| \le (n!)(2kJ)^n\|O\|. \tag{D.24}$$

We thus have

$$\|k_{\tilde{X}}\| \leq \sum_{m=0}^{n} \frac{(2kJ)^m}{(m+1)^2} m! \tau^m, \tag{D.25}$$

where we use $|\sum_{\sigma \in S_{m+1}}| = (m+1)!$, $\theta(\sigma)!(m-\theta(\sigma))!/m! = \binom{m}{\theta(\sigma)}^{-1} \leq 1$, and $\|h_X^{(i_1)}\| = 1$ regardless of j as $\{h_X^{(j)}\}$ are Pauli words. As a consequence, we obtain

$$\|K\| \leq \sum_X \|k_{\tilde{X}}\| \leq H_{\max} \sum_{m=0}^{n} \frac{(2kJ)^m}{(m+1)^2} m! \tau^m, \tag{D.26}$$

where $H_{\max}$ is the maximum number of terms in $H^{(j)} = \sum_X h_X^{(j)}$ (defined in the main text).

E Proof of Theorem 1

We here provide a detailed proof showing that there exists $\tau_0 = \Theta(1/N)$ such that the HVA with $\sum_{i,j} \theta_{i,j} = t_R + t_L \leq \tau_0$ has large gradient components $\partial_{n,m} C$. Here, we consider the cost function C given by the expectation value of a local observable O acting on at most k_O sites and an initial state ρ_0 which gives $|\operatorname{Tr}\{\rho_0 [H^{(m)}, O]\}| = \Theta(1)$.

Our proof consists of three steps. First, we show that the error approximating the HVA to local Hamiltonian evolution from the FM expansion is $\mathcal{O}(1/N^2)$. Next, we derive all factors ($\|H_R\|$, $\|[H_L, O]\|$, etc.) in Proposition 1 from the FM expansion. We then complete the proof by combining steps to show that there exists $\tau_0 = \Theta(1/N)$ such that $|\partial_{n,m} C|$ is lower bounded by a constant.

E.1 Polynomially decaying bound of the error from the truncated FM expansion

Let us first analyze the error term in Proposition 3 for $t_R, t_L \leq c/N$ with $n = 1$. We note that $n = 1$ requires $n_0 = \lfloor 1/(32kJt_{R,L}) \rfloor \geq 1$, which is satisfied for $N \geq N_0 := 32ckJ$. In addition, we assume $kJ \geq 1$, which is true in our setting [see Eq. (27)]. Then, the error from the truncated FM expansion [the RHS of Eq. (29)] is given by

$$\begin{aligned} \epsilon &= \Big[6c \times 2^{-\lfloor N/(32ckJ) \rfloor} + \frac{4c^3 (2kJ)^2}{9} \frac{1}{N^2}\Big] \frac{H_{\max}}{N} \\ &\leq r \Big[6c \times 2^{-\lfloor N/(32ckJ) \rfloor} + \frac{4c^3 (2kJ)^2}{9} \frac{1}{N^2}\Big], \end{aligned} \tag{E.27}$$

where r is a constant such that $H_{\max} \leq rN$ (which is from $H_{\max} = \mathcal{O}(N)$).

We now use the following lemma to find N_1 such that the error is $\mathcal{O}(1/N^2)$ for $N \geq N_1$.

Lemma 2. *For a given $\kappa_1, \kappa_2, \alpha > 0$ and*

$$N_1 := \max\Big\{\frac{8}{\alpha}, -\frac{4}{\alpha} \log\Big[\Big(\frac{\alpha e}{8}\Big)^2 \frac{\kappa_2}{2\kappa_1}\Big]\Big\}, \tag{E.28}$$

the inequality

$$\kappa_1 2^{-\lfloor \alpha N \rfloor} + \frac{\kappa_2}{N^2} \leq 2\frac{\kappa_2}{N^2} \tag{E.29}$$

is satisfied for all $N \geq N_1$.

Proof. We first have

$$\kappa_1 2^{-\lfloor \alpha N \rfloor} - \frac{\kappa_2}{N^2} \leq 2\kappa_1 e^{-\alpha N/2} - \frac{\kappa_2}{N^2} = \frac{2\kappa_1 N^2 e^{-\alpha N/2} - \kappa_2}{N^2}. \tag{E.30}$$

Let us define $f(N) := 2\kappa_1 N^2 \exp[-\alpha N/2] = 2\kappa_1 \exp[-\alpha N/2 + 2\log N]$. For $N \geq 8/\alpha$, we have

$$\log N \leq \frac{\alpha}{8} N + \log[\frac{8}{\alpha e}]. \tag{E.31}$$

Thus,

$$f(N) \leq 2\kappa_1 (\frac{8}{\alpha e})^2 \exp[-\frac{\alpha}{4} N] \tag{E.32}$$

for all $N \geq 8/\alpha$. One sees that the RHS is smaller than κ_2 when

$$N \geq -\frac{4}{\alpha} \log\Big[(\frac{\alpha e}{8})^2 \frac{\kappa_2}{2\kappa_1}\Big], \tag{E.33}$$

which completes the proof. □

We then apply this lemma to obtain the upper bound of Eq. (E.27). Inserting $\alpha = 1/(32ckJ)$, $\kappa_1 = 6c$, and $\kappa_2 = 4c^3(2kJ)^2/9$ gives $N_1 = 128\gamma ckJ$ where $\gamma = \log(4^7 \cdot 3^3/e^2) \approx 11.00$. As $N_1 \geq N_0$, we have $\epsilon \leq \beta(c)/N^2$ for all $N \geq \max\{N_0, N_1\} = 128\gamma ckJ$ with $\beta(c) = 8c^3 r(2kJ)^2/9$. The obtained bounds tells us that the U_L and U_R in Proposition 3 which appear in $\partial_{n,m} C$ approximate to local Hamiltonian evolution. Precisely, for $U_R = e^{-iH^{(m-1)}\theta_{n,m-1}} \cdots e^{-iH^{(1)}\theta_{i,1}}$ and $U_L = e^{-iH^{(q)}\theta_{p,q}} \cdots e^{-iH^{(m)}\theta_{n,m}}$, there are $2k$-local Hamiltonians H_R, H_L such that $\|U_{R,L} - e^{-iH_{R,L}t_{R,L}}\| \leq \beta(c)/N^2$ if $t_R, t_L \leq c/N$ for $N \geq 128\gamma ckJ$ where $t_R = \theta_{1,1} + \cdots + \theta_{n,m-1}$ and $t_L = \theta_{n,m} + \cdots + \theta_{p,q}$.

E.2 Condition of the constant for large gradients

We next find an upper bound of c (for $\tau_0 = c/N$) from the complete expression of time t_c in Proposition 1. From Eq. (D.26) with the first order expansion ($n = 1$), we have

$$\|H_{R,L}\| \leq H_{\max}\Big(1 + \frac{kJ}{2} t_{R,L}\Big). \tag{E.34}$$

Using this inequality, we find

$$\|H_R\| \leq H_{\max}\Big(1 + \frac{kJ}{2} t_R\Big), \qquad \|[H_L, O]\| \leq 2l\|O\|\Big(1 + \frac{kJ}{2} t_L\Big), \tag{E.35}$$

where we obtain the second inequality by combining Eq. 16 and Eq. D.25. Here, $l = |\{X : [k_{\tilde{X}}, O] \neq 0\}| \geq 1$ is a constant for a given lattice, which follows from the fact that $k_{\tilde{X}}$ acts on at most $2k$ nearby sites and O is a local operator.

In addition, we have

$$\|H^{(m)}\| \leq H_{\max}, \qquad \|[H^{(m)}, O]\| \leq 2s\|O\| \tag{E.36}$$

where $s = |\{X : [h_X, O] \neq 0\}| \geq 1$ is also a constant. We used the fact that each h_X is a Pauli string to obtain the second inequality.

As $t_R, t_L \leq \tau_0 = c/N$ (from $t_R, t_L \geq 0$ and $t_R + t_L \leq \tau_0$), we obtain for $N \geq N_0 \geq 32ckJ$,

$$\|H_R\| \leq \frac{65}{64} H_{\max} := \mu H_{\max}, \qquad \|[H_L, O]\| \leq 2l\|O\| \times \frac{65}{64} := 2\mu l\|O\| \tag{E.37}$$

where $\mu = 65/64$.

Using the fact that $\|H_{\max}\| \leq rN$, Proposition 1 yields

$$t_c \geq \frac{g}{8\mu r\|O\| \max\{\mu l, s\}} \frac{1}{N}, \tag{E.38}$$

where $g = |\operatorname{Tr}[\rho_0[H^{(m)}, O]]|$ is the magnitude of the gradient when the circuit is trivial ($U_R = U_L = \mathbb{1}$). Thus $t_R + t_L \leq t_c$ is satisfied for all c such that

$$c \leq \frac{g}{8\mu r\|O\| \max\{\mu l, s\}}. \tag{E.39}$$

We still note that the current condition implies large gradients only when $U_{R,L}$ are exact time-evolution operators. As there is an approximation error from the FM expansion, we consider this factor in the following subsection.

E.3 Bounding gradient with the FM truncation error

We introduce the following lemma to see how much an error from unitary approximation affects the gradients.

Lemma 3. *For a density matrix $\rho \geq 0$ and $\mathrm{Tr}[\rho] = 1$, Hermitian operators A, $\tilde{A}$, and unitary operators U, and $\tilde{U}$,*

$$|\mathrm{Tr}[U\rho U^\dagger A] - \mathrm{Tr}[\tilde{U}\rho\tilde{U}^\dagger\tilde{A}]| \leq \|A\|\|U-\tilde{U}\| + \|A-\tilde{A}\| + \|\tilde{A}\|\|U^\dagger - \tilde{U}^\dagger\|. \tag{E.40}$$

Proof.

$$\begin{aligned}
|\mathrm{Tr}[U\rho U^\dagger A] - \mathrm{Tr}[\tilde{U}\rho\tilde{U}^\dagger\tilde{A}]| &= |\mathrm{Tr}[\rho(U^\dagger AU - \tilde{U}^\dagger\tilde{A}\tilde{U})]| \\
&\leq \|U^\dagger AU - \tilde{U}^\dagger\tilde{A}\tilde{U}\| = \|U^\dagger AU - U^\dagger A\tilde{U} + U^\dagger A\tilde{U} - \tilde{U}^\dagger\tilde{A}\tilde{U}\| \\
&\leq \|A\|\|U-\tilde{U}\| + \|U^\dagger A - \tilde{U}^\dagger\tilde{A}\| \\
&\leq \|A\|\|U-\tilde{U}\| + \|A-\tilde{A}\| + \|\tilde{A}\|\|U^\dagger - \tilde{U}^\dagger\|.
\end{aligned}$$

□

Let us now apply this lemma to $\partial_{n,m}C = \mathrm{Tr}\{U_R\rho_0 U_R^\dagger[H^{(m)}, U_L^\dagger O U_L]\}$ with $U = U_R$, $\tilde{U} = e^{-iH_R t_R}$, $A = [H^{(m)}, U_L^\dagger O U_L]$, and $\tilde{A} = [H^{(m)}, e^{iH_L t_L} O e^{-iH_L t_L}]$. Denoting ϵ by the error from the FM expansion, i.e., $\|e^{-iH_R t_R} - U_R\| \leq \epsilon$ and $\|e^{-iH_L t_L} - U_L\| \leq \epsilon$, we obtain

$$\begin{aligned}
&|\mathrm{Tr}\{U_R\rho_0 U_R^\dagger[H^{(m)}, U_L^\dagger O U_L]\} - \mathrm{Tr}\{e^{-iH_R t_R}\rho_0 e^{iH_R t_R}[H^{(m)}, e^{iH_L t_L} O e^{-iH_L t_L}]\}| \\
&\leq \epsilon\|[H^{(m)}, U_L^\dagger O U_L]\| + \|[H^{(m)}, U_L^\dagger O U_L] - [H^{(m)}, e^{iH_L t_L} O e^{-iH_L t_L}]\| + \epsilon\|[H^{(m)}, e^{iH_L t_L} O e^{-iH_L t_L}]\| \\
&\leq 4\epsilon\|H^{(m)}\|\|O\| + 2\|H^{(m)}\|\|U_L^\dagger O U_L - e^{iH_L t_L} O e^{-iH_L t_L}\| \\
&\leq 8\epsilon\|H^{(m)}\|\|O\|,
\end{aligned} \tag{E.41}$$

where we used $\|[A,B]\| \leq 2\|A\|\|B\|$ to obtain the third line and $\|UOU^\dagger - \tilde{U}O\tilde{U}^\dagger\| \leq 2\|O\|\|U-\tilde{U}\|$ for the last inequality. As we obtained a bound $\epsilon \leq \beta(c)/N^2$ for $N \geq N_1(c)$ (final result in Sec. E.1) and $\|H^{(m)}\| \leq H_{\max} \leq rN$, the error is upper bounded by $8r\beta(c)\|O\|/N$ for a sufficiently large N. Thus, for $N \geq 32r\beta(c)\|O\|/g$, we can bound the error to be less than $g/4$. As Proposition 1 implies $|\mathrm{Tr}\{e^{-iH_R t_R}\rho_0 e^{iH_R t_R}[H^{(m)}, e^{iH_L t_L} O e^{-iH_L t_L}]\}| \geq g/2$, we have $|\mathrm{Tr}\{U_R\rho_0 U_R^\dagger[H^{(m)}, U_L^\dagger O U_L]\}| \geq g/4$ under this condition.

We summarize the overall result as follows. For the HVA for N qubits, a local operator O and an initial state ρ_0 are given. We assume there is a constant $g > 0$ and m such that $|\mathrm{Tr}\{\rho_0[H^{(m)}, O]\}| \geq g$ regardless of N. Then we fix

$$c = \frac{g}{8\mu r\|O\|\max\{\mu l, s\}}, \tag{E.42}$$

where r, l, s are constants obtained from the properties of $\{H^{(m)}\}$, and $\mu = 65/64$. Then for $N_{\min} = \max\{128\gamma ckJ, 32r\beta(c)\|O\|/g\}$,

$$\left|\frac{\partial C}{\partial\theta_{n,m}}\right| \geq \frac{1}{4}g \tag{E.43}$$

is satisfied for all $N \geq N_{\min}$ if $t_R + t_L \leq c/N$.

F Vanishing gradient after a finite time evolution

In the main text and previous Appendix, we argued that there exists $\tau_0 = \Theta(1/N)$ such that the HVA with constraints $\theta_{i,j} \geq 0$ and $\sum_{i,j}\theta_{i,j} \leq \tau_0$ does not have vanishing gradients if $|\mathrm{Tr}[\rho_0[H^{(m)}, O]]| \neq 0$

for some m. In this subsection, we provide an example whose gradient component vanishes where the sum of parameters is a constant. This implies that if there is $\tilde{\tau}_0$ such that the gradient is bounded by a constant when $\sum_{ij}\theta_{i,h} \leq \tilde{\tau}_0$, $\tilde{\tau}_0$ must be smaller than this constant.

We consider the Ising model with transverse and longitudinal fields whose Hamiltonian is given by $\mathcal{H} = -\sum_i Z_i Z_{i+1} - h\sum_i X_i - g\sum_i Z_i$. The HVA for this model can be written as

$$|\psi(\{\theta_{i,j}\})\rangle = \prod_{i=p}^{1} e^{-i\theta_{i,3}\sum_i Z_i} e^{-i\theta_{i,2}\sum_i X_i} e^{-i\theta_{i,1}\sum_i Z_i Z_{i+1}} |+\rangle^N \tag{F.44}$$

where the initial state $\rho_0 = |\psi_0\rangle\langle\psi_0|$ with $|\psi_0\rangle = |+\rangle^{\otimes N}$. Consider a local observable $O = Y_1$ and gradient for $\theta_{p,3}$ which is given by

$$\partial_{p,3} C = i\,\mathrm{Tr}\{U_R \rho_0 U_R^\dagger [\sum_i Z_i, Y_1]\} = -2\,\mathrm{Tr}\{U_R \rho_0 U_R^\dagger X_1\}. \tag{F.45}$$

As $|\mathrm{Tr}[\rho_0[\sum_i Z_i, Y_1]]| = 2$, Theorem 1 implies that there is $\tau_0 = \Theta(1/N)$ such that any parameters satisfying $\theta_{i,j} \geq 0$ and $\sum_{i,j}\theta_{i,j} \leq \tau_0$ give an $\Theta(1)$ gradient. We now consider a parameter set with $\theta_{i,2} = \theta_{i,1} = 0$ for all i. This gives $|\psi(\{\theta_{i,j}\})\rangle = U_R|\psi_0\rangle = (\cos\Upsilon|+\rangle - i\sin\Upsilon|-\rangle)^{\otimes N}$ where $\Upsilon := \sum_{i,j}\theta_{i,j} = \sum_i \theta_{i,3}$. Thus for $\Upsilon = \pi/4$, we obtain $\partial_{p,3}C = -2\langle y;+|^{\otimes N} X_1 |y;+\rangle^{\otimes N} = 0$. This implies that we do not expect that the condition of the theorem is relaxed to $\tau_0 \geq \pi/4 = \Theta(1)$.