Gradient scalability and Taylor surrogation of quantum cost landscapes

Sabri Meyer [1,*] Francesco Scala [1] Francesco Tacchino [2] and Aurelien Lucchi [1]
[1]*Department of Mathematics and Computer Science, University of Basel, 4051 Basel, Switzerland*
[2]*IBM Quantum, IBM Research Europe—Zurich, 8803 Rüschlikon, Switzerland*

(Received 19 September 2025; accepted 21 May 2026; published 22 June 2026)

Variational quantum algorithms (VQAs) are promising candidates for near-term quantum computing, yet they face scalability challenges due to barren plateaus, where gradients vanish exponentially in the system size. Recent conjectures suggest that avoiding barren plateaus might inherently lead to classical simulability, thus limiting the opportunities for quantum advantage. In this work, we advance the theoretical understanding of the relationship between gradient scalability at initialization and computational complexity of VQAs. We first present the *Taylor surrogate*, a classical simulation technique matching Pauli path runtime guarantees on near-Clifford regions, with run-time advantages in specific regimes. Leveraging the Taylor surrogate, we prove that beyond the previously established classically simulable regions the computational complexity is at least superpolynomial. Next, we introduce the *linear Clifford encoder* (LCE), a classically efficient ansatz modifier that ensures constant-scaling gradients within landscape regions close to Clifford circuits. Finally, numerical experiments on LCE-modified landscapes provide preliminary empirical evidence of a transition zone where the constant-scaling gradients may start to decay polynomially in superpolynomially complex regions rather than exponentially. These findings suggest speculative instances where nonvanishing gradients and superpolynomial complexity could potentially coexist, vindicating the need for future formal proofs.

DOI: 10.1103/l9bp-dccf

I. INTRODUCTION

Variational quantum algorithms (VQAs) [1] are promising tools for near-term quantum computers with a wide variety of possible applications in quantum chemistry, condensed matter physics, and quantum machine learning (QML) [2–10]. However, the scalability of these algorithms is severely hindered by the phenomenon of barren plateaus [11,12], where gradients vanish exponentially with the number of qubits, thus requiring an exponential number of shots for gradient estimation during optimization. As a result, demonstrating indisputable advantages of VQAs over their classical counterparts remains an open question. A critical challenge arises when attempting to mitigate this issue: It has been conjectured that avoiding barren plateaus may inherently lead to classical simulability and thus limits quantum advantage [13]. Despite considerable efforts in this direction [14–16], it remains unknown whether barren plateaus can be avoided for arbitrary ansätze. Specifically, it is unclear if gradient scalability can be maintained without restricting the parametrized quantum circuit (PQC) architecture or limiting observables to local operators, both of which are often necessary to ensure the system remains classically hard to simulate [13].

This work focuses on studying the interplay between gradient scalability at initialization and the computational complexity of quantum cost functions for arbitrary PQC ansätze. By Taylor expanding the optimization landscape around Clifford circuits, we establish a rigorous mathematical framework to analyze both the cost of classical simulation and the behavior of gradients. Specifically, we propose a generalization of the quadratic Clifford expansion [17] and leverage the Gottesman-Knill theorem [18,19] to efficiently compute higher-order Taylor coefficients. This approach yields what we call the *Taylor surrogate* [20] for classically simulating expectations with near-Clifford PQCs. Our method is capable of surrogating the same regimes as the Pauli path technique [21] while offering provable computational speedups in special cases. Furthermore, we analyze the cost of simulating circuits beyond known classically tractable limits [22–24]. By deriving superpolynomial lower bounds as a function of the model parameters, we offer theoretical insights into the boundaries of classical hardness.

The efficiency of computing derivatives near Clifford points provides a direct path to managing gradient scalability. We introduce the *linear Clifford encoder* (LCE), a classically efficient ansatz modifier that systematically manages linear Taylor coefficients to ensure constant-scaling gradient norms at initialization. This result holds within local "patches" of the landscape near Clifford circuits, even for large-scale models, at the cost of incorporating a small number of Clifford gates into the ansatz.

Our Taylor framework further reveals a phase transition in computational complexity as the size of these surrogate patches increases. This threshold separates regimes of

*Contact author: sabri.meyer@unibas.ch

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

polynomial complexity, where LCE ensures constant-scaling gradients at initialization, from those exhibiting superpolynomial hardness. While we do not prove theoretical guarantees for the absence of barren plateaus beyond this threshold, numerical experiments indicate the existence of a "transition patch" where gradients may begin to decay polynomially rather than exponentially. We present this regime strictly as a numerical indication and an open hypothesis for future rigorous proof. Furthermore, for certain global observables known to suffer from barren plateaus [25], LCE-assisted VQA experiments provide encouraging evidence of significantly enhanced training dynamics with faster convergence. Although the value for practical problems remains a subject for future study, these experiments hint at the possibility of barren-plateau-free landscape regimes that nonetheless possess high computational complexity.

The rest of this document is structured as follows. In Sec. II, we introduce the necessary concepts to lay the foundation for our main theoretical results, presented in Sec. III. There, we mathematically address the challenge of balancing gradient scalability and computational complexity. Section IV discusses numerical experiments, providing evidence of large landscape patches without barren plateaus. Finally, we summarize our findings in Sec. V. A compact overview comparing our theoretical results with related works is found in Appendix A.

II. PRELIMINARIES

In this section, we establish the four key ideas that we will leverage to yield our main results: the stabilizer formalism, the expressive power of PQCs, barren plateau mitigation, and a classical Taylor surrogate.

A. Stabilizer formalism

The stabilizer formalism efficiently describes a large class of quantum states, named *stabilizer states*, by tracking the Pauli operators that leave them invariant [26]. This underlies the Gottesman-Knill theorem [18,19], which guarantees efficient classical simulation for circuits made solely of Clifford operators. Here, we briefly recall the definitions of the Pauli and Clifford groups (see also Refs. [26,27]).

Let $N \geqslant 1$ denote the number of qubits. Here, the N-qubit Pauli group is defined as the unitary subgroup

$$\mathcal{P}_N := \{\pm P, \pm iP : P \in \{I, X, Y, Z\}^{\otimes N}\}, \tag{1}$$

and its elements (modulo global phase) are referred to as (unsigned) Pauli strings. The N-qubit Clifford group is the subgroup of the unitary group $\mathrm{U}(2^N)$ that normalizes $\mathcal{P}_N$ (modulo global phases):

$$\mathrm{Clifford}(N) := \{Q \in \mathrm{U}(2^N) : Q\,\mathcal{P}_N\,Q^\dagger = \mathcal{P}_N\}. \tag{2}$$

It is generated by the ensemble $\{H, S, \mathrm{CX}\}$. The space of stabilizer states is then

$$\mathrm{Stab}(N) := \{Q|0\rangle : Q \in \mathrm{Clifford}(N)\}, \tag{3}$$

where $|0\rangle \equiv |0\rangle^{\otimes N}$. For further details, including a review of related simulation techniques, see Appendixes B and D.

B. Variational quantum algorithms

An ideal N-qubit quantum system resides in a Hilbert space $\mathcal{H}_N$ spanned by $\{|b\rangle : b \in \{0,1\}^N\}$. We focus on a PQC [28] defined by the reversed product of unitaries,

$$U(\theta) = \prod_{k=D}^{1} R_{V_k}(\theta_k)\,W_k := R_{V_D}(\theta_D)W_D \cdots R_{V_1}(\theta_1)W_1, \tag{4}$$

with $D \geqslant 1$ parametrized single-qubit rotation gates and Clifford operators W_k. In particular, D corresponds to the circuit depth. Here, $R_{V_k}(\theta_k) := \exp(-i\frac{\theta_k}{2}V_k)$, with $V_k \in \{X, Y, Z\}$ acting on single qubits, using the notation $V_j := I_{2^{j-1}} \otimes V \otimes I_{2^{N-j}}$. We then define $|\psi(\theta)\rangle := U(\theta)|0\rangle$. Note that for $\theta = 0$, $U(0)$ is Clifford and $|\psi(0)\rangle \in \mathrm{Stab}(N)$. Without loss of generality, we may only consider expanding around $\theta = 0$.

We consider a nonidentity observable

$$H = \sum_{i\in\mathcal{I}} c_i P_i, \tag{5}$$

with Pauli strings $P_i \in \mathcal{P}_N \setminus \{I\}$ and coefficients $c_i \in \mathbb{R}$ satisfying $c_i \in \mathcal{O}(1)$ as $N \to \infty$ with indexing set $\mathcal{I} \subseteq \{I, X, Y, Z\}^{\otimes N} \setminus \{I_{2^N}\}$ uniquely identifying P_i. The expectation value (cost function) is

$$C(\theta) := \langle\psi(\theta)|H|\psi(\theta)\rangle = \sum_{i\in\mathcal{I}} c_i \langle\psi(\theta)|P_i|\psi(\theta)\rangle, \tag{6}$$

which is smooth, 2π periodic in each parameter, and bounded by $\|H\|$. A typical VQA [1] utilizes a quantum processor to estimate the cost $C(\theta)$, which a classical optimizer (such as gradient descent) uses to update the parameters θ, e.g., in variational quantum eigensolver (VQE) [2] tasks that seek the minimizer,

$$\theta_* = \arg\min_{\theta\in[-\pi,\pi]^D} C(\theta). \tag{7}$$

If the ansatz $U(\cdot)$ is sufficiently expressive, then $C(\theta_*)$ provides a good estimate of the lowest eigenvalue of H. The expressivity of a PQC refers to its ability to explore $\mathcal{H}_N$ in an unbiased manner [29]. Notably, expressivity is also closely linked to the so-called barren plateau phenomenon [11,30].

C. Barren plateaus

Here and in the following, we make use of the standard Landau notations to describe asymptotic behavior of functions. Specifically, asymptotically as $D \to \infty$, $\mathcal{O}$ denotes an upper bound, Ω a lower bound, and Θ a tight bound. These notations are formally introduced at the beginning of the Appendix A.

We typically initialize $\theta \sim \mathbb{P}_\theta$, and the choice of the probability distribution $\mathbb{P}_\theta$ is crucial to avoid the barren plateau problem [11,12,31–33]. Barren plateaus occur whenever the expected squared gradient norm vanishes as

$$\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] = \mathcal{O}(2^{-\eta N}) \tag{8}$$

for some $\eta > 0$, while the expected gradient satisfies $\mathbb{E}_\theta[\nabla_\theta C(\theta)] = 0$, making optimization intractable. Intuitively, high expressivity of PQCs is known to be one of the major causes of barren plateaus [29,30], while it is also known to potentially limit classical simulability [21,23,34].

The gradient is sufficient to characterize an essentially flat loss landscape (i.e., no curvature on average): Indeed, by combining Chebychev's inequality with the higher-order parameter-shift rule, the probability of any $\mathcal{O}(1)$-order derivative being at least a certain constant is bounded by the variance of the gradient and thus exponentially suppressed for all constants if Eq. (8) holds true [35].

Various mitigation strategies [16,36–44] exist that establish polynomial lower bounds of the form

$$\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] = \Omega(\mathrm{Poly}(N)^{-1}), \tag{9}$$

but consistently appear to result in classical simulability [13]. Our contribution introduces the LCE technique, which ensures constant-scaling gradients on classically simulable landscape patches. In this work, we assume that $D = \Theta(\mathrm{Poly}(N))$, meaning that the number of parameters D scales polynomially with the number of qubits N. The polynomial is determined by the specifics of the circuit architecture, such as its layer depth.

D. Classical Taylor surrogate patches

For complexity benchmarking, we consider a *classical surrogate* of a quantum learning model [20] that classically reproduces its input-output relations, with high accuracy, efficiency, and low probability of failure. Following ideas from Pauli path methods [21,24,45,46], small-angle techniques [32,33], and quadratic Clifford expansion [17], we approximate $C(\theta)$ using a truncated Taylor expansion of arbitrary order $m-1$ around $\theta = 0$:

$$C(\theta) \approx \sum_{\|\alpha\|_1 < m} (D_\theta^\alpha C)(0)\, \frac{\theta^\alpha}{\alpha!} =: C_m(\theta), \tag{10}$$

where for each multi-index $\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathbb{N}_0^D$ we use the notations $D_\theta^\alpha := \partial_{\theta_1}^{\alpha_1} \cdots \partial_{\theta_D}^{\alpha_D}$, $\theta^\alpha := \theta_1^{\alpha_1} \cdots \theta_D^{\alpha_D}$, $\alpha! := \alpha_1! \cdots \alpha_D!$, and $\|\alpha\|_1 := |\alpha_1| + \cdots + |\alpha_D|$ denoting the ℓ_1 norm of α.

As we will show below, the Taylor coefficients $(D_\theta^\alpha C)(0)$ can be classically simulated by combining higher-order parameter-shift rules [35,47,48] and the Gottesman-Knill theorem [18,19], with a complexity of at most $\mathcal{O}(|\mathcal{I}|\, D^{\|\alpha\|_1})$. Thus, $C_m(\theta)$ serves as an efficient polynomial surrogate of $C(\theta)$ for near-Clifford patches, where parameters are initialized as

$$\mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}, \tag{11}$$

with $\sigma > 0$ determining the patch size.

While small patches limit PQC expressivity (since a Clifford ensemble covers only a discrete subset of $\mathcal{H}_N$), increasing non-Clifford resources eventually render classical simulation intractable for stabilizer-based techniques [21–23].

We argue that other popular classical simulation methods are in general unsuitable for near-Clifford simulation, especially since we allow arbitrarily deep and complex PQCs, which is the reason why we focus on applying Gottesman-Knill in this work: Since Clifford circuits allow an arbitrary amount of entanglement, this generally causes tensor network simulations to be hard due to an exponential growth of the bond dimension in the entanglement entropy [49–51]. Likewise, an arbitrarily deep PQC ansatz generally leads to an exponentially large dynamical Lie algebra, thus causing Lie algebraic simulation techniques to be infeasible [34]. To the best of our knowledge, it is unknown whether there exist more efficient simulation techniques to classically surrogate near-Clifford landscape patches than the Pauli path surrogate [21] or our proposed Taylor surrogate. A brief comparison with the Pauli path method is provided in Appendix A.

Our framework aims to maximize the landscape patch size while maintaining both gradient scalability and classical intractability. An illustration of our approach is summarized in Fig. 1.

III. THEORETICAL RESULTS

This section presents the main theoretical contributions of our work. The central idea is to employ the Gottesman-Knill theorem [18,19] to classically simulate the Taylor coefficients of our Taylor surrogate. These can be used to systematically adapt the Clifford structure of the PQC, ensuring constant gradient norms within the surrogate patch. This technique, termed LCE, can be implemented efficiently using classical resources for large-scale PQCs.

After establishing the LCE transformation, we present complexity-theoretical insights showing that our Taylor surrogate simulation technique can also be leveraged to measure the circuit computational complexity as a function of the initialization strategy. We further identify a critical patch size where run-time complexity transitions from polynomial to superpolynomial, indicating a phase transition in computational complexity. Combined with the LCE theory, we argue that these results could in principle be leveraged to prove the existence of superpolynomially complex landscape patches where gradient norms still decay only polynomially. While these results outline a clear theoretical pathway, we explicitly emphasize that the existence of a transition patch with polynomially decaying gradients remains a conjecture supported by our numerical evidence rather than a mathematically proven guarantee.

A. Backpropagation of high-order Clifford structures in the Heisenberg picture

We begin by outlining the core technical contribution and intuition behind our method. Since our proposed Taylor surrogate expands around Clifford circuits [recalling from Eq. (4) that $U(0)$ is Clifford], we introduce the notion of *Clifford structures*—specific ensembles of Clifford operators that arise from the evaluation of the Taylor coefficients $(D_\theta^\alpha C)(0)$. These structures generate, via Heisenberg backpropagation [52], a set of efficiently computable Pauli strings and can be used to classically benchmark [18,19] optimization landscape patches prior to the deployment of the corresponding quantum algorithm on quantum processors.

As shown in Appendix E, the Taylor coefficients take the explicit parameter-shift form [35,47,48]

$$\left(D_\theta^\alpha C\right)(0) = \frac{1}{2^{\|\alpha\|_1}} \sum_{j \leqslant \alpha} (-1)^{\|j\|_1} \binom{\alpha}{j} C(\xi_{j,\alpha}), \tag{12}$$

where $\xi_{j,\alpha} := \frac{\pi}{2}(\alpha - 2j)$ are equidistant parameter points for fixed multi-indices $\alpha \in \mathbb{N}_0^D$, and $0 \leqslant j \leqslant \alpha$, giving rise to

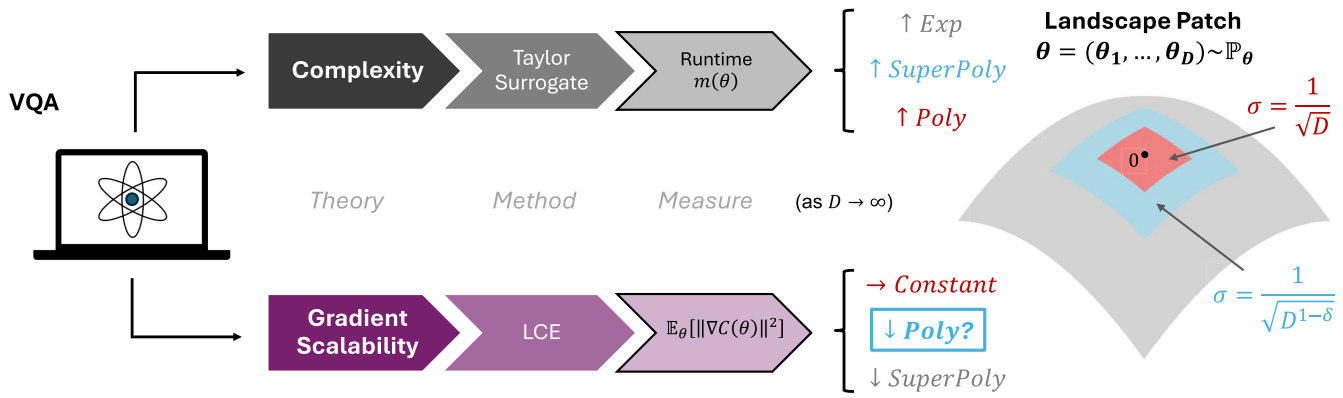


FIG. 1. Illustration of our decomposition of the theoretical link between complexity and gradient scalability in VQAs. These two theories relate to each other with the landscape patch size $\sigma > 0$ around $\theta = 0$. The LCE and the classical Taylor surrogate reveal how gradient scalability and computational complexity interact, exposing a superpolynomially complex transition patch (blue), where for sufficiently small $\delta > 0$, gradients are conjectured to decay polynomially with the model size, a hypothesis currently supported solely by our numerical experiments. The scaling results hold true for observables satisfying $\|H\| = \mathcal{O}(1)$ and $c_{i_0} = \Omega(1)$ for some $i_0 \in \mathcal{I}$.

Clifford configurations exclusively. Indeed, for each shift $j \leqslant \alpha$, we establish in Appendix J that evaluating the PQC in Eq. (4) at $\xi_{j,\alpha}$ results in a Clifford operator $U(\xi_{j,\alpha})$, enabling Gottesman-Knill-based simulation [18,19]. This leads to an algorithm evaluating Eq. (12) in polynomial time $\mathcal{O}(|\mathcal{I}|\, D^{\|\alpha\|_1})$ if $\|\alpha\|_1 = \mathcal{O}(1)$, motivating the following.

Definition 1 (High-order Clifford structures). For a variational ansatz $U(\theta)$ composed of Clifford gates and single-qubit Pauli rotations with $D \geqslant 1$ parameters, the Clifford structure up to order $m-1$ is the collection defined as

$$\{U(\xi_{j,\alpha}) : j \leqslant \alpha,\ \|\alpha\|_1 < m\},$$

where $\xi_{j,\alpha} := \frac{\pi}{2}(\alpha - 2j)$, and each $U(\xi_{j,\alpha})$ is a Clifford unitary transformation.

Since we expand around $\theta = 0$, $U(0)$ always belongs to the Clifford structure. For instance, the *linear Clifford structure* derived from the standard parameter-shift rule [47,48] is given by the Clifford ensemble $\{U(0), U(\pm\frac{\pi}{2}e_k) : k \in [D]\}$, with unit vectors e_k. The linear Clifford structure is used to build the LCE in the next section.

Due to the transitive action of the Clifford group on $\mathcal{P}_N$, we obtain for each $j \leqslant \alpha$:

$$\begin{aligned} C(\xi_{j,\alpha}) &= \sum_{i\in\mathcal{I}} c_i \langle 0|U(\xi_{j,\alpha})^\dagger P_i U(\xi_{j,\alpha})|0\rangle \\ &= \sum_{i\in\mathcal{I}} c_i \langle 0|P_i^{j,\alpha}|0\rangle, \end{aligned} \tag{13}$$

where $P_i^{j,\alpha} := U(\xi_{j,\alpha})^\dagger P_i U(\xi_{j,\alpha}) \in \mathcal{P}_N \setminus \{I\}$ defines a list of Heisenberg-evolved Pauli strings. These evolved strings are efficiently computable via the Aaronson-Gottesman check matrix algorithm [19], producing Pauli strings [26] indexed by $i \in \mathcal{I}$ [cf. Eq. (5)]. In Appendix E, we also provide three different visual representations of the Clifford structure as per Definition 1 in order to help the intuition of the reader.

To analyze the influence of $U(\cdot)$ on the optimization landscape, we combine Eqs. (12) and (13):

$$\begin{aligned} \left(D_\theta^\alpha C\right)(0) &= \frac{1}{2^{\|\alpha\|_1}} \sum_{j\leqslant\alpha} (-1)^{\|j\|_1} \binom{\alpha}{j} C(\xi_{j,\alpha}) \\ &= \frac{1}{2^{\|\alpha\|_1}} \sum_{j\leqslant\alpha} (-1)^{\|j\|_1} \binom{\alpha}{j} \sum_{i\in\mathcal{I}} c_i \langle 0|P_i^{j,\alpha}|0\rangle. \end{aligned} \tag{14}$$

Thus, each Taylor coefficient depends on Pauli coefficients $\{c_i\}$ and Heisenberg-evolved Pauli strings $P_i^{j,\alpha}$ generated by the Clifford structure. Together with prefactors $(-1)^{\|j\|_1}\binom{\alpha}{j}$, manipulating the set of Pauli strings

$$\left\{P_i^{j,\alpha} : j \leqslant \alpha,\ \|\alpha\|_1 < m,\ i \in \mathcal{I}\right\}, \tag{15}$$

e.g., by adding or removing Clifford gates in $U(\cdot)$, allows one to influence summation cancellations and contributions in Eq. (14), for example, to ensure a nonzero gradient component $(\partial_{\theta_k} C)(0) \neq 0$. Since Eq. (12) can be evaluated using Gottesman-Knill [18,19], this establishes an efficient classical framework for engineering the Clifford architecture of $U(\cdot)$ in order to achieve better gradients at $\theta = 0$.

Figure 2(b) illustrates this framework: Each Taylor coefficient $(D_\theta^\alpha C)(0)$ induces an α-order Clifford structure that backpropagates through the observable [52], producing a list of Pauli strings $\{P_i^{j,\alpha}\}_{i,j}$, which encode the α-order contribution to the surrogate patch.

B. The linear Clifford encoder

Having introduced the mathematical framework, this prompts a central question: Can we systematically engineer the Clifford architecture of $U(\cdot)$ to enhance landscape properties (e.g., gradient scalability) without sacrificing expressivity of the ansatz? In the following, we demonstrate that this is, in principle, achievable. We now propose a technique that leverages the use of first-order Clifford structures (cf. Definition 1) to ensure trainability on near-Clifford patches. Proofs are deferred to Appendix F.

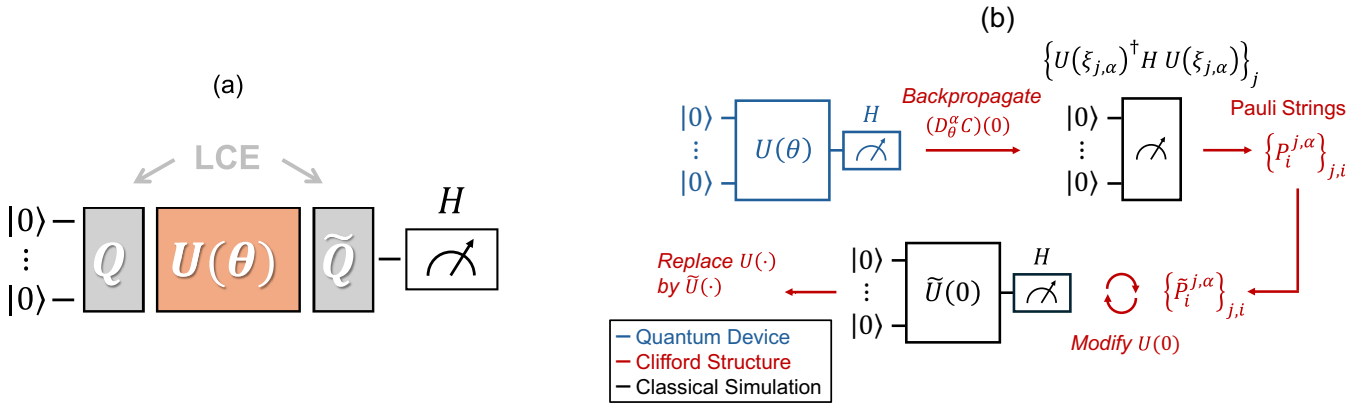


FIG. 2. (a) An illustration of the pair $(Q, \tilde{Q})$ of Clifford operators used in the LCE construction. The Clifford Q translates the initial state, while $\tilde{Q}$ performs a change of basis on H. (b) Classical framework for engineering Clifford structures.

The previously discussed Taylor surrogation framework allows the explicit construction of a pair of Clifford operators $(Q, \tilde{Q})$ defining the transformation $\tilde{U}(\cdot) := \tilde{Q}U(\cdot)Q$ and yielding an *LCE-transformed cost* $\tilde{C}(\cdot)$, where LCE stands for *linear Clifford encoder*, as shown in Fig. 2(a). Even though this transformation inherently changes the Clifford architecture of the PQC ansatz, we prove in Appendix H that LCE maintains the minimizer in Eq. (7) (possibly with another parameter configuration) given that the PQC is sufficiently expressive.

In brief, the LCE procedure consists of the following three key steps:

(1) Choose an arbitrary parameter coordinate $k \in [D]$ and a Pauli decomposition index $i_0 \in \mathcal{I}$ with $c_{i_0} = \Omega(1)$.

(2) Construct the first LCE Clifford $\tilde{Q}$ that induces a relative phase of ± 1 between the two Pauli expectation in the parameter shift in order to avoid cancellation. This can be viewed as a change of basis of the observable $H \mapsto \tilde{Q}^\dagger H \tilde{Q}$ in the Heisenberg picture.

(3) Construct the second LCE Clifford Q that diagonalizes the previously $\tilde{Q}$-backpropagated Pauli term $\tilde{P}_{i_0} \mapsto Q^\dagger \tilde{P}_{i_0} Q \in \langle Z \rangle_N$. This can be viewed as a change of the initial state $|0\rangle \mapsto Q|0\rangle$ in the Schrödinger picture.

As a matter of fact, the two Clifford operators of LCE are used to explicitly control arbitrary linear Taylor coefficient $(\partial_{\theta_k} C)(0)$ (i.e., $\alpha = e_k$ for some unit vector e_k). Hence, given an arbitrary $k \in [D]$, our first main result shows how to use the first-order Clifford structure to transform any cost $C(\cdot) \mapsto \tilde{C}(\cdot)$ such that the gradient component satisfies $(\partial_{\theta_k} \tilde{C})(0) = \Omega(1)$, independently from D or N.

The following theorem employs LCE to control the summation in Eq. (14) and isolate a fixed coefficient c_{i_0} of the observable H. Observe that the operators $(Q, \tilde{Q})$ depend on the choice of k and i_0.

Theorem 1 (Informal). Consider an arbitrary observable H indexed by $\mathcal{I}$ in Eq. (5). Then, for every $k \in [D]$ and $i_0 \in \mathcal{I}$, there exists a pair of Clifford operators $(Q, \tilde{Q})$ such that the gradient component $(\partial_{\theta_k} \tilde{C})(0)$ is equal to

$$\beta_{i_0}(H) := \pm c_{i_0} + \frac{1}{2} \sum_{i \in \mathcal{I} \setminus \{i_0\}} c_i \big\{ \langle 0 | \tilde{P}_i^{0, e_k} | 0 \rangle - \langle 0 | \tilde{P}_i^{e_k, e_k} | 0 \rangle \big\},$$

defining $\tilde{P}_i^{j,\alpha} := Q^\dagger U(\xi_{j,\alpha})^\dagger \tilde{Q}^\dagger P_i \tilde{Q} U(\xi_{j,\alpha}) Q$ and $\xi_{j,\alpha} := \frac{\pi}{2}(\alpha - 2j)$ for $0 \leqslant j \leqslant \alpha$.

Note that the choice of the coefficient c_{i_0} is independent of k, thus allowing a total of $D \cdot \#\{i_0 \in \mathcal{I} : c_{i_0} = \Omega(1)\}$ possible candidates for the LCE construction. The proof of the theorem (cf. Appendix F) shows how to construct $(Q, \tilde{Q})$ explicitly using only a few Clifford gates. Explicit examples illustrating how LCE is used to modify the PQC are provided in Appendix C. The key idea is to choose the first Clifford $\tilde{Q}$ such that it ensures a relative phase between the difference of the Pauli expectations in the parameter shift in order to avoid them from canceling to zero if they are both nonzero. Next, one constructs the second Clifford Q such that both Pauli strings in the parameter shift are diagonal giving rise to nonzero Pauli expectations ± 1. The full gradient is then evaluated by classically computing the expectations of the transformed Pauli strings by using the stabilizer formalism as shown in Appendix B. Furthermore, multiple gradient components may simultaneously benefit from the LCE, provably up to $\frac{D}{2}$ many components for structured PQCs, as discussed in Appendix F.

Importantly, it is possible to explicitly construct the LCE such that $\beta_{i_0}(H) = 1$ if $H = P$ is a single Pauli string (cf. Theorem F1). For general observables H, understanding the scaling behavior of $\beta_{i_0}(H)$ is more subtle due to potential cancellations among the Pauli terms. The lemma below shows that the probability of complete cancellation, which is $\beta_{i_0}(H) = 0$, decreases exponentially with the number of qubits N.

Lemma 1. Let $H = \sum_{i \in \mathcal{I}} c_i P_i$ be a random observable with $P_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{P}_N \setminus \{I\})$. Then, for each $i_0 \in \mathcal{I}$, we have the conditional probability that

$$\mathbb{P}\big(\beta_{i_0}(H)^2 = c_{i_0}^2 \big| P_{i_0}\big) \geqslant 1 - \mathcal{O}(|\mathcal{I}|\, 2^{-N}),$$

as $N \to \infty$.

Intuitively, LCE is applied to isolate the Pauli coefficient c_{i_0}, while the remaining terms in $\beta_{i_0}(H)$ vanish with exponentially high probability. We condition the probability on P_{i_0} since this Pauli string is fixed for the LCE construction as shown in the proof of Theorem 1. If $c_{i_0} = \Omega(1)$, then this almost surely guarantees a constant gradient lower bound, as long as $|\mathcal{I}| = \mathcal{O}(\text{Poly}(N))$. Numerical experiments in Sec. IV (cf. Fig. 4) and Appendix K demonstrate that the

probability of complete cancellation is indeed negligible, and the expected behavior closely matches the case $H = P$ where $\beta_{i_0}(H) = 1$, even if we additionally assume independent and identically distributed (i.i.d.) coefficients $c_i \sim \mathrm{Unif}([-1, 1])$.

C. Escaping barren plateaus with the linear Clifford encoder

We now combine the preceding results to show that the Clifford operators $(Q, \tilde{Q})$ of LCE guarantee constant-scaling gradients in a neighborhood of the origin. Specifically, when combined with the Taylor surrogate in Eq. (10), Theorem 1 yields a constant lower bound in Eq. (9) for arbitrary Pauli decompositions H that contain at least one coefficient with constant scaling, as specified in Eq. (5).

Theorem 2 (Informal). Let H be an arbitrary, uniformly randomly sampled observable as in Eq. (5) with Pauli terms indexed by $\mathcal{I}$. Let $\delta > 0$ and initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components and $\sigma = \mathcal{O}(\|H\|^{-1} D^{-\frac{1+\delta}{2}})$. Then, for each $k \in [D]$ and $i_0 \in \mathcal{I}$ with $c_{i_0} = \Omega(1)$, there exists a pair of Clifford operators $(Q, \tilde{Q})$ such that, with probability at least $1 - \mathcal{O}(|\mathcal{I}|\, 2^{-N})$ (conditioned on P_{i_0}), as $D \to \infty$,

$$\mathbb{E}_{\theta \sim \mathbb{P}_\theta}\left[\left(\partial_{\theta_k} \tilde{C}(\theta)\right)^2\right] = \Omega(1),$$

where $\tilde{C}(\cdot)$ denotes the LCE-transformed cost.

Theorem 2 shows how to apply LCE (cf. Theorem 1) in order to achieve a constant lower bound on the expected gradient square norm, independent of the circuit depth of $U(\theta)$. This is achieved by initializing the parameters on a sufficiently small patch, with a size σ that depends on the number of parameters D and the operator norm $\|H\|$ of the observable. In the case, where $\|H\| = \mathcal{O}(1)$, the effective patch size σ for escaping barren plateaus thus scales as $\mathcal{O}(D^{-1/2})$, which is consistent with the findings of previous works [21,33,44,53]. Otherwise, if $\|H\|$ depends on D (or N), then the patch size σ scales in proportion to $\|H\|^{-1}$ as shown in Theorem 2. That is, our analysis only assumes that

$$c_i = \Omega(1), \quad \text{for some} \quad i \in \mathcal{I}, \tag{16}$$

and we additionally assume that $\|H\| = \mathcal{O}(1)$ only to make our results (particularly the patch sizes) comparable with Refs. [21,33,44,53], even though our results hold true for arbitrary $\|H\|$. Furthermore, note that rescaling the observable [such that $\|H\| = \mathcal{O}(1)$] does not affect its commutative structure and only rescales the energy levels and the energy gaps accordingly. In Appendix J 1, we show that Eq. (16) contains a wide range of Hamiltonians, including cases for which no efficient (probabilistic) classical algorithm exists that can solve the underlying Hamiltonian ground energy decision problem [54], unless BPP = QMA. When combined with Theorem 2, this particularly implies that any nonclassical observable of the form in Eq. (16) remains compatible with LCE to explore all directions with $\Omega(1)$ initial descent magnitudes within $\mathcal{O}(D^{-\frac{1}{2}})$ patches. Crucially, this remains effective for arbitrary model sizes $D, N \to \infty$ of the parametrized circuit $U(\theta)$.

Unlike previous studies limited to specific circuit geometries and local operators [14], our proposed framework works for arbitrary ansätze. It remains valid for any observable with a constant-scaling Pauli term, providing a more universal theoretical foundation. Related work by Park *et al.* [16] achieves constant scaling gradients as well, but on smaller patches, and their analysis relies on stronger assumptions on the observable. A summary that compares additional barren-plateau mitigation methods on small landscape regions with the LCE technique is provided in Table II.

TABLE I. Comparing the computational complexity of the Taylor surrogate and the Pauli path surrogate of Lerch *et al.* [21] as a function of the measure $\|\theta\|_1$ of non-Cliffordness, with provable speedups on sufficiently small patches. The bold entries highlight the time advantage of the Taylor surrogate.

Norm $\|\theta\|_1$	Times $t_{\mathrm{Taylor}}, t_{\mathrm{Pauli}}$
$\Theta(1)$	$\mathcal{O}(\mathrm{Poly}(D)) : \mathbf{t_{Taylor} < t_{Pauli}}$
$qD^{1/2}$ for some $q \in (0, 1)$	$\mathcal{O}(\mathrm{Poly}(D)) : \mathbf{t_{Taylor} < t_{Pauli}}$
$cD^{1/2}$ for all $c > q$	$\mathcal{O}(\mathrm{Poly}(D)) : t_{\mathrm{Taylor}} > t_{\mathrm{Pauli}}$
$\Theta(D)$	$t_{\mathrm{Taylor}} = \mathcal{O}(2^{2D}), t_{\mathrm{Pauli}} = \mathcal{O}(2^D)$

Since the solution θ_* is *a priori* unknown, a reasonable initialization strategy is to randomly sample $k \in [D]$ and $i \in \mathcal{I}$ for the LCE construction within $\mathbb{P}_\theta$ patches. Inside $\mathcal{O}(D^{-\frac{1}{2}})$ patches specifically, the total number of effective LCE samples is equal to $D\#\{i_0 \in \mathcal{I} : c_{i_0} = \Omega(1)\}$; one for each direction $k \in [D]$, and one for each Pauli coefficient $c_{i_0} = \Omega(1)$ isolated from the remaining $i \in \mathcal{I}$. In general, the descent magnitude of such samples may depend on the specific choice of k and i_0 before training.

Importantly, however, while a random walk lacks directional guidance, LCE uses structured gradient information to drive efficient optimization from the start. In fact, note that for any given $k \in [D]$ and $i_0 \in \mathcal{I}$, LCE has the effect of globally reshaping the loss landscape itself (via a translation and change of basis as elaborated in the next paragraph) so that the gradient at initialization captures a stronger, more informative descent signal. More intuitively, LCE ensures parameter initialization at a point with a steep slope along that direction. This effect of LCE on the loss landscape is visually demonstrated in Appendix F.

Crucially, we must ensure that the LCE transformation $C(\cdot) \mapsto \tilde{C}(\cdot)$ does not compromise expressivity of the PQC and does not significantly change the value of the minimum $\tilde{C}(\theta_*)$ [while the minimizer in Eq. (7) may change]. Heuristically, this is plausible: The operator Q merely changes the initial stabilizer state $|0\rangle \mapsto |\psi_0\rangle = Q|0\rangle$, corresponding to a landscape translation, while $\tilde{Q}^\dagger H \tilde{Q}$ constitutes a basis change, mapping H to another Pauli string if $H \in \mathcal{P}_N$. More formally, we prove in Appendix H that both expressivity and global optimality are preserved under LCE transformations via a geometric notion of *ε-state expressivity*. This implies that the LCE-transformed PQC ansatz is able to explore the same regions as the original Hilbert space $\{|\psi(\theta)\rangle : \theta \in [-\pi, \pi]^D\}$, even though the optimal point θ_* might be different.

While LCE preserves the global minima, the risk of getting trapped in local minima remains. In such cases, quantum optimal control [55] and the theory of overparametrization [56] may help transform local minima into global ones [57]. We leave the study of this topic to future work.

Having introduced the LCE technique, which allows us to circumvent barren plateaus on $\mathcal{O}(D^{-1/2})$-sized surrogate

TABLE II. Comparison of barren-plateau mitigation on small patches. Models within the $\mathcal{O}(D^{-1/2})$ patch are classically simulable (cf. Theorem 4), while only our method provides numerical experiments with evidence of a larger landscape patch (cf. Figs. 5 and 6) where no classical surrogate is currently known to exist. To simplify the comparison, we may assume that $\|H\| = \mathcal{O}(1)$.

	Gradient bounds within $\mathcal{O}(D^{-\frac{1}{2}})$ patches	Gradient bounds beyond $\mathcal{O}(D^{-\frac{1}{2}})$ patches	Model assumptions
Linear Clifford encoder	$\Omega(1)$ on $\mathcal{O}(\|H\|^{-1}D^{-\frac{1}{2}+\delta})$	$\Omega(\mathrm{Poly}(N)^{-1})$?	LCE-transformed expectation $C(\theta) = \mathrm{Tr}\,(U(\theta)^\dagger\|0\rangle\langle 0\|U(\theta)H)$ with arbitrary PQC $U(\theta)$, and any observable H with $c_i = \Omega(1)$ for some $i \in \mathcal{I}$.
Small angles [33]	$\Omega(\mathrm{Poly}(N)^{-1})$ on $\mathcal{O}(\|H\|^{-1}D^{-\frac{1}{2}})$	None	Expectation value of the form $\mathrm{Tr}\,(U(\theta)^\dagger\|0\rangle\langle 0\|U(\theta)H)$, and local observable H.
Park *et al.* [16]	$\exists\gamma > 0:\quad \Omega(1)$ on γD^{-1}	None	Expectation cost function $\mathrm{Tr}\,(U(\theta)^\dagger\|\psi_0\rangle\langle\psi_0\|U(\theta)H)$ with $\|\psi_0\rangle$ such that $\|(\nabla_\theta C)(0)\| = \Omega(1)$, imposing single Pauli or local observable H.
Warm start [53]	$\Omega(\mathrm{Poly}(N)^{-1})$ on $\mathcal{O}(D^{-\frac{1}{2}})$	None	Warm-started cost function $1 - \|\langle\psi_0\|U(\theta)^\dagger e^{-i\delta t H}U(\theta^*)\|\psi_0\rangle\|^2$ with $\delta t = \Theta(\|H\|^{-1})$ for an observable H, and arbitrary PQC $U(\theta)$.

patches, we now ask whether larger patch sizes are achievable. The next section explores how computational complexity scales with increasing parameter norms $\|\theta\|_1$, providing a measure of non-Cliffordness, relating patch sizes to classical simulability, and revealing a phase transition in computational complexity.

D. Phase transitions in computational complexity

Let us now consider again the task of approximating a target PQC with the Taylor surrogate $C_m(\theta)$ from Eq. (10), with truncation threshold $m = m(\theta)$. The truncation threshold directly determines its computational complexity, as m directly determines the number of terms in the summation in Eq. (10) (cf. Lemma I4). We analyze the minimal threshold required to satisfy a surrogate error tolerance via two approaches:

(1) *Worst-case error analysis*: $m = m(\theta)$ is deterministic for a fixed configuration $\theta \in [-\pi, \pi]^D$.

(2) *Mean-squared error analysis*: $m = m(\mathbb{P}_\theta)$ is probabilistic, depending on an initialization strategy $\mathbb{P}_\theta$.

This separation, previously explored by Lerch *et al.* [21], reveals distinct trade-offs in simulability and gradient scalability. We show in Appendix I that the worst-case analysis characterizes computational complexity for arbitrary patch sizes, while mean-squared error is limited to $\mathcal{O}(D^{-1/2})$-sized patches. Notably, this threshold marks the boundary between polynomial and superpolynomial regimes (cf. Appendix J) and identifies the largest known patch where barren plateaus can be provably mitigated [21,33,44,53] (cf. Theorem 2).

1. Worst-case error analysis

In Appendix I, we show that the worst-case error obeys

$$|C(\theta) - C_m(\theta)| \leqslant \frac{\|H\|}{m!}\|\theta\|_1^m. \tag{17}$$

Thus, for any fixed $\theta \in [-\pi, \pi]^D$, one can determine the minimal $m = m(\theta)$ needed to satisfy the error bound in Eq. (17) as a function of $\|\theta\|_1$, which in this context serves as a way to measure the level of non-Cliffordness of the configuration. For the Taylor surrogate, this leads to the following ℓ_1 characterization of a superpolynomial computational phase transition (cf. Appendix J).

Theorem 3 (Informal). If $\|\theta\|_1 = \mathcal{O}(1)$, the run-time of evaluating $C_m(\theta)$ is at most polynomial, namely, $\mathcal{O}(|\mathcal{I}|D^m)$. If $\|\theta\|_1 = o(D)$ strictly increases sublinearly as $D \to \infty$, and $\#\{k \in [D] : \exists\xi \in \frac{\pi}{2}\mathbb{Z} : |\theta_k - \xi| > \Omega(D^{-1})\} = \Omega(D)$, then the run-time becomes at least superpolynomial,

$$\mathrm{Time}\, C_m(\theta) = \Omega\left(\mathrm{Poly}(D)^{-1}\left(\frac{D}{\|\theta\|_1}\right)^{e\|\theta\|_1}\right),$$

and is at most $\mathcal{O}(|\mathcal{I}|D\,2^{2D})$ in the worst case.

Appendix D 4 (cf. Fig. 9) demonstrates how to compare the polynomial complexity across various PQC models and how to identify computational overheads.

For $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components and $\sigma = D^{-r}$ for $r \in [0, 1)$, the law of large numbers yields

$$\mathbb{E}_\theta[m(\theta)] = \Theta(D^{1-r}) \tag{18}$$

as $D \to \infty$. Thus, the Taylor surrogate captures computational complexity across all patch sizes $\mathcal{O}(D^{-r})$ parametrized by $r \geqslant 0$. Combining the lower bound $m(\theta) = \Omega(\|\theta\|_1)$ as $D \to \infty$ (cf. Lemma I3) with Theorem 3 thus reveals a *phase transition in computational complexity* at $r = 1$, separating polynomial from superpolynomial regimes. As $r \to 0$, the run-time saturates to exponential scaling.

The complexity of computing the Taylor surrogate $C_m(\theta)$ is essentially determined by the number of Taylor coefficients appearing in the truncated series in Eq. (10). However, the *effective* complexity of $C_m(\theta)$ is generally smaller since many of the Taylor coefficients could vanish. In practice, it is *a priori* unknown which coefficients vanish, meaning that all of them have to be computed in any case, thus ruling out

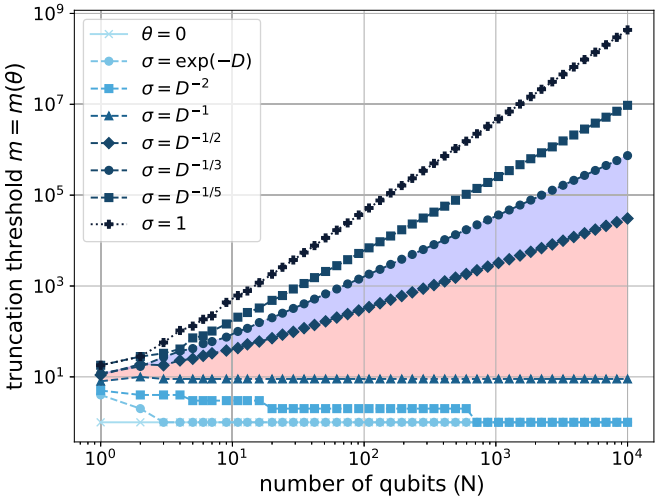


FIG. 3. Worst-case analysis truncation threshold $m = m(\theta)$ of a model with $D = 2N(N+1)$ parameters, as predicted by the expected truncation threshold in Eq. (18). Here, we set $\|H\| = 1$ and $\varepsilon = 10^{-6}$. The red region is easy to simulate classically (cf. Theorem 4), while the blue region is not (cf. Theorem 3).

the possibility of exploiting this fact for obtaining a more powerful surrogate.

2. *Mean-squared error anaylsis*

For $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components, the mean-squared error satisfies

$$\mathbb{E}_\theta[(C(\theta) - C_m(\theta))^2] \leqslant c\,\|H\|^2 \frac{(D\sigma^2)^m}{m^m}, \tag{19}$$

for some constant $c > 0$ if $m = \mathcal{O}(1)$ (cf. Appendix I). This implies that σ cannot exceed $\Omega(D^{-1/2})$ without incurring significant error. Thus, mean-squared analysis fails beyond $\mathcal{O}(D^{-1/2})$ patches, while worst-case analysis remains valid across the full parameter space.

Figure 3 illustrates the scaling of $m(\theta)$ when simulating a model with $D = 2N(N+1)$ parameters, such as the linear-depth `mHEA` (minimalistic Hardware Efficient Ansatz [58,59]) defined in Appendix C, with patch sizes $\mathcal{O}(\sigma)$ under various distributions $\mathbb{P}_\theta$. The red-shaded region corresponds to the regime where barren plateaus are provably avoided (cf. Theorem 2). The blue-shaded region marks the regime requiring at least superpolynomial computational resources for the Taylor surrogate. As further shown in Theorem 4, the mean-squared error analysis collapses the red-shaded region to constant scaling $m(\mathbb{P}_\theta) = \mathcal{O}(1)$, implying classical simulability of $\mathcal{O}(D^{-1/2})$ patches—consistent with the findings of Lerch *et al.* [21].

To ensure a small mean-squared error, we distinguish between two initialization strategies:

(1) *Asymptotic local initialization:*

For $\sigma = \mathcal{O}(\|H\|^{-1/m} D^{-(1+\delta)/2})$ with $\delta > 0$, the error decays as $\mathcal{O}(D^{-\delta m})$.

(2) *Nonasymptotic local initialization:*

For $\sigma = \mathcal{O}(q\|H\|^{-1/m} D^{-1/2})$ with $0 < q < 1$, the error decays as $\mathcal{O}(q^{2m})$.

The first initialization strategy mitigates the problem of barren plateaus on small landscape patches (cf. Theorem 2), while the second facilitates run-time comparison with state-of-the-art classical simulation methods [21]. Both yield polynomial-time surrogates since the first initialization strategy results in a smaller patch than the second.

Theorem 4 (Informal). Let $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components, according to Eq. (19) with constant $c > 0$. Assume further that $\|H\| = \mathcal{O}(1)$. Then, with probability at least $1 - \rho$, the computational cost of simulating $C(\theta)$ via $C_m(\theta)$ up to an error $\varepsilon > 0$, is of polynomial order,

$$\mathcal{O}\left(|\mathcal{I}|\, D^{\frac{1}{2\ln(\frac{1}{q})}\ln(\frac{c}{\rho\varepsilon^2})}\right)$$

for nonasymptotic local initialization, i.e., $\sigma = \mathcal{O}(qD^{-1/2})$ for some $0 < q < 1$.

For general $\|H\|$, Theorem 4 establishes a classical surrogate for all patches $\sigma = \mathcal{O}(\|H\|^{-1} D^{-1/2})$. In the special case where $\|H\| = \Theta(1)$ is normalized, this aligns with the main results of Lerch *et al.* [21], who study a classical surrogate based on the Pauli path formalism [24]. Notably, as shown in Appendix J (cf. Corollary J1), our Taylor surrogate outperforms such a method on sufficiently small surrogate patches (i.e., $0.35 < q < 1$) under nonasymptotic local initialization. A tabular comparison of the two surrogate complexities in terms of $\|\theta\|_1$ [which characterizes the patch size via $\sigma = \mathcal{O}(D^{-r})$ as per Eq. (18)] is shown in Table I. The Taylor surrogate outperforms the Pauli surrogate on sufficiently small patches $\sigma = qD^{-1/2}$ as per Corollary J1, whereas the Pauli path technique is preferable beyond these patches. While Ref. [21] does not contain a complexity result as a continuous function of $\|\theta\|_1$ and instead characterizes complexity only at $r \in \{0, \frac{1}{2}, 1\}$ for $\theta \sim \mathrm{Unif}([-\sigma, \sigma]^D)$ with $\sigma = \mathcal{O}(D^{-r})$, our analysis extends to the entire spectrum $r \geqslant 0$ via Eq. (18), and additionally includes superpolynomial complexity lower bounds beyond the classical patch (cf. Theorem 3). Table III provides a more detailed comparison with the Pauli path surrogate [21] of small landscape patches.

IV. NUMERICAL EXPERIMENTS

We now present numerical experiments supporting our theoretical results. First, we benchmark the gradient scalability of LCE-transformed PQCs using the Taylor surrogate to confirm that proper Clifford design yields constant-scaling gradients (cf. Theorem 2). We then extend these insights to LCE-assisted VQE [2,4] optimization, leveraging improved gradients on near-Clifford patches.

A. Initial gradient scalability

We evaluate the scaling of the gradient norm $\|(\nabla_\theta C)(0)\|$ for increasing qubit numbers N focusing on the `mHEA` [58] model—a minimalistic Hardware-Efficient Ansatz based on Qiskit's `EfficientSU2` [59], comprising Y/Z rotations and circular entanglement. Appendix C shows additional PQC models with their respective LCE modifications. Note that Theorem 2 applies in general even to deep PQCs.

TABLE III. Comparing classical surrogates of small landscape patches.

	Worst-case simulation of $\mathcal{O}(D^{-1})$ patches	Mean squared error (MSE) simulation of $\mathcal{O}(D^{-1/2})$ patches	Model assumptions
Taylor surrogate	$\mathcal{O}(\lvert\mathcal{I}\rvert D^{m(\theta,\varepsilon)})$	$\mathcal{O}\left(\lvert\mathcal{I}\rvert\, D^{\frac{1}{2\log\left(\frac{1}{q}\right)}\log\left(\frac{c}{\rho\varepsilon^2}\right)}\right)$	Expectation $\mathrm{Tr}\,(U(\theta)^\dagger\lvert 0\rangle\langle 0\rvert U(\theta)H)$ with arbitrary PQC $U(\theta)$, generated by Clifford and single-qubit Pauli-rotation gates.
Pauli path surrogate [21]	$\mathcal{O}\left(\lvert\mathcal{I}\rvert D^{\log\left(\frac{1}{\varepsilon}\right)}\right)$	$\mathcal{O}\left(\lvert\mathcal{I}\rvert D^{\log\left(\frac{1}{\rho\varepsilon^2}\right)}\right)$	Expectation $\mathrm{Tr}(\mathcal{E}_\theta(\rho)H)$ with density matrix ρ, and parametrized channel $\mathcal{E}_\theta$, imposing classically feasible derivatives $D^\alpha_\theta C(\theta)$ with certain bounds.

Figure 4 compares initial gradient norms at $\theta = 0$ with and without LCE, employing `mHEA` with linear circuit depth $L = N$. Gradients are computed analytically via Eq. (14). Each run uses a randomly sampled observable H, and a random LCE construction. Panel (a) of Fig. 4 focuses on uniformly random single Pauli strings $H = P$. Without LCE, constructive contributions in Eq. (14) vanish with increasing N, leading to zero expected gradient for sufficiently large N. With LCE, gradient norms remain constant in N. Panels (b) and (c) of Fig. 4 extend results to random observables with $|\mathcal{I}| = \lfloor\sqrt{N}\rfloor$ Pauli terms. For observables with uniformly sampled Pauli strings P_i and $c_i = 1$ for all $i \in \mathcal{I}$ (called random unweighted H), the scaling compares to the single Pauli case. For observables with uniformly sampled Pauli strings P_i and i.i.d. $c_i \sim \mathrm{Unif}([-1, 1])$ for all $i \in \mathcal{I}$ (called random weighted H), gradient norms tend to behave more randomly due to random Pauli coefficients. This confirms that the LCE transformation remains effective for arbitrary observables. As previously mentioned, evaluation at $\theta = 0$ is classically efficient by Gottesman-Knill [18,19], allowing simulation even for large N. Specifically, evaluating the gradient norm $\|(\nabla_\theta C)(0)\|$ takes $\mathcal{O}(|\mathcal{I}|D^2)$ operations: The vanilla parameter-shift rule [47,48] requires $\mathcal{O}(|\mathcal{I}|D)$ operations to evaluate $(\partial_{\theta_k} C)(0)$ (cf. Lemma E2), of which there are exactly D distinct gradient components $k \in [D]$. Appendix K contains gradient norm scaling experiments with up to $N = 32$ qubits for `mHEA` with $L = 1$ layer.

Figure 5 repeats the experiment of the left panel in Fig. 4 on near-Clifford patches, varying patch sizes via Gaussian initialization $\sigma = D^{-r}$. Each run uses a randomly sampled parameter, a random single Pauli string observable $H = P$, and a random LCE construction. The red-shaded region corresponds to classically efficient regimes (cf. Fig. 3); blue-shaded region to superpolynomial regimes as per Theorem 3, where gradients may decay polynomially in a transition zone just beyond the red regime. Within the constant-scaling regime (red-shaded region), LCE ensures gradient scalability independent of the model size. For too large σ, gradients decay exponentially. This empirically suggests the existence of a transition patch (contained in the blue-shaded region) where gradients asymptotics smoothly transition from constant to polynomial decay [cf. Eq. (9)].

Interestingly, the improved gradient scaling behavior due to LCE closely resembles that shown in Fig. 3, suggesting a deeper link between gradient scalability and complexity. Note that the non-LCE trends appearing noisier are likely an artefact of the log scale on fluctuations of smaller orders of magnitude. We refer to Appendix K for comparisons across additional hardware-efficient PQC models.

B. Variational quantum eigensolvers

We now perform an ablation study on VQE optimization dynamics. While the theory provided in the previous sections focuses solely on initialization, here we provide experiments investigating the effects of LCE during optimization. When LCE is applied at initialization, we select a random parameter $k \in [D]$ and a random Pauli index $i_0 \in \mathcal{I}$ with $c_{i_0} = \Omega(1)$ to construct a single LCE transformation (as elab-

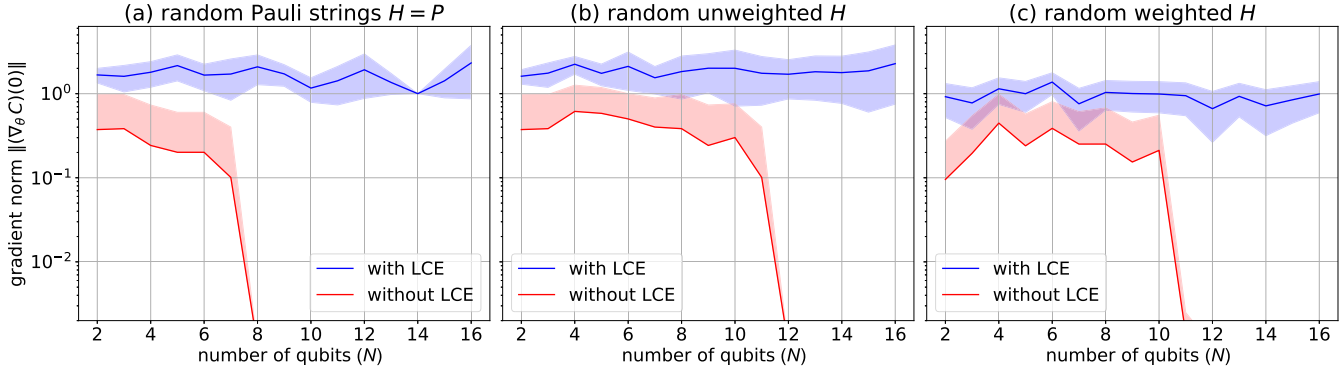


FIG. 4. Scaling behavior of the initial gradient norm at $\theta = 0$ for the `mHEA` model with $L = N$ layers. The panels vary across types of random observables. The experiment highlights the effect of the LCE transformation. We average over ten independent runs.

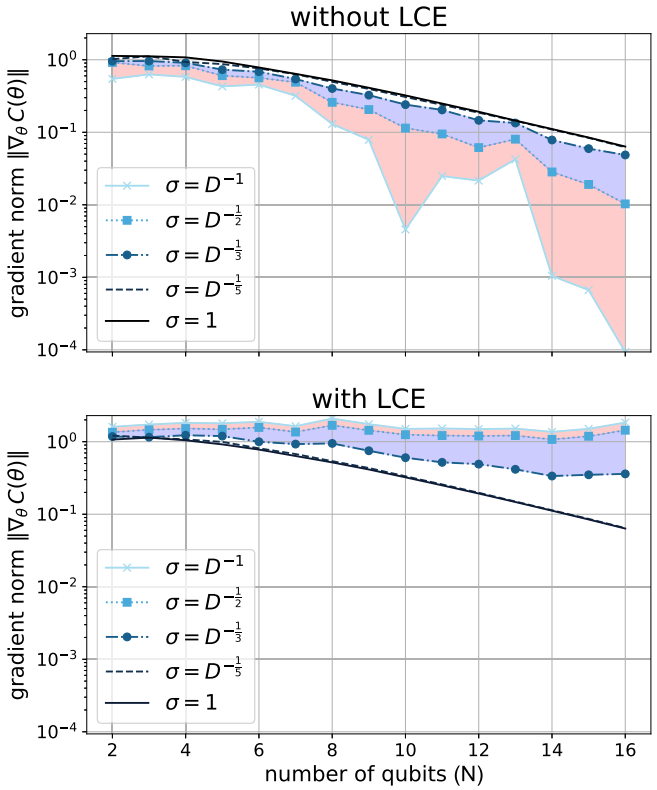


FIG. 5. Scaling behavior of the gradient norm on various landscape patches for the mHEA model with $L = N$ layers. We average over 50 independent runs of random $\theta \sim \mathbb{P}_\theta$ and uniformly sampled Pauli strings $H = P$.

orated in Sec. III B) that will be kept fixed during the entire optimization procedure. This transformation can be viewed as a transformation of the optimization landscape $C(\cdot) \mapsto \tilde{C}(\cdot)$, where the minima are close in value to the one in the original landscape $C(\cdot)$, but with different parameter coordinates (see Appendix H and Fig. 11). In particular, we do not apply an LCE transformation at each iteration during optimization. Since the LCE technique applies to initialization only, we empirically test whether improved gradients translate to practical optimization benefits—especially under barren-plateau-prone observables like global ones [25] such as $H_{\text{global}} := Z^{\otimes N}$.

Figure 6 shows VQE optimization of H_{global} employing the mHEA ansatz ($L = 5, N = 15$) via vanilla gradient descent [update rules defined as $\theta_{t+1} = \theta_t - \eta \nabla_\theta C(\theta_t)$ with constant learning rate $\eta > 0$] over five runs with Gaussian initializations $\theta \sim \mathcal{N}(0, \sigma^2 I_D)$ across varying patch sizes. The top row displays the cost minimization, highlighting faster convergence with LCE across all patch dimensions. The central row shows gradient norms, reflecting improved trainability both at initialization (cf. Theorem 2) and throughout the minimization. The bottom row tracks the parameter norm $\|\theta_t\|_2$ [which takes into account 2π periodicity of $C(\cdot)$ since the parameter trajectories remain within the hypercube $[-\pi, \pi]^D$ far from the periodic boundaries], indicative of classical hardness (cf. Theorem 3), showing that indeed this increases when enlarging the patch size. The numerical simulations continue to demonstrate successful convergence on the larger patch size $\sigma = \mathcal{O}(D^{-1/3})$, which is likely due to the improved initial gradients arising from LCE. As the patch size σ increases, the gap between standard VQE and VQE with LCE gradually decreases, eventually approaching the situation where convergence rates appear to become more similar [cf. $\sigma = \mathcal{O}(D^{-1/5})$], as for this patch size LCE seems to not have an impact on the gradients (see Fig. 5). This confirms that the benefits of LCE are most significant when employed on patches of moderate size. Additional experiments are compiled in Appendix K, where we show that changing learning rates leaves the effect of LCE invariant, and that LCE is compatible with finite sampling noise [60].

While the ground state of the $Z^{\otimes N}$ observable is not of direct practical interest, it offers a suitable testbed for barren-plateaus mitigation [12,25,37]. In Appendix K, we provide VQE experiments on the Heisenberg model [61], which is a more structured observable composed of local terms. In such a case, the benefits of LCE tend to disappear, most likely due to the absence of barren plateaus implied by the locality of the observable when employing shallow circuits [25]. Further investigation is required to find a class of practically relevant observables with a Pauli basis structure benefiting from LCE gradient improvement.

V. CONCLUSION

In this work, we investigated the interplay between gradient scalability at initialization and computational complexity in VQAs [1]. By leveraging our Taylor surrogate framework, we connected these two concepts and provided a rigorous analysis of the regimes where classical simulability and favorable gradient statistics coexist.

To address the challenge of barren plateaus, we introduced the LCE. This technique utilizes the stabilizer formalism [18,19,26] to modify the first-order structure of PQCs, ensuring constant-scaling gradient norms within local, classically simulable near-Clifford landscape patches. The LCE is implemented as an efficient classical algorithm that modifies the PQC ansatz, making it a practical tool for improving the gradient statistics of large-scale models at initialization. Our numerical results provide encouraging evidence that the LCE can enhance training dynamics in VQE tasks [2] for certain global observables that are typically prone to vanishing gradients [25].

On the complexity front, we introduced a *Taylor surrogate* that generalized the quadratic Clifford expansion [17] and leveraged the Gottesman-Knill theorem [18,57] to efficiently compute the high-order Taylor coefficients, thus establishing a classical surrogate [20] for near-Clifford landscape patches. We showed that the Taylor surrogate matches Pauli path runtime guarantees [21] with provable speedups on sufficiently small patches. Furthermore, we identified a *computational phase transition* as PQCs deviate from the near-Clifford regime. This threshold, characterized by the ℓ_1 norm of the parameters (or the random initialization patch size), identifies the boundary where our Taylor surrogate shifts from polynomial to superpolynomial complexity, as established by our lower bound analysis.

Below this threshold, where the landscape patch is classically simulable, LCE ensures constant-scaling gradients. Beyond that threshold, numerical experiments provide evidence that the gradients may begin to decay polynomially rather than exponentially. Notably, we introduce a framework

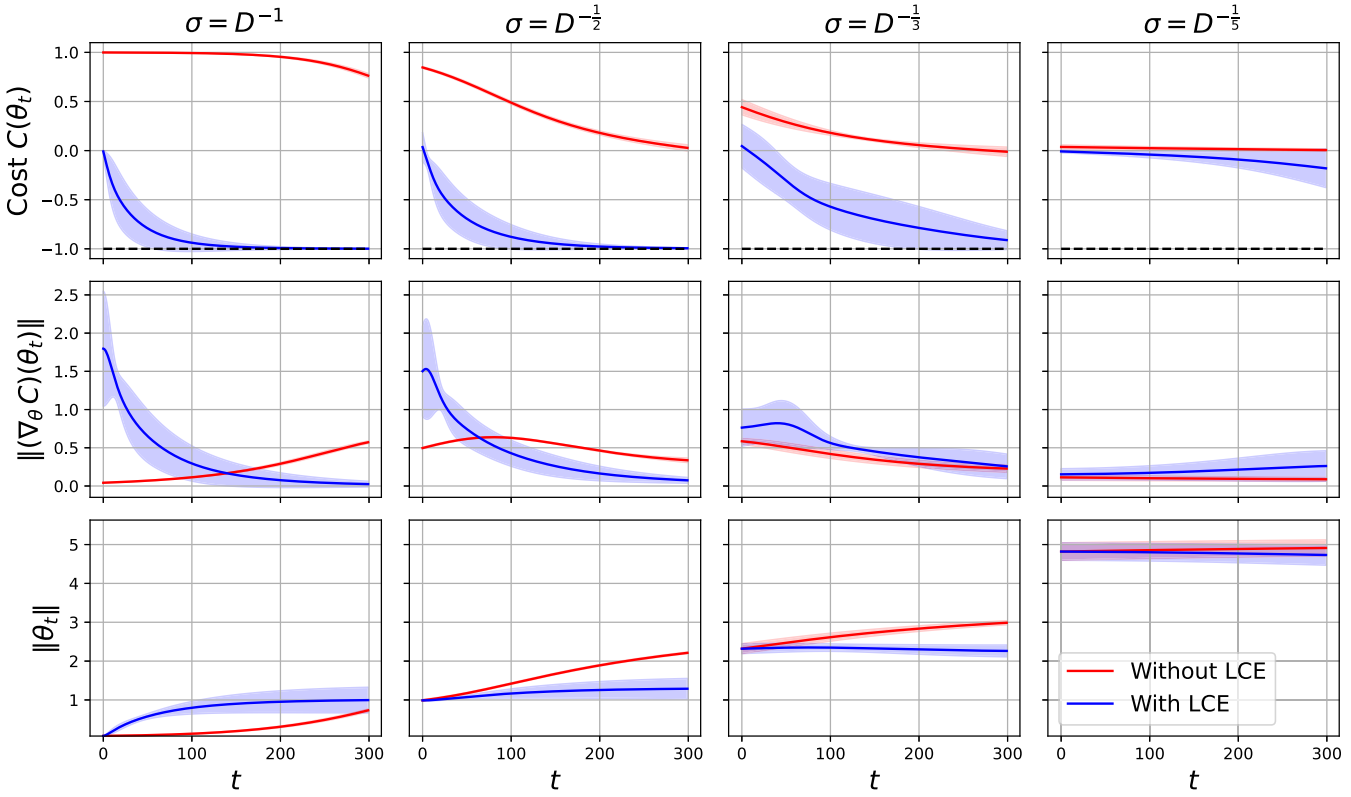


FIG. 6. VQE Optimization using vanilla gradient descent with a learning rate of 0.01, averaging over five runs of random Gaussian initialization across varying patch sizes. We find the ground energy of the global observable $Z^{\otimes N}$ with $N = 15$, employing the `mHEA` ansatz with $L = 5$ layers.

that applies to arbitrary PQCs and all observables with at least one constant-scaling Pauli term. Therefore, this extends beyond the restricted circuit geometries and local operators discussed in recent literature [14].

Despite these insights, several open questions remain. A theoretical understanding of the training dynamics with LCE, particularly the model behavior beyond initialization [56,62–66], is lacking. Importantly, we do not yet have convergence guarantees for finding the ground energy of Hamiltonians. In particular, it remains unclear whether LCE can be applied successfully for practically relevant observables. Additionally, the LCE technique is currently incompatible with QML tasks that rely on quantum feature maps [67], which typically disrupt the Clifford structure. Extending our framework to such settings—possibly by encoding data into observables [68]—may extend some of our analytical guarantees to data-driven settings. Moreover, while LCE addresses first-order structure, the role of higher-order Clifford encoders remains unexplored. Most importantly, a mathematical proof confirming our hypothesis of polynomially decaying gradients beyond the classical landscape patch remains an open problem. Bridging these gaps would offer a promising path toward identifying practical quantum advantage in the prefault tolerance era.

ACKNOWLEDGMENTS

We are grateful to Matthis Lehmkühler for his valuable input on the classical simulation of PQCs via Taylor surrogates and for discussions on the phenomenon of barren plateaus. We also thank Christa Zoufal, David Sutter, Stefan Woerner, Zoë Holmes, Ricard Puig, and Sacha Lerch for their constructive feedback, Jiří Černý for his insightful comments, and Enea Monzio Compagnoni and Jim Zhao for discussions on optimization. S.M., F.S., and A.L. acknowledge the support from SNSF Grant No. 214919. F.T. acknowledges support from SNSF Grant No. 225229 RESQUE.

DATA AVAILABILITY

The data that support the findings of this article are openly available [78].

APPENDIX A: COMPARING BARREN-PLATEAU MITIGATION TECHNIQUES AND CLASSICAL SURROGATES

We use the following Landau notations to characterize asymptotic behavior of a function $f : \mathbb{R}_{\geqslant 1} \to \mathbb{R}$ in terms of some $g : \mathbb{R}_{\geqslant 1} \to \mathbb{R}$, as $x \to \infty$:

$$f(x) = \mathcal{O}(g(x)) : \Longleftrightarrow \lim_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| \in [0, \infty),$$

$$f(x) = \Omega(g(x)) : \Longleftrightarrow \lim_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| \in (0, \infty], \tag{A1}$$

$$f(x) = \Theta(g(x)) : \Longleftrightarrow \lim_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| \in (0, \infty),$$

$$f(x) = o(g(x)) : \Longleftrightarrow \lim_{x\to\infty} \frac{f(x)}{g(x)} = 0, \tag{A2}$$

assuming that there exists $x_0 \geqslant 1$ such that $g(x) \neq 0$ for all $x \geqslant x_0$. Furthermore, we say that $f(x)$ and $g(x)$ are *asymptotically equivalent* if the much stronger condition

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1 \tag{A3}$$

holds true. In that case, we write $f(x) \sim g(x)$ as $x \to \infty$.

In order to provide an overview of our results in gradient scalability and classical computational complexity, and how they fit into the literature, we compare related barren-plateau mitigation techniques and the performance of classical surrogates with related works, respectively.

1. Comparing barren-plateau mitigation techniques

Refer to Table II in order to compare barren-plateau mitigation techniques. In this work, we established the LCE that ensures constant-scaling gradient lower bounds on $\mathcal{O}(D^{-\frac{1}{2}})$ patches (cf. Theorem 2) and is currently the only known mitigation technique providing empirical evidence of polynomially decaying gradient lower bounds beyond that patch. Notably, Park *et al.* [16] also established constant lower gradient bounds, where their assumption is satisfied with the special case where the initial state $|\psi_0\rangle = \tilde{Q}U(0)Q|0\rangle$ matches the LCE transformation. However, they achieve these asymptotics solely within the $\mathcal{O}(D^{-1})$ patch, and they also impose locality of the observable, or $H = P$ being a single Pauli string. Puig *et al.* [53] use a warm start approach in order to achieve polynomially decaying gradients on the larger $\mathcal{O}(D^{-\frac{1}{2}})$ but do not extend their analysis beyond that patch size. Additionally, we compare with uniform small-angle initialization of parameters, where Wang *et al.* [33] establish polynomially decaying gradients on patches $\mathcal{O}(D^{-\frac{1}{2}})$. Likewise, they require locality of the observable H. A Gaussian small-angles initialization strategy is studied by Zhang *et al.* [32], where they employ gradients evaluated at $\theta = 0$ to lower bound the expected gradient norm, thus bearing resemblance to our Taylor approximation technique (cf. Theorem G1). Similarly, they achieve polynomially decaying asymptotics on $\mathcal{O}(D^{-\frac{1}{2}})$ patches and require locality of the observable.

2. Comparing classical surrogates

Table III compares our Taylor surrogate performance to that relying on the Pauli path technique in the work of Lerch *et al.* [21]. Both techniques are able to simulate the patch $\mathcal{O}(D^{-1})$ in polynomial time up to a certain approximation error $\varepsilon > 0$. In our case, the degree of the polynomial is given by the explicitly determined truncation threshold $m(\theta, \varepsilon)$ of the Taylor surrogate (cf. Lemma I3). When allowing a small probability of failure $\rho > 0$, both surrogates are able to extend classical simulability to the larger patch $\mathcal{O}(D^{-\frac{1}{2}})$ (cf. Theorem 4). This is achieved with the mean-squared error

$$\mathbb{E}_\theta[(C(\theta) - C_m(\theta))^2] \leqslant c\,\|H\|^2 \frac{(D\sigma^2)^m}{m^m}, \tag{A4}$$

for some $c > 0$ (cf. Appendix I). When initializing the parameters with $\sigma = q\,\|H\|^{-\frac{1}{m}} D^{-\frac{1}{2}}$ for some sufficiently small $0.35 < q < 1$, we are able to prove that the Taylor surrogate outperforms the Pauli path approach (cf. Corollary J1), where the efficiency of Gottesman-Knill [18,19] starts to emerge.

APPENDIX B: STABILIZER FORMALISM AND THE GOTTESMAN-KNILL THEOREM

We presume a familiarity of the reader with basic concepts in group theory. The content of this Appendix may be seen as a summary and a slightly alternative point of view of the introductory works on stabilizer formalism in Ref. [26]. We begin by refreshing the necessary terminology, and conclude with a proof sketch of the version of the Gottesman-Knill theorem [18,19] that we employ in this work. For a more in-depth review of the Pauli and Clifford theory, we encourage the reader to look into Ref. [27].

1. Quantum states in an exponential Hilbert space

Consider an N-qubit quantum system on which we intend to perform error-free computation. That system may be described by a Hilbert space

$$\mathcal{H}_N := \left\{ |\psi\rangle = \sum_{b\in\{0,1\}^N} a_b\,|b\rangle : a_b \in \mathbb{C} \right\}, \tag{B1}$$

spanned by the computational basis $\{|b\rangle : b \in \{0,1\}^N\}$, where valid quantum states satisfy the normalization condition,

$$\sum_{b\in\{0,1\}^N} |a_b|^2 = 1 \tag{B2}$$

for all $b \in \{0,1\}^N$.

The Hilbert space $\mathcal{H}_N$ is exponentially large and generally difficult to fully describe classically. Of course, there are states that are given by polynomially many amplitudes a_b, which remain classically feasible. However, a groundbreaking insight leads to the discovery of *stabilizer states*, which are quantum states spanned by exponentially many amplitudes that may be classically described using polynomial resources. These states may be derived with basic concepts from group theory.

2. Stabilizer group

Stabilizer formalism is motivated from a concept in group theory called *stabilizer*. In the context of quantum computing, we say that a unitary operator $U \in \mathrm{U}(2^N)$ *stabilizes* a quantum state $|\psi\rangle \in \mathcal{H}_N$ if $U|\psi\rangle = |\psi\rangle$, i.e., $|\psi\rangle$ is an eigenstate of U with eigenvalue $+1$. From a group-theoretic perspective, the *stabilizer* of a quantum state $|\psi\rangle \in \mathcal{H}_N$ is defined as

$$\mathrm{Stab}(|\psi\rangle) := \{U \in \mathrm{U}(2^N) : U|\psi\rangle = |\psi\rangle\}, \tag{B3}$$

consisting of all unitary operators that stabilize the state $|\psi\rangle$. It can be shown that $\mathrm{Stab}(|\psi\rangle) \subset \mathcal{H}_N$ defines an Abelian subgroup for every $|\psi\rangle \in \mathcal{H}_N$. Similarly, we say that $|\psi\rangle$ is a *stabilizer* of a unitary operator $U \in \mathrm{U}(2^N)$ if $U|\psi\rangle = |\psi\rangle$.

3. N-qubit Pauli group

The Pauli matrices are defined as

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},$$
$$\text{and} \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{B4}$$

where i denotes the imaginary unit. The multiplicative group dynamics of $\mathcal{P}_1 := \langle X, Y, Z\rangle$ (modulo phases $\pm 1, \pm i$) is fully determined by the *Cayley table*,

	X	Y	Z
X	I	iZ	$-iY$
Y	$-iZ$	I	iX
Z	iY	$-iX$	I

.

In particular, we have the commutator rules

$$[P, Q] = 0 \quad \text{or} \quad \{P, Q\} = 0 \tag{B5}$$

for all $P, Q \in \mathcal{P}_1$.

The following table represents the set of stabilizer states for each element in the single-qubit Pauli group $\mathcal{P}_1$ (with $-I$ being the sole pathological case):

Operator	I	X	Y	Z	$-I$	$-X$	$-Y$	$-Z$
Stabilizers	$\mathcal{H}_N$	$\lvert+\rangle$	$\lvert i\rangle$	$\lvert 0\rangle$	$\emptyset$	$\lvert-\rangle$	$\lvert -i\rangle$	$\lvert 1\rangle$

,

where the quantum states

$$|0\rangle := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |1\rangle := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (Z \text{ axis}),$$
$$|\pm\rangle := \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle) \qquad (X \text{ axis}), \tag{B6}$$
$$|\pm i\rangle := \frac{1}{\sqrt{2}}(|0\rangle \pm i|1\rangle) \qquad (Y \text{ axis})$$

are known as the main axes of the *Bloch sphere*.

A generalization of the group $\mathcal{P}_1$ leads to the definition of the N-qubit Pauli group:

$$\mathcal{P}_N := \left\{ \phi \cdot \bigotimes_{j=1}^{N} P_j : P_j \in \{I, X, Y, Z\},\ \phi \in \{\pm 1, \pm i\} \right\}, \tag{B7}$$

and it can be shown that this defines a finite group with $|\mathcal{P}_N| = 4^{N+1}$ elements. The elements of $\mathcal{P}_N$ are known as *Pauli strings* or *Pauli words*.

In the Clifford theory, setting the stage of the Gottesman-Knill theorem discussed further below, we are particularly interested in the stabilizer group consisting of Pauli strings only. For that purpose, we introduce the N-qubit *Pauli stabilizer group* of a given quantum state $|\psi\rangle \in \mathcal{H}_N$, which is defined as

$$\text{Stab}_{\mathcal{P}_N}(|\psi\rangle) := \text{Stab}(|\psi\rangle) \cap \mathcal{P}_N. \tag{B8}$$

The intersection of two groups forms a group, and $\text{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ is Abelian since $\text{Stab}(|\psi\rangle)$ is Abelian.

4. The Clifford group and stabilizer states

The N-qubit Clifford is defined to be the subgroup of $2^N \times 2^N$ unitary operators that map tensor products of Pauli matrices to tensor products of Pauli matrices through conjugation modulo global phases, $U(1) = \{e^{i\theta} : \theta \in [-\pi, \pi]\}$:

$$\text{Clifford}(N) := \{Q \in \text{U}(2^N) : Q\,\mathcal{P}_N\,Q^\dagger = \mathcal{P}_N\} / U(1). \tag{B9}$$

In other words, these operators normalize the N-qubit Pauli group $\mathcal{P}_N$ modulo global phases. The Clifford group is well known to be finite, and generated by H, S, and CX gates, i.e., $\text{Clifford}(N) = \langle H, S, \text{CX}\rangle_N$ (modulo global phases).

The *stabilizer states* are simply given by the set of all states achieved by applying a Clifford operator to the zero state, i.e.,

$$\text{Stab}(N) := \{Q|0\rangle : Q \in \text{Clifford}(N)\}. \tag{B10}$$

It can be shown that for each $|\psi\rangle \in \text{Stab}(N)$ there exists a set of basis states $S \equiv \{b\}_{b\in S} \subseteq \{0, 1\}^N$ and $k_b \in \mathbb{N}_0$ such that

$$|\psi\rangle = \frac{1}{\sqrt{|S|}} \sum_{b\in S} i^{k_b} |b\rangle \tag{B11}$$

defines an equal superposition of basis states with relative phases in $\{e^{i\phi} : \phi \in \frac{\pi}{2}\mathbb{Z}\}$. Stabilizer states may always be efficiently described classically with the help of the Gottesman-Knill theorem, as elaborated next.

5. The Gottesman-Knill theorem

We focus on the following version of the Gottesman-Knill theorem [18,19]. We provide a proof sketch in the sense that we skip minor technical details in order to focus on the main idea behind the argumentation. The goal is to develop an intuitive formal understanding on how the group theory in the stabilizer formalism grants access to exponential savings in the classical description of stabilizer states. For more details, see Ref. [26].

The version of Gottesman-Knill (cf. Theorem B1) that we are going to prove states that every stabilizer state $|\psi\rangle \in \text{Stab}(N)$ is *uniquely* determined by its underlying Pauli stabilizer group $\text{Stab}_{\mathcal{P}_N}(|\psi\rangle) = \langle g_1, \ldots, g_N\rangle$, generated by N independent generators $g_1, \ldots, g_N \in \mathcal{P}_N$. The generators $g_1, \ldots, g_N$ are efficiently described classically with the help of *check matrices* [19,26], which establishes that stabilizer states are indeed easy to simulate classically. The proof of Theorem B1 below breaks down into three main steps:

(1) *Zero state description*: First, we establish by induction that the zero state $|0\rangle \in \text{Stab}(N)$ may always be described in terms of group generators $Z_1, \ldots, Z_N$ of the corresponding Pauli stabilizer group $\text{Stab}_{\mathcal{P}_N}(|0\rangle)$. That is, $\text{Stab}_{\mathcal{P}_N}(|0\rangle) = \langle Z_1, \ldots, Z_N\rangle$.

(2) *Zero state evolution*: Next, we employ the previous step to prove that any stabilizer state $|\psi\rangle \in \text{Stab}(N)$ may be fully described by evolving the zero state through the underlying Clifford operator. Indeed, by definition there exists $C \in \text{Clifford}(N)$ such that $|\psi\rangle = C|0\rangle$. With the help of the previous step, we then establish that the generators of $\text{Stab}_{\mathcal{P}_N}(|\psi\rangle) = \langle g_1, \ldots, g_N\rangle$ are simply obtained by evolving the generators of $\text{Stab}_{\mathcal{P}_N}(|0\rangle)$ through C in the Heisenberg picture, i.e., $g_j = CZ_jC^\dagger$ for all $j \in [N]$. This establishes that each stabilizer state is fully described by N generators.

(3) *Stabilizer state uniqueness:* Finally, it remains to prove that the generators of $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ are uniquely determined. This is achieved by introducing a certain projector operator $P_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} : \mathcal{H}_N \to V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$, which projects the full Hilbert space $\mathcal{H}_N$ onto a certain linear subspace $V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$ characterized by the Pauli stabilizer group $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$. Afterward, by using the fact that nonidentity Pauli operators are traceless, we establish that the dimension of the subspace $V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$ is equal to one. Thus, by normalization of quantum state, this shows that the group generators $g_1, \ldots, g_N$ of $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ are sufficient to uniquely determine the stabilizer state $|\psi\rangle$.

Theorem B1 (Gottesman-Knill). Every stabilizer state $|\psi\rangle \in \mathrm{Stab}(N)$ is uniquely determined by its Pauli stabilizer group, $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$. Moreover, the group $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ is spanned by N pairwise independent generators.

Proof sketch. Let $|\psi\rangle \in \mathrm{Stab}(N)$ be arbitrary. Then, by definition, $|\psi\rangle = C|0\rangle$ for some $C \in \mathrm{Clifford}(N)$. By induction, one establishes that the group

$$\mathrm{Stab}_{\mathcal{P}_N}(|0\rangle) = \langle Z_1, \ldots, Z_N\rangle \tag{B12}$$

is uniquely generated by the pairwise independent Pauli strings $Z_1, \ldots, Z_N$, where we use the conventional notations $V_j := I_{2^{j-1}} \otimes V \otimes I_{2^{N-j}}$. We use this as a basis to show that

$$\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle) = \langle CZ_1C^\dagger, \ldots, CZ_NC^\dagger\rangle \tag{B13}$$

is uniquely generated by the pairwise independent Pauli strings $CZ_1C^\dagger, \ldots, CZ_NC^\dagger$, where $C \in \mathrm{Clifford}(N)$. To prove Eq. (B13), first note that

$$\begin{aligned} P \in \mathrm{Stab}_{\mathcal{P}_N}(C|0\rangle) &\iff PC|0\rangle = C|0\rangle \iff (C^\dagger PC)|0\rangle \\ &= |0\rangle \iff \underbrace{C^\dagger PC}_{\in \mathcal{P}_N} \in \mathrm{Stab}_{\mathcal{P}_N}(|0\rangle), \end{aligned} \tag{B14}$$

again using $C \in \mathrm{Clifford}(N)$. Next, we apply Eq. (B12) to establish that there exist exponents $e_1, \ldots, e_N \in \{0, 1\}$ such that

$$\begin{aligned} &C^\dagger PC \in \mathrm{Stab}_{\mathcal{P}_N}(|0\rangle) = \langle Z_1, \ldots, Z_N\rangle \\ &\iff C^\dagger PC = Z_1^{e_1} \cdots Z_N^{e_N} \\ &\iff P = C\left(\prod_{j=1}^N Z_j^{e_j}\right)C^\dagger = \prod_{j=1}^N (CZ_jC^\dagger)^{e_j} \\ &\iff P \in \langle CZ_1C^\dagger, \ldots, CZ_NC^\dagger\rangle, \end{aligned} \tag{B15}$$

using $C^\dagger C = CC^\dagger = I$. Combined with Eq. (B14), this proves the equality in Eq. (B13). Next, we show that $CZ_1C^\dagger, \ldots, CZ_NC^\dagger$ are independent generators. For that purpose, we assume for absurd that the contrary is true. Then, there exists an index $i \in [N]$ such that

$$\mathrm{Stab}_{\mathcal{P}_N}(C|0\rangle) = \langle CZ_jC^\dagger\rangle_{j\in[N]} = \langle CZ_jC^\dagger\rangle_{j\in[N]\setminus\{i\}}. \tag{B16}$$

However, by the previous equivalence in Eq. (B15), this would imply $Z_i \in \langle Z_j\rangle_{j\in[N]} = \langle Z_j\rangle_{j\in[N]\setminus\{i\}}$, which is absurd.

It remains to show that the group $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ (and thus the set of its generators) uniquely determines the stabilizer state $|\psi\rangle$. Consider the space

$$\begin{aligned} V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} := \{|v\rangle \in \mathcal{H}_N : P|v\rangle = |v\rangle \\ \text{for all } P \in \mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)\}. \end{aligned} \tag{B17}$$

It can be shown that $V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} \subset \mathcal{H}_N$ defines a linear subspace, and it consists of all states that are fixed by elements of the Pauli stabilizer group $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ (in particular, by its generators). We are going to establish that its dimension equals $\dim V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} = 1$. Indeed, this completes the proof because by normalization of quantum states, together with the one dimensionality, Eq. (B17) implies the existence of a unique quantum state $|\psi\rangle \in \mathcal{H}_N$.

By Eq. (B13), there exist Pauli generators $g_1, \ldots, g_n \in \mathcal{P}_N$ such that $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle) = \langle g_1, \ldots, g_N\rangle$. Consider the operator

$$P_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} := \frac{1}{2^N}\prod_{j=1}^N (I + g_j). \tag{B18}$$

Then, we note that $P_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} : \mathcal{H}_N \to V_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$ defines a projector. Indeed, for each $j \in [N]$ we have

$$\begin{aligned} &\left(\frac{I+g_j}{2}\right)^2 = \frac{1}{4}(I + 2g_j + g_j^2) = \frac{I+g_j}{2} \\ &\text{and} \quad \left(\frac{I+g_j}{2}\right)|\psi\rangle = \frac{|\psi\rangle + g_j|\psi\rangle}{2} = |\psi\rangle, \end{aligned} \tag{B19}$$

using $g_j^2 = I$ and $g_j|\psi\rangle = |\psi\rangle$ since $g_j \in \langle g_j\rangle_{j\in[N]} = \mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)$ is a Pauli string that stabilizes $|\psi\rangle$. In other words, for each $j \in [N]$, $\frac{1}{2}(I + g_j)$ is the projector onto the $+1$ eigenspace of g_j. Taking the product, this shows that $P_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$ defines the claimed projector, and it can be shown that it is surjective.

The crucial observation now is that this projector decomposes as

$$\begin{aligned} P_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} &= \frac{1}{2^N}\prod_{j=1}^N (I + g_j) \\ &= \frac{1}{2^N}\sum_{e_1,\ldots,e_n\in\{0,1\}} g_1^{e_1}\cdots g_N^{e_N} \\ &= \frac{1}{2^N}\sum_{g\in\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} g, \end{aligned} \tag{B20}$$

where the second equality is obtained by expanding the product, and the fact that all g_j commute pairwise (else they must anticommute, which is impossible because of $g_jg_i|\psi\rangle = -g_ig_j|\psi\rangle = -|\psi\rangle$). The last equality simply follows from $\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle) = \langle g_1, \ldots, g_N\rangle$. This observation immediately implies that taking its trace yields

$$\mathrm{Tr}\, P_{\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} = \frac{1}{2^N}\sum_{g\in\mathrm{Stab}_{\mathcal{P}_N}(|\psi\rangle)} \mathrm{Tr}\, g = \frac{1}{2^N}\underbrace{\mathrm{Tr}\, I}_{=2^N} = 1, \tag{B21}$$

using that $\mathrm{Tr}\, g = 0$ if and only if $g \neq I$. On the other hand, since the trace of an operator equals the sum of its eigenvalues,

we find

$$1 = \operatorname{Tr} P_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)} = \sum_{k=1}^{2^N} \underbrace{\lambda\big(\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)\big)}_{\in\{0,1\}} = \dim \operatorname{Image}\big(P_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)}\big) = \dim V_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)}. \tag{B22}$$

Here, we used the fact that the eigenvalues of a projector are either 0 or 1. The sum over all eigenvalues of the (linear) projector operator $P_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)} : \mathcal{H}_N \to V_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$, thus counts the dimensions that are not in its kernel, and thus counts the dimension of its image, which precisely equals the dimension of the subspace $V_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$, proving Eq. (B17). This completes the proof sketch. ■

This version of Gottesman-Knill proves that there exist states $|\psi\rangle$ (stabilizer states) with 2^N nonzero amplitudes, that are classically fully characterized by only N generators of the underlying Pauli stabilizer group $\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)$. Aaronson and Gottesman developed an efficient method, known as the *check matrix algorithm* [19], establishing that a stabilizer state $C|0\rangle$ requires space resources of order $\mathcal{O}(D_C^2)$, where D_C is the number of elementary gates in the circuit $C \in \operatorname{Clifford}(N)$. Furthermore, they showed that the time resources required to obtain $C|0\rangle$ from $|0\rangle$ is of order $\mathcal{O}(D_C)$. In our setting, where we consider a PQC $U(\theta)$ generated by $D_C \geqslant 1$ elementary Clifford gates and $D \geqslant 1$ parametrized single-qubit (generally non-Clifford) rotation gates, we may generally assume that $D_C = c\,D$ for some constant factor $c > 0$, so that simulating $U(0)|0\rangle$ requires $\mathcal{O}(D^2)$ space resources, since $U(0) \in \operatorname{Clifford}(N)$.

Next, we show that simulating the expectation $\langle 0|U(0)^\dagger H U(0)|0\rangle$ yields a run-time of order $\mathcal{O}(D\,|\mathcal{I}|)$, where $|\mathcal{I}|$ is the number of Pauli terms in the observable H.

6. Computing expectations with the stabilizer formalism

We may use the Gottesman-Knill theorem (Theorem B1) in order to prove that Hamiltonian expectations of the form $\langle 0|C^\dagger H C|0\rangle$ may be computed in polynomial time for each Clifford operator $C \in \operatorname{Clifford}(N)$. The result particularly holds true for PQCs evaluated at Clifford configurations, i.e., $C = U(0)$.

Corollary B1. Let $|\psi\rangle \in \operatorname{Stab}(N)$ be a stabilizer state, and $P \in \mathcal{P}_N$ a Pauli string. Let $g_1, \ldots, g_N \in \mathcal{P}_N$ be the generators of Pauli stabilizer group $\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)$. Then,

$$\langle\psi|P|\psi\rangle = \begin{cases} \operatorname{sign} P, & \text{if } [P, g_j] = 0 \text{ for all } j \in [N], \\ 0, & \text{if there exists } j \in [N] \\ & \quad \text{such that } \{P, g_j\} = 0, \end{cases}$$

where $\operatorname{sign} P \in \{\pm 1\}$ denotes the sign of the Pauli string.

Proof. Since $P \in \mathcal{P}_N$ is a Pauli string, it has eigenvalues in $\{\pm 1\}$. Therefore, we have

$$\langle\psi|P|\psi\rangle = p(+1) \times 1 + p(-1) \times (-1) = p(+1) - p(-1), \tag{B23}$$

where $p(\pm 1)$ is the probability of measuring the outcome ± 1 when the system is prepared in the state $|\psi\rangle$. By linearity and the cyclic property of the trace operator, we first establish that

$$\begin{aligned} p(+1) - p(-1) &= \langle\psi|P|\psi\rangle = \operatorname{Tr}(P|\psi\rangle\langle\psi|) \\ &= \operatorname{Tr}\left(\frac{I+P}{2}|\psi\rangle\langle\psi|\right) - \operatorname{Tr}\left(\frac{I-P}{2}|\psi\rangle\langle\psi|\right), \end{aligned} \tag{B24}$$

and by Born's rule, it immediately follows that

$$p(\pm 1) = \operatorname{Tr}\left(\frac{I \pm P}{2}|\psi\rangle\langle\psi|\right). \tag{B25}$$

Therefore, the projectors $\frac{1}{2}(I \pm P)$ determine the probability distributions $p(\pm 1)$ (of the outcomes ± 1) needed for the computation of the expectation value $\langle\psi|P|\psi\rangle$. It turns out that certain commutation rules of the Pauli operators involved suffice to explicitly determine these probabilities. More precisely, if the following holds true:

$$\begin{cases} [P, g_j] = 0 \text{ for all } j \in [N] \Rightarrow p(\operatorname{sign} P) = 1 \text{ and } p(-\operatorname{sign} P) = 0 & \text{(deterministic outcome)}, \\ \text{there exists } j \in [N] \text{ such that } \{P, g_j\} = 0 \Rightarrow p(+1) = p(-1) = \frac{1}{2} & \text{(coin flip outcome)}, \end{cases} \tag{B26}$$

then combining Eq. (B26) with Eq. (B24) would yield

$$\langle\psi|P|\psi\rangle = p(+1) - p(-1) \overset{\text{(B26)}}{=} \begin{cases} \operatorname{sign} P, & \text{if } [P, g_j] = 0 \text{ for all } j \in [N], \\ 0, & \text{if there exists } j \in [N] \text{ such that } \{P, g_j\} = 0. \end{cases} \tag{B27}$$

Therefore, it remains to prove Eq. (B26). First, let us handle the case where $[P, g_j] = 0$ for all $j \in [N]$. Recall the definition of the linear subspace $V_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)}$ in Eq. (B17) introduced in the proof of Theorem B1. By assumption, we have for all $j \in [N]$ that

$$g_j P|\psi\rangle = P g_j|\psi\rangle = P|\psi\rangle \Rightarrow P|\psi\rangle \in V_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)}. \tag{B28}$$

We are going to use this observation in order to establish that $p(\operatorname{sign} P) = 1$. Indeed, since $\dim V_{\operatorname{Stab}_{\mathcal{P}_N}(|\psi\rangle)} = 1$ (as shown in Theorem B1), we know that there exists a phase $\phi \in \mathbb{C}$ such that $P|\psi\rangle = \phi|\psi\rangle$. Since $P \in \mathcal{P}_N$, we must have that $\phi \in \{\pm 1\}$ (eigenvalues of P). As a result, we find that $P|\psi\rangle = \operatorname{sign}(P)|\psi\rangle$. Consequently,

$$\begin{aligned} p(\operatorname{sign} P) &\overset{\text{(B25)}}{=} \operatorname{Tr}\left(\frac{I + \operatorname{sign}(P)\,P}{2}|\psi\rangle\langle\psi|\right) \\ &= \operatorname{Tr}\left(\frac{I + \operatorname{sign}(P)^2 I}{2}|\psi\rangle\langle\psi|\right) \\ &= \operatorname{Tr}(|\psi\rangle\langle\psi|) = 1. \end{aligned} \tag{B29}$$

In particular, $p(-\operatorname{sign} P) = 1 - p(\operatorname{sign} P) = 1 - 1 = 0$, which proves the deterministic outcome in Eq. (B26).

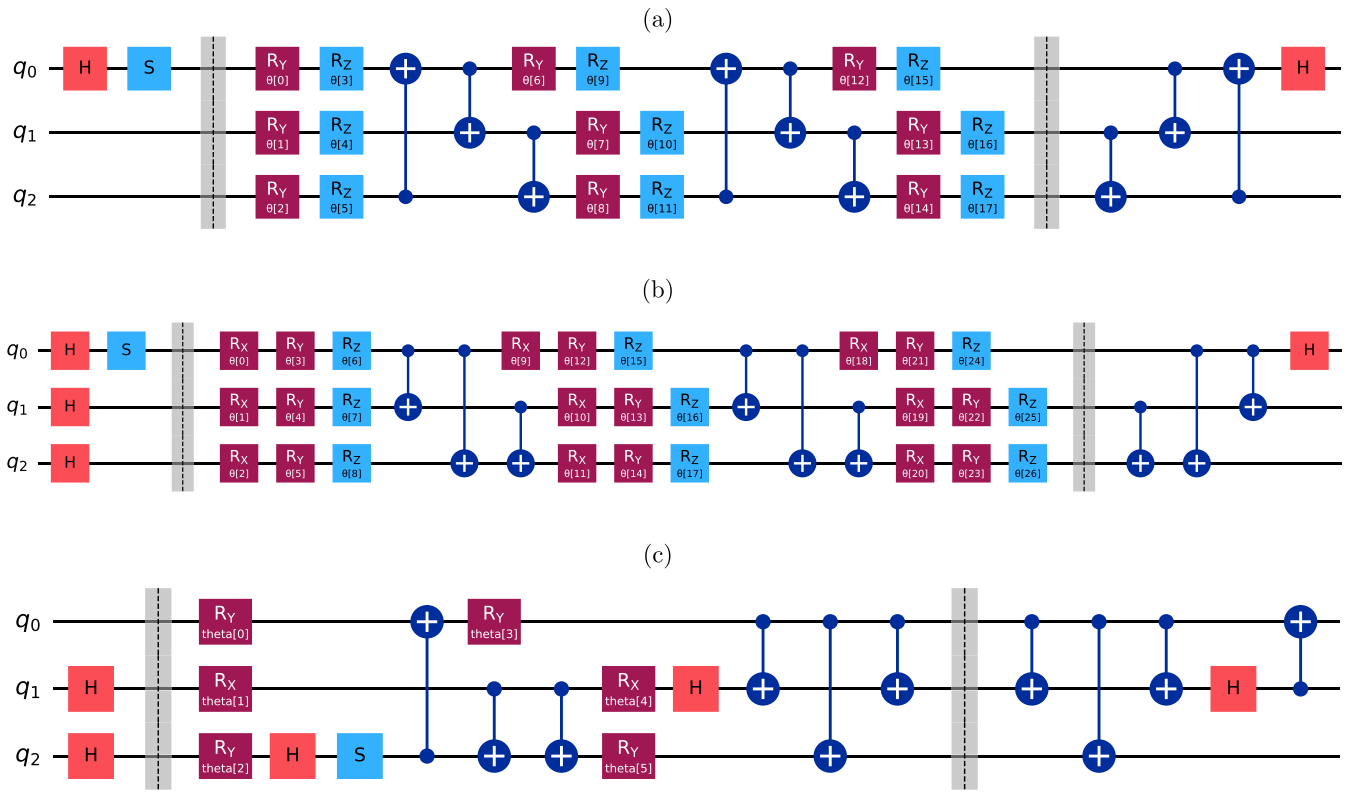


FIG. 7. Examples of two-layer PQCs after LCE transformation. (a) `mHEA` circuit with $N = 3$, $D = 18$, $k = 10$, and $P = XXZ$. (b) `fHEA` circuit with $N = 3$, $D = 27$, $k = 16$, and $P = IIZ$. (c) `rPQC` circuit with $N = 3$, $D = 6$, $k = 4$, and $P = IXI$.

On the other hand, if there exists $j \in [N]$ such that $\{P, g_j\} = 0$, then

$$g_j P|\psi\rangle = -P g_j|\psi\rangle = -P|\psi\rangle, \tag{B30}$$

which in turn, once again making use of the cyclic property of the trace operator, implies that

$$\begin{aligned} p(-1) &\overset{(B25)}{=} \mathrm{Tr}\left(\frac{I-P}{2}|\psi\rangle\langle\psi|\right) = \mathrm{Tr}\left(\frac{I+g_jP}{2}|\psi\rangle\langle\psi|\right) \\ &= \frac{1}{2}\langle\psi|\psi\rangle + \frac{1}{2}\underbrace{\langle\psi|g_j}_{\langle\psi|}P|\psi\rangle = \frac{1}{2}\langle\psi|(I+P)|\psi\rangle \\ &= \mathrm{Tr}\left(\frac{I+P}{2}|\psi\rangle\langle\psi|\right) = p(+1). \end{aligned} \tag{B31}$$

In particular, $p(+1) = p(-1) = \frac{1}{2}$, as claimed. ■

As previously established by Theorem B1, we already know that all states involved only require polynomial classical space resources. Together with Corollary B1, this implies that any cost function, $C(0) = \langle 0|U(0)^\dagger H U(0)|0\rangle$, evaluated at Clifford configurations (without loss of generality $\theta = 0$) may be classically evaluated in polynomial time. Indeed, the time complexity of classically evaluating $\langle\psi|P|\psi\rangle$ for a single Pauli string $P \in \mathcal{P}_N$ scales as $\mathcal{O}(N)$ since at most N of the generators $g_1, \ldots, g_N$ need to be compared with P in order to determine the commutator rules in Eq. (B26). Doing this for each of the Pauli terms in the observable yields a run-time scaling as $\mathcal{O}(N|\mathcal{I}|)$, where $|\mathcal{I}|$ is the number of Pauli terms in the observable H.

In this work, we make the assumption that the ansatz $U(\theta)$ satisfies that $D = \Theta(N)$ (i.e., the number of parameters is a multiple of the number of qubits) so that the time complexity of evaluating $C(0) = \langle 0|U(0)^\dagger H U(0)|0\rangle$ scales as $\mathcal{O}(D|\mathcal{I}|)$ as established in our complexity theory results.

APPENDIX C: PQC ANSATZ ARCHITECTURES

In this work, we consider three PQC model ansätze with $L \geqslant 1$ layers:

(1) `mHEA` [58]: Minimalistic Hardware Efficient Ansatz based on Qiskit's `EfficientSU2` [59] with parametrized Y- and Z-single-qubit rotation layers, and circular entanglements. The number of parameters satisfies $D = 2N(L+1)$.

(2) `fHEA`: Full Hardware Efficient Ansatz based on Qiskit's `EfficientSU2` [59] with parametrized X-, Y-, and Z-single-qubit rotation layers, and full entanglements. The number of parameters satisfies $D = 3N(L+1)$.

(3) `rPQC`: We construct a layerwise random PQC as follows: Independently, for each layer, we apply a layer of single-qubit H gates followed by a layer of single-qubit S gates. Each single-qubit gate appears with 0.5 probability. Then, we apply N uniformly random pairs of CX gates. The number of parameters satisfies $D = NL$.

Figure 7 illustrates representative examples of the PQC models used in our experiments after applying the LCE transformation $(Q, \tilde{Q})$. The circuit starts with the Clifford Q to diagonalize the Pauli string, and ends with $\tilde{Q}$, used to induce a relative phase between two parameter-shifted terms $C(\pm\pi/2\, e_k)$. Panel (a) shows `mHEA` with $N = 3$ qubits and

TABLE IV. Comparing state-of-the-art simulation techniques with our classical Taylor surrogate method. These are applied to simulate near-Clifford circuits. The comparison is made in terms of the worst-case run-time required for exactly computing $C(\theta)$, assumptions on the quantum model ansatz, and the analysis performed of the computational complexity as a function of θ.

	Exact run-time	Circuit restrictions	Complexity in terms of θ
Low-rank stabilizer decomposition [22,23]	$\mathcal{O}(\lvert\mathcal{I}\rvert\, 2^{0.48D})$	3-locality (cf. Eq. (D6))	Bounding $m(\theta)$ (cf. Fig. 8)
Clifford perturbation theory [21,24]	$\mathcal{O}(\lvert\mathcal{I}\rvert\, 2^{D})$	None	None
Truncated Taylor series	$\mathcal{O}(\lvert\mathcal{I}\rvert\, 2^{2D})$	None	$m = \Theta(\lVert\theta\rVert_1)$ (cf. Lemma I3)

$D = 18$ parameters, LCE transformed using $k = 10$ and Pauli string $P = XXZ$. Panel (b) presents `fHEA` with $N = 3$ qubits and $D = 27$ parameters, LCE transformed using $k = 16$ and $P = IIZ$. Panel (c) depicts `rPQC` with $N = 3$ qubits, $D = 6$ parameters, and LCE transformed using $k = 4$ and $P = IXI$.

APPENDIX D: COMPARING STATE-OF-THE-ART SIMULATION TECHNIQUES

In general, the computational resources needed to evaluate the cost $C(\theta)$ in Eq. (6) are characterized by the number $\mathcal{O}(|\mathcal{J}|)$ of operations in an estimator of the form

$$C(\theta) \approx \sum_{j\in\mathcal{J}} f_j(\theta), \tag{D1}$$

where $f_j : [-\pi, \pi]^D \to \mathbb{R}$ are certain parameter-dependent functions.

1. Naive evaluation of the exponential quadratic form

Recall that, if the n-qubit system is represented by a Hilbert space $\mathcal{H}_N$ with state vectors of the general form

$$|\psi\rangle = \sum_{b\in\{0,1\}^N} a_b\,|b\rangle,$$
$$\text{with amplitudes } a_b \in \mathbb{C} : \sum_{b\in\{0,1\}^N} |a_b|^2 = 1 \tag{D2}$$

in the computational basis, then the cost is naively expressed as the exponential quadratic form (exponentially large matrix, and exponentially many terms depending on the state vector),

$$C(\theta) \overset{(6)}{=} \langle\psi(\theta)|H|\psi(\theta)\rangle = \sum_{b,b'\in\{0,1\}^N} a_b^* a_{b'} \langle b|H|b'\rangle$$
$$= \sum_{b,b'\in\{0,1\}^N} \sum_{i\in\mathcal{I}} a_b^* a_{b'} \langle b|P_i|b'\rangle. \tag{D3}$$

In that case, we find $|\mathcal{J}| = |\mathcal{I}|\,\#\{0,1\}^{2N} = 2^{2N}$, thus implying a worst-case run-time of $\mathcal{O}(|\mathcal{I}|\,2^{2N})$.

The computational cost of evaluating the quadratic form (6) inherently depends on the specific parameter configuration θ. In general, estimating quantum expectation values is exponentially hard [22–24], but depending on the values of the parameters we can classically approximate them more efficiently. For example, if $\theta = 0$, then we can simulate $C(\theta = 0)$ in polynomial time due to the Gottesman-Knill theorem [18,19].

In what follows, we provide a brief comparison of state-of-the-art simulation techniques for approximating the expectation value $C(\theta)$ as a function of θ. A short overview is found in Table IV.

2. Low-rank stabilizer decomposition

Taking into consideration a parametrized state $|\psi(\theta)\rangle = U(\theta)|0\rangle$, setting the parameters to 0 results in a stabilizer state [18,19] $|\psi(0)\rangle \in \mathrm{Stab}(N)$. Such states can be efficiently represented and manipulated using Aaronson-Gottesman's check matrix algorithm [19] leveraging the Gottesman-Knill theorem [18] (cf. Theorem B1).

This observation led to the classical simulation technique of Bravyi *et al.* [22,23], which is based on a *low-rank stabilizer decomposition* of $|\psi(\theta)\rangle$. Specifically,

$$|\psi(\theta)\rangle \approx \sum_{\ell=1}^{m} \gamma_\ell(\theta)\,|\phi_\ell\rangle, \tag{D4}$$

where each $|\phi_\ell\rangle$ defines a stabilizer state, efficiently described within the stabilizer formalism, and $\gamma_\ell(\theta)$ are parameter-dependent complex coefficients.

The computational cost of calculating the expectation value in Eq. (D3) is directly characterized by the decomposition number $m = \chi(\theta)$ (as a function of θ), also known as the *stabilizer rank*. Also, note that $m = \sqrt{|\mathcal{J}|}$ in Eq. (D1) since taking the expectation value requires two copies of the low-rank stabilizer decomposition, hence an extra square root. In particular, starting from a Clifford configuration $\theta = 0$, as $U(\theta)$ approaches a circuit containing an increasing number of non-Clifford gates, the number of terms in the decomposition (D4) grows, ultimately leading to an exponential scaling of the simulation cost [cf. Eq. (D3)].

Using the stabilizer decomposition, the expectation value can then be approximated as

$$C(\theta) \approx \sum_{\ell,\ell'=1}^{m} \gamma_\ell(\theta)\,\gamma_{\ell'}(\theta)\,\langle\phi_\ell|H|\phi_{\ell'}\rangle, \tag{D5}$$

with the approximation error decreasing as m increases. As of this writing, the most efficient algorithm for implementing Eq. (D5) is reported by Bravyi *et al.* [23], achieving a worst-case computational complexity of $\mathcal{O}(|\mathcal{I}|\,2^{0.48D})$.

Unfortunately, their analysis assumes a circuit architecture limited to 3-local qubit gates (Proposition 1 in Ref. [23]), thereby imposing a light-cone constraint that reduces the computational complexity of their method. More precisely, they require that

$$|\psi(\theta)\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle, \tag{D6}$$

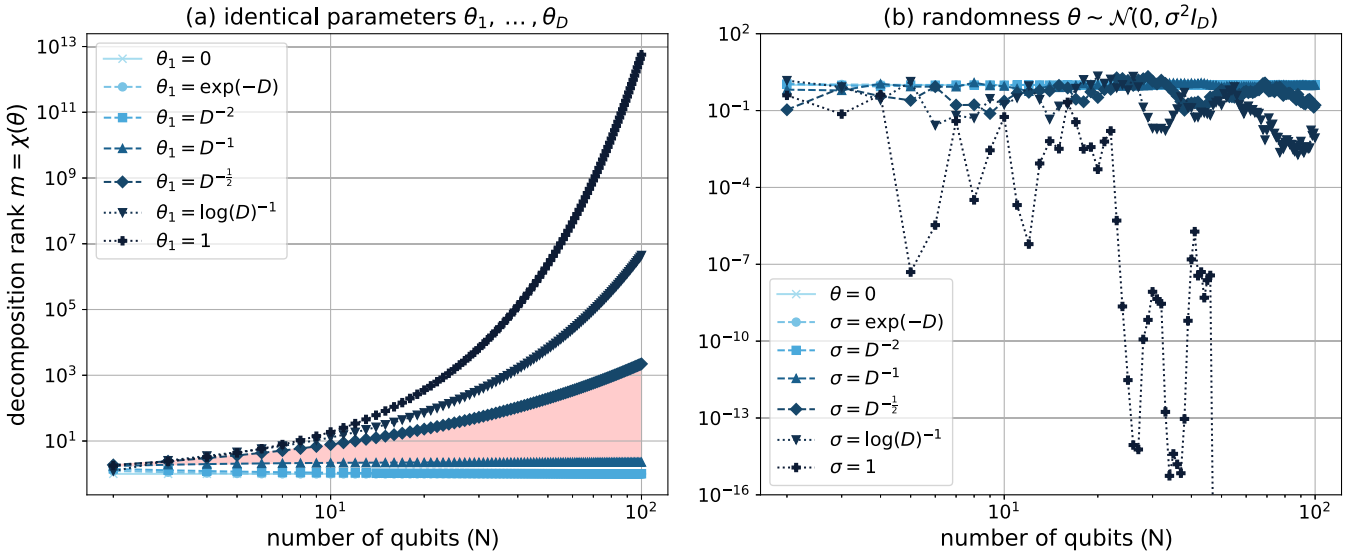


FIG. 8. Stabilizer rank $m = \chi(\theta)$ as a function of qubit number N for a model with $D = N$ parameters. (a) Deterministic parameters show a complexity phase transition and a barren-plateau-free region (red). (b) Randomly initialized parameters yield uninformative scaling of m with the number of qubits N.

with each state $|\psi_j\rangle$ describing a system of at most three qubits. Consequently, this strongly limits the expressive power of the ansatz model.

In Fig. 8, we analyze the stabilizer rank $m = \chi(\theta)$ [Eq. (14) in Ref. [23]] as a function of the number of qubits N, under the same experimental setting as Fig. 3. In panel (a), where all parameters are deterministically initialized as $\theta_1 = \cdots = \theta_D$, we observe a clear phase transition in computational complexity, consistent with Fig. 3. The red-shaded region indicates the barren-plateau-free regime characterized in Theorem 2. In contrast, panel (b) shows that for randomly initialized parameters $\theta \sim \mathcal{N}(0, \sigma^2 I_D)$, the decomposition rank $m = \chi(\theta)$ exhibits no meaningful dependence on N, since $m \geqslant 1$ is always a positive integer. This demonstrates that the low-rank stabilizer decomposition technique fails to capture the statistical behavior of $\chi(\theta)$ under random initialization $\mathbb{P}_\theta$. The key issue is that the formula used in Ref. [23] assumes parameter correlations that are not present under i.i.d. initializations, thereby violating the conditions required for the law of large numbers. In contrast, our formulation $m = m(\|\theta\|_1)$ satisfies these assumptions asymptotically (cf. Lemma I3). As a result, the framework of Bravyi *et al.* is unsuitable for testing initialization strategies based on random sampling.

3. Pauli path formalism

Begušić *et al.* [24] take a fundamentally different approach to the problem. Instead of explicitly describing the quantum state $|\psi(\theta)\rangle$ first, they directly expand the expectation value [Eq. (D3)] by iteratively applying commutator rules. Specifically, for each Pauli-rotation gate R_{V_k} with $k \in [D]$ in the circuit ansatz, and each Pauli term P_i with $i \in \mathcal{I}$ in the cost function, they use

$$R_{V_k}(\theta_k)^\dagger P_i R_{V_k}(\theta_k) = \begin{cases} P_i, & [P_i, V_k] = 0, \\ \cos\frac{\theta_k}{2} P_i + i \sin\frac{\theta_k}{2} V_k P_i, & \{P_i, V_k\} = 0, \end{cases} \tag{D7}$$

to propagate each Pauli-rotation gate through the observable, where $[\cdot,\cdot]$ and $\{\cdot,\cdot\}$ denote the commutator and anticommutator, respectively.

This results in an alternative series expansion of the expectation value:

$$C(\theta) \approx \sum_{i\in\mathcal{I}} c_i \sum_{\ell=1}^{m} f_\ell(\theta)\, \langle 0|\tilde{P}_\ell|0\rangle, \tag{D8}$$

where $f_\ell(\theta)$ are sinusoidal functions of θ and $\tilde{P}_\ell$ are newly propagated Pauli strings. More concretely, the functions $f_\ell(\theta)$ are products of sine and cosine functions so that if $f_\ell(\theta)$ contains many sine factors, it will be negligibly small if $\theta \approx 0$ due to $\sin\theta \approx \theta$. This motivates the idea behind the approximation in Eq. (D8), where one truncates the terms containing many sine factors, which do not contribute much to the error. This simulation technique is used for surrogating near-Clifford patches in the works of Lerch *et al.* [21].

In this approach, the decomposition number m depends on how frequently the Pauli-rotation gates in the circuit ansatz fail to commute with the Pauli strings in the observable. Their simulation strategy has a worst-case complexity of $\mathcal{O}(|\mathcal{I}|\, 2^D)$, but if all involved Pauli matrices are uniformly random, the average-case scaling improves to $\mathcal{O}(|\mathcal{I}|\, 2^{0.59D})$ [24].

Although it exhibits worse scaling compared to Bravyi *et al.* [22,23] this may be due to their analysis imposing no restrictions on the circuit model ansatz. However, their approach does not provide a closed-form expression for the truncation number $m = K(\theta)$ as a function of θ, making it less convenient for characterizing time complexity in terms of θ. This may be due to the technical challenges of their perturbative expansion in the Heisenberg picture, which makes it difficult to analytically capture the dependency $m = K(\theta)$. In contrast, our approach offers a simple characterization of $m = m(\theta)$ in Lemma I3, though at the expense of increased computational complexity when applied to simulating $C(\theta)$.

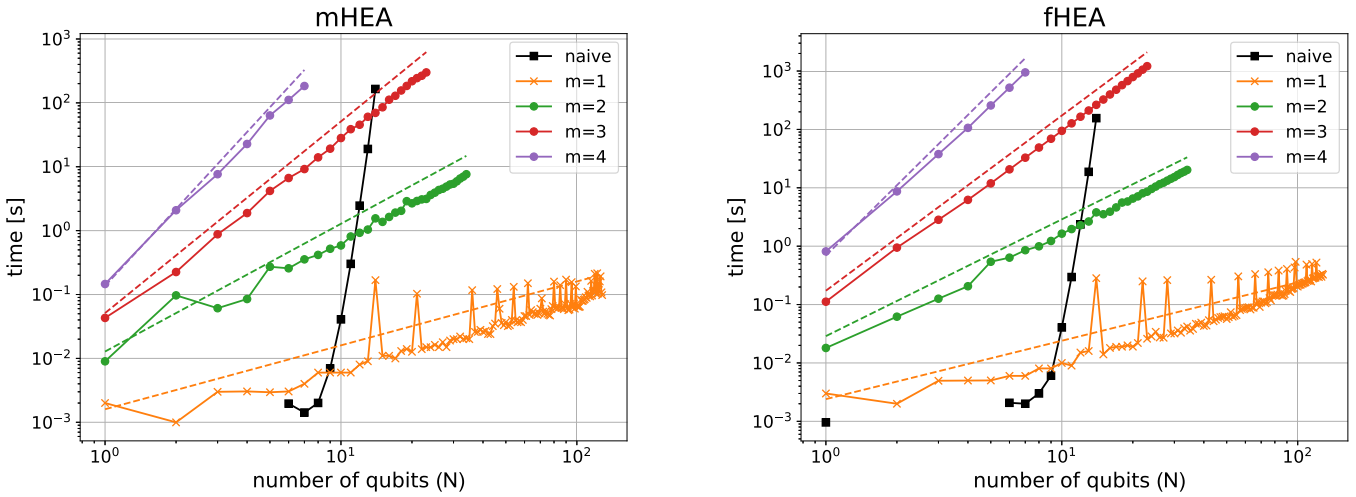


FIG. 9. Simulation time of $C_m(\theta)$ using `mHEA` (left) and `fHEA` (right) for increasing qubit number N and Taylor orders $m \leqslant 4$. Due to the additional number of parameters, `fHEA` has a computational overhead of 1.5^m compared to `mHEA`.

4. Truncated Taylor series

Mitarai *et al.* [17] first introduced a Taylor approach for approximating the expectation value (D3) classically. Their method corresponds to the special case of a truncated Taylor expansion with $m = 3$:

$$C(\theta) \approx \sum_{\|\alpha\|_1 < m} \left(D_\theta^\alpha C\right)(0)\, \frac{\theta^\alpha}{\alpha!}, \tag{D9}$$

where for each multi-index $\alpha = (\alpha_1, \ldots, \alpha_D)$ we use the notations $D_\theta^\alpha := \partial_{\theta_1}^{\alpha_1} \cdots \partial_{\theta_D}^{\alpha_D}$, $\theta^\alpha := \theta_1^{\alpha_1} \cdots \theta_D^{\alpha_D}$, $\alpha! := \alpha_1! \cdots \alpha_D!$, and $\|\alpha\|_1 := |\alpha_1| + \cdots + |\alpha_D|$ denoting the ℓ_1 norm of α.

This expansion provides an explicit polynomial approximation of $C(\theta)$, enabling efficient classical evaluations for small values of θ. Lerch *et al.* [21] also employed a truncated series of this form, not as a simulation method, but rather as part of their analysis in proving their results [cf. their Eq. (C6) and Theorems 1 and 2].

One particularly interesting difference regarding the decomposition complexity $|\mathcal{J}|$ is that the Taylor truncation threshold m is used to upper bound a set of multi-indices $\{\alpha \in \mathbb{N}_0^D : \|\alpha\|_1 < m\}$, while the other methods do not iterate over multi-indices. This distinction allows us to characterize complexity more easily as a function of θ by focusing solely on the multi-index upper bound $m = m(\theta)$ as evident from the proofs of Lemma I3 and Theorem 3, respectively. Additionally, this explains why the scaling of our multi-index threshold $m = m(\theta)$ in Fig. 3 is simpler compared to the full decomposition threshold used by Bravyi *et al.* [23] in Fig. 8, because an upper bound for multi-indices grows much smaller compared to the direct number of multi-indices. Furthermore, the simplicity of our formula ensures consistency with the law of large numbers, unlike the formula in Eq. (14) of Ref. [23] by Bravyi *et al.*, which is not a sum of independent random variables.

Figure 9 illustrates the cost of classically simulating $C_m(\theta)$ under the two PQC models `mHEA` and `fHEA` (introduced in Appendix C), for various system sizes N and Taylor expansion orders $m \leqslant 4$. The simulations use uniformly random Pauli strings $H = P_i$ (in particular, $|\mathcal{I}| = 1$) and initialization $\mathbb{P}_\theta \sim \mathcal{N}(0, \sigma^2 I_D)$ with $\sigma = D^{-1}$. Theorem 3 states that the time complexity of evaluating $C_m(\theta)$ is a polynomial scaling as $\mathcal{O}(D^m)$. Using that `mHEA` and `fHEA` each satisfy $D = 2N(L+1)$ and $D = 3N(L+1)$, respectively, we see that `fHEA` has a computational overhead of $(3/2)^m = 1.5^m$. For large m, computing 3^m takes significantly more time and space resources than 2^m. The special case $m = 1$ corresponds to the Gottesman-Knill theorem [18,19]. The dashed lines represent the theoretical prediction as per Theorem 3. As expected, higher-order approximations ($m > 0$) lead to increased computational cost, especially as N grows.

The following Appendixes present the proofs of our theoretical results, organized thematically for clarity. We begin by defining the PQC ansatz employed in our numerical experiments. This is followed by proofs of the high-order parameter-shift rules and cost function regularities. Subsequently, we introduce the LCE and demonstrate how it can be leveraged to mitigate barren plateaus. We then establish that the LCE transformation preserves both the expressivity of the ansatz and the global structure of the optimization landscape. The subsequent Appendixes address computational complexity. We first derive the minimal truncation threshold required for the Taylor surrogate and use this to characterize the associated computational cost. Finally, we provide additional numerical experiments to support our theoretical findings.

APPENDIX E: HIGH-ORDER PARAMETER-SHIFT RULES AND COST REGULARITY

In this Appendix, we study the regularity properties of the cost function $C(\cdot)$. Recall from Eq. (6) that

$$C(\theta) := \langle\psi(\theta)|H|\psi(\theta)\rangle = \sum_{i\in\mathcal{I}} c_i \langle\psi(\theta)|P_i|\psi(\theta)\rangle. \tag{E1}$$

We first establish smoothness and explicitly demonstrate differentiability using the standard parameter-shift rule [47,48]. We then extend this to the higher-order parameter-shift rule [35] and leverage this structure to derive strong regularity bounds for the cost function.

For a fixed multi-index $\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathbb{N}_0^D$, we define the class $\mathcal{C}^\alpha$ to be the space of functions $f : [-\pi, \pi]^D \to \mathbb{R}$ for which the higher-order derivative $D_\theta^\alpha f$ of f exists and is continuous. Note that the cost function $C(\cdot)$ defines a smooth function, i.e.,

$$C(\cdot) \in \bigcap_{\alpha \in \mathbb{N}_0^D} \mathcal{C}^\alpha =: \mathcal{C}^\infty. \tag{E2}$$

Indeed, this immediately follows from the fact that the vector-valued mapping $\theta \mapsto |\psi(\theta)\rangle = U(\theta)|0\rangle$ is smooth [since $U(\theta)$ is a product involving smooth factors $\exp(-i\theta_k/2\, V_k)$ as per Eq. (4)], and that $C(\theta) = \langle\psi(\theta)|H|\psi(\theta)\rangle$ is a quadratic form in the Hilbert space $\mathcal{H}_N$ (that is, a multivariate polynomial).

The standard parameter-shift rule [47,48] shows explicitly that $C(\cdot)$ is differentiable.

Lemma E1 (Vanilla parameter-shift rule). Consider an expectation of the form in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters. Let $\theta \in [-\pi, \pi]^D$ and $k \in [D]$ be arbitrary. Then, $C(\cdot)$ is continuous, and

$$\partial_{\theta_k} C(\theta) = \frac{1}{2}\left\{C\left(\theta + \frac{\pi}{2} e_k\right) - C\left(\theta - \frac{\pi}{2} e_k\right)\right\},$$

where e_k denotes the kth unit vector.

Proof. Let $\theta \in [-\pi, \pi]^D$ and $k \in [D]$ be arbitrary. By definition of $U(\cdot)$, we may decompose

$$\begin{aligned} U(\theta) &= \prod_{j=D}^{1} U_j(\theta_j) W_j \\ &= \underbrace{\left(\prod_{j=D}^{k+1} U_j(\theta_j) W_j\right)}_{=:U_+} U_k(\theta_k) \underbrace{\left(W_k \prod_{j=k-1}^{1} U_j(\theta_j) W_j\right)}_{=:U_-} \\ &= U_+ U_k(\theta_k) U_-, \end{aligned} \tag{E3}$$

where $U_j(\theta_j) := \exp(-i\frac{\theta_j}{2} V_j)$ with each $V_j \in \{X, Y, Z\}$ specifying the Pauli-rotation axis, and $W_j \in \text{Clifford}(N)$, respectively. Note that $U_\pm$ are independent of θ_k. In particular,

$$\begin{aligned} \partial_{\theta_k} C(\theta_k) &= \partial_{\theta_k} \langle 0|U(\theta)^\dagger H U(\theta)|0\rangle \\ &= \langle 0|U_-^\dagger \partial_{\theta_k}\{U_k(\theta_k)^\dagger U_+^\dagger H U_+ U_k(\theta_k)\} U_-|0\rangle. \end{aligned} \tag{E4}$$

Mitarai *et al.* [47] made the crucial observation that (redefining $H_+ := U_+^\dagger H U_+$)

$$[V_k, H_+] = i\left\{U_k\left(\frac{\pi}{2}\right) H_+ U_k\left(-\frac{\pi}{2}\right) - U_k\left(-\frac{\pi}{2}\right) H_+ U_k\left(\frac{\pi}{2}\right)\right\}, \tag{E5}$$

which directly follows from the fact that $U_k(\pm\frac{\pi}{2}) = \frac{1}{\sqrt{2}}(I \mp iV_k)$, obtained by applying Euler's identity $U_k(\theta_k) = \exp(-i\frac{\theta_k}{2} V_k) = \cos\frac{\theta_k}{2} I - i \sin\frac{\theta_k}{2} V_k$. Also, note that $V_k = V_k^\dagger$ and $V_k^2 = I$. Indeed, Eq. (E5) allows us to compute the derivative

$$\begin{aligned} &\partial_{\theta_k}\{U_k(\theta_k)^\dagger U_+^\dagger H U_+ U_k(\theta_k)\} \\ &\quad = \left\{\partial_{\theta_k} U_k(\theta_k)\right\}^\dagger H_+ U_k(\theta_k) + U_k(\theta_k)^\dagger H_+ \left\{\partial_{\theta_k} U_k(\theta_k)\right\} \\ &\quad = \frac{i}{2} U_k(\theta_k)^\dagger V_k H_+ U_k(\theta_k) - \frac{i}{2} U_k(\theta_k)^\dagger H_+ V_k U_k(\theta_k) \\ &\quad = \frac{i}{2} U_k(\theta_k)^\dagger [V_k, H_+] U_k(\theta_k) \\ &\quad \overset{\text{(E5)}}{=} -\frac{1}{2}\left\{U_k(\theta_k)^\dagger U_k\left(\frac{\pi}{2}\right) H_+ U_k\left(-\frac{\pi}{2}\right) U_k(\theta_k)\right. \\ &\qquad \left. - U_k(\theta_k)^\dagger U_k\left(-\frac{\pi}{2}\right) H_+ U_k\left(\frac{\pi}{2}\right) U_k(\theta_k)\right\} \\ &\quad = \frac{1}{2}\left\{U_k\left(\theta_k + \frac{\pi}{2}\right)^\dagger H_+ U_k\left(\theta_k + \frac{\pi}{2}\right)\right. \\ &\qquad \left. - U_k\left(\theta_k - \frac{\pi}{2}\right)^\dagger H_+ U_k\left(\theta_k - \frac{\pi}{2}\right)\right\}, \end{aligned} \tag{E6}$$

using $\partial_{\theta_k} U_k(\theta_k) = -\frac{i}{2} V_k U_k(\theta_k)$, and in the last equality, $U_k(\pm\theta_k)^\dagger = U_k(\mp\theta_k)$ and $U_k(\theta_k \pm \frac{\pi}{2}) = U_k(\theta_k) U_k(\pm\frac{\pi}{2})$, respectively. Returning to Eq. (E4), it now easily follows that

$$\begin{aligned} \partial_{\theta_k} C(\theta_k) &= \langle 0|U_-^\dagger \partial_{\theta_k}\{U_k(\theta_k)^\dagger H_+ U_k(\theta_k)\} U_-|0\rangle \\ &= \frac{1}{2}\langle 0|\left\{U_-^\dagger U_k\left(\theta_k + \frac{\pi}{2}\right)^\dagger U_+^\dagger H U_+ U_k\left(\theta_k + \frac{\pi}{2}\right) U_-\right. \\ &\qquad \left. - U_-^\dagger U_k\left(\theta_k - \frac{\pi}{2}\right)^\dagger U_+^\dagger H U_+ U_k\left(\theta_k - \frac{\pi}{2}\right) U_-\right\}|0\rangle \\ &= \frac{1}{2}\left\{\langle 0|U\left(\theta + \frac{\pi}{2} e_k\right)^\dagger H U\left(\theta + \frac{\pi}{2} e_k\right)|0\rangle\right. \\ &\qquad \left. - \langle 0|U\left(\theta - \frac{\pi}{2} e_k\right)^\dagger H U\left(\theta - \frac{\pi}{2} e_k\right)|0\rangle\right\} \\ &= \frac{1}{2}\left\{C\left(\theta + \frac{\pi}{2} e_k\right) - C\left(\theta - \frac{\pi}{2} e_k\right)\right\} \end{aligned} \tag{E7}$$

as desired. ∎

The next lemma generalizes the previous result to arbitrary high-order derivatives $D_\theta^\alpha C(\cdot)$, and may be considered an explicit proof of the regularity property in Eq. (E2). The α-order parameter-shift rule gives rise to the illustration in Fig. 10, which depicts three views of the α-order Clifford structure as oper Definition 1. The leftmost picture shows an equidistant grid of points $\xi_{j,\alpha} = \frac{\pi}{2}(\alpha - 2j)$ in parameter space stemming from the α-order parameter-shift rule in Eq. (12). The central one depicts the ensemble of Clifford circuits $U(\xi_{j,\alpha})$ obtained by evaluating $U(\cdot)$ at each grid point. Finally, the rightmost one shows the corresponding Heisenberg-evolved list of

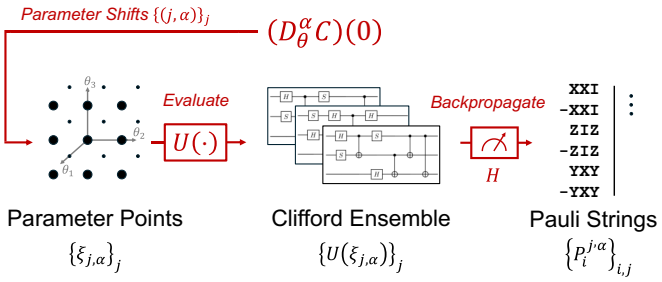


FIG. 10. Three possible interpretations of the α-order Clifford structure.

Pauli strings $P_i^{j,\alpha}$, classically encoding the α-order Clifford structure.

Lemma E2. Consider an expectation of the form Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters. Let $\theta \in [-\pi, \pi]^D$ and $\alpha \in \mathbb{N}_0^D$ be arbitrary but fixed. Then, the higher-order derivative of $C(\theta)$ satisfies

$$D_\theta^\alpha C(\theta) = \frac{1}{2^{\|\alpha\|_1}} \sum_{j \leqslant \alpha} (-1)^{\|j\|_1} \binom{\alpha}{j} C\Big(\theta + \frac{\pi}{2}(\alpha - 2j)\Big), \tag{E8}$$

where $\binom{\alpha}{j} := \prod_{k=1}^D \binom{\alpha_k}{j_k}$, and the indexing set $\{j \leqslant \alpha\}$ defined as $\{j \in \mathbb{N}_0^D : j_k \leqslant \alpha_k \, \forall k \in [D]\}$. The multi-index expressions $\alpha - 2j$ are to be understood pointwise. In particular, $(D_\theta^\alpha C)(\cdot)$ is continuous. If $\theta = 0$, the computational cost of evaluating Eq. (E8) is of order $\mathcal{O}(D\,|\mathcal{I}|\,2^{\|\alpha\|_1})$, which is polynomial if $\|\alpha\|_1 = \mathcal{O}(\log D)$. Furthermore, $C(\cdot)$ satisfies the strong regularity property that for each $\alpha \in \mathbb{N}_0^D$,

$$\sup_{\theta \in [-\pi,\pi]^D} |D_\theta^\alpha C(\theta)| \leqslant \|H\|,$$

where $\|H\|$ denotes the operator norm of the observable in Eq. (5).

Proof. Let $\theta \in [-\pi, \pi]^D$ and $\alpha \in \mathbb{N}_0^D$ be arbitrary. By the vanilla parameter-shift rule, Lemma E1, we know that

$$\partial_{\theta_k} C(\theta) = \frac{1}{2}\Big\{C\Big(\theta + \frac{\pi}{2}e_k\Big) - C\Big(\theta - \frac{\pi}{2}e_k\Big)\Big\}, \tag{E9}$$

for all $k \in [D]$, where e_k denotes the kth unit vector. This is precisely the formula for the central finite difference of $C(\cdot)$. Inductively over $\alpha_k \geqslant 1$, one easily establishes the higher-order generalization of the central finite difference formula, i.e.,

$$\partial_{\theta_k}^{\alpha_k} C(\theta) = \frac{1}{2^{\alpha_k}} \sum_{j_k=0}^{\alpha_k} (-1)^{j_k} \binom{\alpha_k}{j_k} C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k\Big), \tag{E10}$$

for all $k \in [D]$. Indeed, the case $\alpha_k = 1$ follows directly from Eq. (E9). Assume that Eq. (E10) is true for some fixed $\alpha_k \geqslant 1$. Then, by induction,

$$\begin{aligned}
\partial_{\theta_k}^{\alpha_k+1} C(\theta) &= \partial_{\theta_k}\big\{\partial_{\theta_k}^{\alpha_k} C(\theta)\big\} \overset{\text{(E10)}}{=} \frac{1}{2^{\alpha_k}} \sum_{j_k=0}^{\alpha_k} (-1)^{j_k} \binom{\alpha_k}{j_k} \partial_{\theta_k} C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k\Big) \\
&= \frac{1}{2^{\alpha_k+1}} \sum_{j_k=0}^{\alpha_k} (-1)^{j_k} \binom{\alpha_k}{j_k} \Big\{C\Big(\theta + \pi\Big(\frac{\alpha_k+1}{2} - j_k\Big)e_k\Big) - C\Big(\theta - \pi\Big(\frac{\alpha_k+1}{2} - j_k\Big)e_k\Big)\Big\} \\
&= \frac{1}{2^{\alpha_k+1}} \sum_{j_k=0}^{\alpha_k+1} (-1)^{j_k} \binom{\alpha_k+1}{j_k} C\Big(\theta + \pi\Big(\frac{\alpha_k+1}{2} - j_k\Big)e_k\Big),
\end{aligned} \tag{E11}$$

where the second equality makes use of Eq. (E9) as follows:

$$\begin{aligned}
\partial_{\theta_k} C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k\Big) &\overset{\text{(E9)}}{=} \frac{1}{2}\Big\{C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k + \frac{\pi}{2}e_k\Big) - C\Big(\theta - \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k - \frac{\pi}{2}e_k\Big)\Big\} \\
&= \frac{1}{2}\Big\{C\Big(\theta + \pi\Big(\frac{\alpha_k+1}{2} - j_k\Big)e_k\Big) - C\Big(\theta - \pi\Big(\frac{\alpha_k+1}{2} - j_k\Big)e_k\Big)\Big\}.
\end{aligned} \tag{E12}$$

Next, note that for each pair $k' \neq k$ in $[D]$,

$$\begin{aligned}
\partial_{\theta_{k'}}^{\alpha_{k'}} \partial_{\theta_k}^{\alpha_k} C(\theta) &= \frac{1}{2^{\alpha_k}} \sum_{j_k=0}^{\alpha_k} (-1)^{j_k} \binom{\alpha_k}{j_k} \partial_{\theta_{k'}}^{\alpha_{k'}} C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k\Big) \\
&= \frac{1}{2^{\alpha_k}} \sum_{j_k=0}^{\alpha_k} (-1)^{j_k} \binom{\alpha_k}{j_k} \frac{1}{2^{\alpha_{k'}}} \sum_{j_{k'}=0}^{\alpha_{k'}} (-1)^{j_{k'}} \binom{\alpha_{k'}}{j_{k'}} C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k + \pi\Big(\frac{\alpha_{k'}}{2} - j_{k'}\Big)e_{k'}\Big) \\
&= \frac{1}{2^{\alpha_k+\alpha_{k'}}} \sum_{j_k=0}^{\alpha_j} \sum_{j_{k'}=0}^{\alpha_{k'}} (-1)^{j_k+j_{k'}} \binom{\alpha_k}{j_k} \binom{\alpha_{k'}}{j_{k'}} C\Big(\theta + \pi\Big(\frac{\alpha_k}{2} - j_k\Big)e_k + \pi\Big(\frac{\alpha_{k'}}{2} - j_{k'}\Big)e_{k'}\Big).
\end{aligned} \tag{E13}$$

Applying these calculations successively for each $k \in [D]$, one finally obtains the multivariate higher-order formula

$$D_\theta^\alpha C(\theta) = \partial_{\theta_1}^{\alpha_1} \cdots \partial_{\theta_D}^{\alpha_D} C(\theta) = \frac{1}{2^{\alpha_1+\cdots+\alpha_D}} \sum_{j_1=0}^{\alpha_1} \cdots \sum_{j_D=0}^{\alpha_D} (-1)^{j_1+\cdots+j_D} \binom{\alpha_1}{j_1} \cdots \binom{\alpha_D}{j_D} C\left(\theta + \pi \sum_{k=1}^{D} \left(\frac{\alpha_k}{2} - j\right) e_k\right)$$

$$= \frac{1}{2^{\|\alpha\|_1}} \sum_{\substack{j \in \mathbb{N}_0^D \\ j_k \leqslant \alpha_k \, \forall k}} (-1)^{|j|} \binom{\alpha}{j} C\left(\theta + \frac{\pi}{2} \sum_{k=1}^{D} (\alpha_k - 2j_k) e_k\right) \tag{E14}$$

of the central finite difference, as desired.

In order to show the uniform bound, let $j \in \mathbb{N}_0^D$ be arbitrary with $j_k \leqslant \alpha_k$ for all $k \in [D]$. Let us define the shifted parameter $\theta_{j,\alpha} := \theta + \frac{\pi}{2} \sum_{k=1}^{D} (\alpha_k - 2j_k) e_k$. Then, we apply submultiplicativity of norms in order to establish the bound

$$|C(\theta_{j,\alpha})| = \left|\left\langle \psi_{\theta_{j,\alpha}} \middle| H \middle| \psi_{\theta_{j,\alpha}} \right\rangle\right| \leqslant \|H\| \underbrace{\left|\left\langle \psi_{\theta_{j,\alpha}} \middle| \psi_{\theta_{j,\alpha}} \right\rangle\right|}_{=1} = \|H\|, \tag{E15}$$

using the notation $|\psi_{\theta_{j,\alpha}}\rangle = U(\theta_{j,\alpha})|0\rangle$. Since this holds for arbitrary j and α as above, by the previous computations, we may conclude with the uniform bound

$$\left|D_\theta^\alpha C(\theta)\right| \leqslant \frac{1}{2^{|\alpha|}} \sum_{\substack{j \in \mathbb{N}_0^D \\ j_k \leqslant \alpha_k \, \forall k}} \binom{\alpha}{j} \underbrace{\left|C\left(\theta + \frac{\pi}{2} \sum_{k=1}^{D} (\alpha_k - 2j_k) e_k\right)\right|}_{= |C(\theta_{j,\alpha})| \leqslant \|H\|} \leqslant \|H\| \frac{1}{2^{|\alpha|}} \sum_{\substack{j \in \mathbb{N}_0^D \\ j_k \leqslant \alpha_k \, \forall k}} \binom{\alpha}{j} = \|H\|, \tag{E16}$$

where in the last step follows from the binomial theorem, i.e.,

$$\sum_{\substack{j \in \mathbb{N}_0^D \\ j_k \leqslant \alpha_k \, \forall k}} \binom{\alpha}{j} = \sum_{j_1=0}^{\alpha_1} \cdots \sum_{j_D=0}^{\alpha_D} \prod_{k=1}^{D} \binom{\alpha_k}{j_k} = \prod_{k=1}^{D} \underbrace{\left(\sum_{j_k=0}^{\alpha_k} \binom{\alpha_k}{j_k}\right)}_{=(1+1)^{\alpha_k}} = \prod_{k=1}^{D} 2^{\alpha_k} = 2^{|\alpha|}. \tag{E17}$$

Since $\theta \in \mathbb{R}^D$ and $\alpha \in \mathbb{N}_0^D$ were arbitrary, this completes the first part of the lemma.

It remains to show that, if $\theta = 0$, then the computational cost of evaluating Eq. (E8) is of order $\mathcal{O}(D\,|\mathcal{I}|\,2^{\|\alpha\|_1})$. Each evaluation of $C(\xi_{j,\alpha})$ using Gottesman-Knill's theorem requires $\mathcal{O}(D\,|\mathcal{I}|)$ operations since all unitaries $U(\xi_{j,\alpha})$ involved are Clifford operators (cf. last part of the proof of Theorem 3 in Appendix J). The number of times we need to evaluate $C(\xi_{j,\alpha})$ is given by the cardinality

$$\#\{j \leqslant \alpha\} = \#\left\{j \in \mathbb{N}_0^D : j_k \leqslant \alpha_k \, \forall k \in [D]\right\} = \prod_{k=1}^{D} \underbrace{(\alpha_k + 1)}_{\leqslant 2^{\alpha_k}} \leqslant 2^{\sum_{k=1}^{D} \alpha_k} = 2^{\|\alpha\|_1}. \tag{E18}$$

Therefore, the total cost of evaluating Eq. (E8) is indeed $\mathcal{O}(D\,|\mathcal{I}|\,2^{\|\alpha\|_1})$. ■

APPENDIX F: CONSTRUCTION OF THE LINEAR CLIFFORD ENCODER

We now construct the LCE transformation used to mitigate the barren-plateau problem [11]. The approach involves modifying the PQC via a pair $(Q, \tilde{Q})$ of Clifford operators such that the kth gradient component satisfies $(\partial_{\theta_k} C)(0) = \pm 1$, ensuring it is nonzero at the origin. This is accomplished by propagating the shifted Clifford operators $U(\xi_{j,\alpha})$ through the Pauli observable $H = P$, yielding diagonal Pauli strings $\tilde{P}_j := U(\xi_{j,\alpha})^\dagger P\, U(\xi_{j,\alpha})$ with a relative phase. Afterward, we generalize this result to arbitrary, nonidentity observables of the general form

$$H = \sum_{i \in \mathcal{I}} c_i P_i, \tag{F1}$$

with Pauli strings $P_i \in \mathcal{P}_N \setminus \{I\}$ and real coefficients $c_i \in \mathbb{R}$.

In what follows, we use $\langle A, B\rangle_N$ to denote the multiplicative subgroup of the unitary group $\mathrm{U}(2^N)$, generated by the matrices A and B, modulo the global phase factors $\pm, \pm i$. For example, it is well known [26] that the Clifford group is equal to $\langle H, S, \mathrm{CX}\rangle_N$. In practice, this means that a Clifford circuit is constructed from elementary gates in the ensemble $\{H, S, \mathrm{CX}\}$. As another example, recall that the N-qubit Pauli group is defined as

$$\mathcal{P}_N := \langle X, Y, Z\rangle_N = \{\pm P, \pm iP : P = P_1 \otimes \cdots \otimes P_N,\ \forall j \in [N] : P_j \in \{I, X, Y, Z\}\}. \tag{F2}$$

We generally consider

$$U(\cdot) \in \langle R_X(\cdot), R_Y(\cdot), R_Z(\cdot), \mathrm{Clifford}(N)\rangle_N \tag{F3}$$

to be generated by single-qubit rotation gates and Clifford operators, $\mathrm{Clifford}(N) = \langle H, S, \mathrm{CX}\rangle_N$, acting on an N-qubit system. It is well known that this defines a universal gate set.

1. Explicit construction of the linear Clifford encoder

The following theorem shows how to explicitly construct the LCE pair $(Q, \tilde{Q})$, which plays a central role in our barren-plateau mitigation technique. We begin by handling the case where the observable $H = P$ is a single Pauli string. Afterward, we show that the same construction may be used for arbitrary observables.

Theorem F1. Consider an expectation of the form [Eq. (E1)] with a variational ansatz $U(\theta)$ consisting of Clifford gates and single-qubit Pauli-rotation gates involving $D \geqslant 1$ parameters. Assume that the observable $H = P$ in Eq. (F1) is a single Pauli string. Then, for every $k \in [D]$ there exist two explicitly determined Clifford operators $Q \in \langle H, S, \mathrm{CX}\rangle_N$ and $\tilde{Q} \in \langle H, S\rangle_N$ such that

$$\big(\partial_{\theta_k}\tilde{C}\big)(0) = \tfrac{1}{2}\big\{\langle 0|Q^\dagger \tilde{P}^{0,e_k} Q|0\rangle - \langle 0|Q^\dagger \tilde{P}^{e_k,e_k} Q|0\rangle\big\} = \pm 1,$$

after replacing $U(\theta)$ with $\tilde{Q}\,U(\theta)\,Q$, where $\tilde{P}^{j,\alpha} := U(\xi_{j,\alpha})^\dagger \tilde{Q}^\dagger P\, \tilde{Q}\, U(\xi_{j,\alpha})$, and $\xi_{j,\alpha} := \frac{\pi}{2}(\alpha - 2j)$ for multi-indices $\alpha \in \mathbb{N}_0^D$ and $0 \leqslant j \leqslant \alpha$ (i.e., $0 \leqslant j_k \leqslant \alpha_k$ for all $k \in [D]$).

Proof. Let $k \in [D]$ be arbitrary, and the Clifford pair $(Q, \tilde{Q})$ to be determined.

Step 1: Higher-order parameter-shift rule. First, we apply the higher-order parameter-shift rule, Lemma E2, with $\xi_{0,e_k} = \frac{\pi}{2}e_k$ and $\xi_{e_k,e_k} = -\frac{\pi}{2}e_k$, which results in

$$\begin{aligned}\big(\partial_{\theta_k}\tilde{C}\big)(0) &= \tfrac{1}{2}\big\{\tilde{C}\big(\xi_{0,e_k}\big) - \tilde{C}\big(\xi_{e_k,e_k}\big)\big\}\\ &= \tfrac{1}{2}\big\{\langle 0|Q^\dagger U\big(\xi_{0,e_k}\big)^\dagger \tilde{Q}^\dagger P\, \tilde{Q}\, U\big(\xi_{0,e_k}\big) Q|0\rangle\\ &\quad - \langle 0|Q^\dagger U\big(\xi_{e_k,e_k}\big)^\dagger \tilde{Q}^\dagger P\, \tilde{Q}\, U\big(\xi_{e_k,e_k}\big) Q|0\rangle\big\}\\ &= \tfrac{1}{2}\big\{\langle 0|Q^\dagger \tilde{P}^{0,e_k} Q|0\rangle - \langle 0|Q^\dagger \tilde{P}^{e_k,e_k} Q|0\rangle\big\}. \quad \text{(F4)}\end{aligned}$$

The idea is to engineer the pair $(Q, \tilde{Q})$ of Clifford operators as follows: First, we design $\tilde{Q}$ explicitly for inducing a relative phase, i.e.,

$$\operatorname{sign}\tilde{P}^{0,e_k} \overset{!}{\neq} \operatorname{sign}\tilde{P}^{e_k,e_k}. \quad \text{(F5)}$$

Here, the notation "!" means that we would like to achieve the desired relation by adjusting free parameters (in that case $\tilde{Q}$). Next, after determining $\tilde{Q}$, we explicitly design Q such that it simultaneously diagonalizes $\tilde{P}^{0,e_k}$ and $\tilde{P}^{e_k,e_k}$. That is,

$$Q^\dagger \tilde{P}^{0,e_k} Q,\ Q^\dagger \tilde{P}^{e_k,e_k} Q \overset{!}{\in} \langle Z\rangle_N. \quad \text{(F6)}$$

Engineering the pair $(Q, \tilde{Q})$ that way thus implies

$$\begin{aligned}\langle 0|Q^\dagger \tilde{P}^{0,e_k} Q|0\rangle - \langle 0|Q^\dagger \tilde{P}^{e_k,e_k} Q|0\rangle &\overset{\text{(F6)}}{=} \operatorname{sign}\tilde{P}^{0,e_k} - \operatorname{sign}\tilde{P}^{e_k,e_k}\\ &\overset{\text{(F5)}}{=} \pm 2, \quad \text{(F7)}\end{aligned}$$

which, substituted into Eq. (F4), results in the desired equality.

Step 2: Relative phase via $\tilde{Q}$. Let us begin by designing $\tilde{Q} \in \langle H, S, \mathrm{CX}\rangle_N$. In general, since $U(\cdot)$ is generated according to Eq. (F3), we may decompose

$$U(\xi_{j,e_k}) = W_a R_{V_k}\Big(\pm\frac{\pi}{2}e_k\Big)W_b \quad \text{(F8)}$$

for some Clifford operators $W_a, W_b \in \langle H, S, \mathrm{CX}\rangle_N$. Thus, evolving the input Pauli string P through the Clifford $\tilde{Q}\,U(\xi_{j,e_k})$ in the Heisenberg picture yields the evolved Pauli string,

$$\begin{aligned}\tilde{P}^{j,e_k} &= U(\xi_{j,\alpha})^\dagger \tilde{Q}^\dagger P\, \tilde{Q}\, U(\xi_{j,\alpha})\\ &= W_b^\dagger R_{V_k}\Big(\pm\frac{\pi}{2}e_k\Big)^\dagger W_a^\dagger\, \tilde{Q}^\dagger P\, \tilde{Q}\, W_a\, R_{V_k}\Big(\pm\frac{\pi}{2}e_k\Big)W_b. \quad \text{(F9)}\end{aligned}$$

The only way to obtain a relative phase, specifically different signs as per Eq. (F5), is to design $\tilde{Q}$ systematically such that

$$R_{V_k}\Big(\pm\frac{\pi}{2}e_k\Big)^\dagger W_a^\dagger\, \tilde{Q}^\dagger P\, \tilde{Q}\, W_a\, R_{V_k}\Big(\pm\frac{\pi}{2}e_k\Big) \in \{\pm\Pi, \mp\Pi\} \quad \text{(F10)}$$

for some unsigned Pauli string $\Pi \in \mathcal{P}_N$.

With the help of Table V, one establishes that for every pair (V, Π_1) of unsigned, single-qubit Pauli-operators, there exists an unsigned Pauli operator Π_2 such that $\Pi_1 \neq \Pi_2$ and

$$\begin{aligned}R_V\Big(\pm\frac{\pi}{2}\Big)^\dagger \Pi_1\, R_V\Big(\pm\frac{\pi}{2}\Big) &\in \{\pm\Pi_2, \mp\Pi_2\}\\ &\text{if and only if } \Pi_1 \notin \{I, V\}. \quad \text{(F11)}\end{aligned}$$

We make the ansatz $\tilde{Q} := \tilde{W} W_a^\dagger$ for some Clifford $\tilde{W}$ to be determined, and it will follow that Eq. (F10) holds true if and only if

$$(\tilde{W}^\dagger P \tilde{W})_k \notin \{I, V_k\}. \quad \text{(F12)}$$

Indeed, the factor $W_a^\dagger$ is used to cancel with W_a, and the Clifford operator $\tilde{W}$ remains to be designed such that a relative phase is induced as per Eq. (F5).

We briefly recall the indexing convention for Pauli operators: For a Pauli string $P \in \mathcal{P}_N$, we use the notation $P_k \in \{I, X, Y, Z\}$ to denote the kth component of the tensor product $P = P_1 \otimes \cdots \otimes P_N$. Equivalently, P_k may be interpreted as the N-qubit operator $I_{2^{k-1}} \otimes P_k \otimes I_{2^{N-k}}$. For 2×2 operators A (like H and S), we analogously denote $A_k := I_{2^{k-1}} \otimes A \otimes I_{2^{N-k}}$. These conventions are used to simplify our construction. The goal is now to construct $\tilde{W}$ such that Eq. (F12) is satisfied.

Recall that we have fixed $k \in [D]$ for the LCE construction. First, consider the case where $P_k \neq I$. If $V_k \neq P_k$, then Eq. (F12) holds true with the identity $\tilde{W} := I$. If $V_k = P_k$, then define

$$\tilde{W} := \begin{cases} H_k, & V_k \in \{X, Z\},\\ S_k, & V_k = Y, \end{cases} \quad \text{(F13)}$$

and it follows from Table VII that Eq. (F12) is satisfied.

For the case where $P_k = I$, we are going to decompose $\tilde{W} := \tilde{W}_1\tilde{W}_2$ into two Clifford operators $\tilde{W}_1 \in \langle \mathrm{CX}\rangle_N$ and $\tilde{W}_2 \in \langle H, S\rangle_N$, respectively. Due to the fact that $P \neq I$ is

TABLE V. Update rules for conjugating Pauli strings Π with Clifford operators C.

Clifford C	Input Pauli $\Pi \to$ output Pauli $C^\dagger \Pi C$		
$R_X(\pm\frac{\pi}{2})$	$X \to X$	$Y \to \mp Z$	$Z \to \pm Y$
$R_Y(\pm\frac{\pi}{2})$	$X \to \pm Z$	$Y \to Y$	$Z \to \mp X$
$R_Z(\pm\frac{\pi}{2})$	$X \to \mp Y$	$Y \to \pm X$	$Z \to Z$

TABLE VI. Update rules for conjugating Pauli strings R with Clifford operators C.

Clifford C	Input Pauli $R \to$ output Pauli $C^\dagger R C$		
$\mathrm{CX}_{i,j}$	$X_i \to X_i X_j$	$Y_i \to Y_i X_j$	$Z_i \to Z_i$
	$X_j \to X_j$	$Y_j \to Z_i Y_j$	$Z_j \to Z_i Z_j$

nontrivial, there exists $i \in [N]$ such that $P_i \neq I$. Define

$$\tilde{W}_1 := \begin{cases} \mathrm{CX}_{i,k}, & P_i \in \{X, Y\}, \\ \mathrm{CX}_{k,i}, & P_i = Z, \end{cases} \tag{F14}$$

where the entangling gate $\mathrm{CX}_{i,j}$ has control i and target j.

With the help of Table VI, we see that

$$(\tilde{W}_1^\dagger P \tilde{W}_1)_k = \left\{ \begin{matrix} X_k, & P_i \in \{X, Y\} \\ Z_k, & P_i = Z, \end{matrix} \right\} \neq I. \tag{F15}$$

Hence, we may construct $\tilde{W}_2$ analogous to the previous case, $P_k \neq I$. That is, if $V_k \neq (\tilde{W}_1^\dagger P \tilde{W}_1)_k$, then Eq. (F12) holds true with $\tilde{W}_2 = I$. Otherwise, if $V_k = (\tilde{W}_1^\dagger P \tilde{W}_1)_k$, we may define

$$\tilde{W}_2 := \begin{cases} H_k, & V_k \in \{X, Z\}, \\ S_k, & V_k = Y, \end{cases} \tag{F16}$$

and it follows from Table VII that Eq. (F12) is satisfied.

Step 3: Simultaneous diagonalization via Q. It remains to design $Q \in \langle H, S \rangle_N$. We make the ansatz $Q = Q_1 \otimes \cdots \otimes Q_N$ and define

$$Q_i := \begin{cases} H, & \tilde{P}_i^{0,e_k} = X, \\ SH, & \tilde{P}_i^{0,e_k} = Y, \\ I, & \text{else}. \end{cases} \tag{F17}$$

With the help of Table VII, one obtains that

$$Q^\dagger \tilde{P}^{0,e_k} Q = \bigotimes_{i=1}^{N} \underbrace{Q_i^\dagger \tilde{P}_i^{0,e_k} Q_i}_{\in \{I, \pm Z\}} \in \langle Z \rangle_N \tag{F18}$$

is diagonal. Since $\tilde{P}^{0,e_k}$ and $\tilde{P}^{e_k,e_k}$ only differ by the sign, we also have that $Q^\dagger \tilde{P}^{e_k,e_k} Q \in \langle Z \rangle_N$ is diagonal.

We have thus successfully constructed the pair $(Q, \tilde{Q})$ of Clifford operators such that Eqs. (F5) and (F6) are simultaneously satisfied. Therefore, by Eq. (F4), we have established that

$$\begin{aligned} \big(\partial_{\theta_k} \tilde{C}\big)(0) &= \tfrac{1}{2} \big\{ \langle 0 | Q^\dagger \tilde{P}^{0,e_k} Q | 0 \rangle - \langle 0 | Q^\dagger \tilde{P}^{e_k,e_k} Q | 0 \rangle \big\} \\ &\overset{(F6)}{=} \tfrac{1}{2} \big\{ \operatorname{sign} \tilde{P}^{0,e_k} - \operatorname{sign} \tilde{P}^{e_k,e_k} \big\} \\ &\overset{(F5)}{=} \pm 1, \end{aligned} \tag{F19}$$

which completes the proof of the theorem. ■

TABLE VII. Update rules for conjugating Pauli strings R with Clifford operators C.

Clifford C	Input Pauli $R \to$ output Pauli $C^\dagger R C$		
H	$X \to Z$	$Y \to -Y$	$Z \to X$
S	$X \to -Y$	$Y \to X$	$Z \to Z$
SH	$X \to Y$	$Y \to Z$	$Z \to X$

We refer to the Clifford pair $(Q, \tilde{Q})$ used to transform the circuit $U(\cdot)$ into $\tilde{Q} U(\cdot) Q$ as the LCE, since it is constructed to render the kth linear Taylor coefficient constant, i.e., $(\partial_{\theta_k} C)(0) = \pm 1$.

In Fig. 11, we use a numerical experiment to illustrate geometrically how LCE transforms the loss landscape $C(\cdot) \mapsto \tilde{C}(\cdot)$. We consider a $N = 12$ qubit `mHEA` model with $L = 6$ layers, and focus on the global observable $Z^{\otimes N}$. For the LCE construction, we pick the direction $k = 1$, and evaluate the cost $\tilde{C}(\theta_1 e_1 + \theta_k e_k)$ for all non-LCE directions $k \neq 1$, where e_k denotes the kth unit vector. This results in an equidistant grid $\{\theta_1 e_1 + \theta_k e_k\}$ (we use 30×30 points) on which we evaluate the cost $C(\cdot)$. Then, we average over all non-LCE directions, i.e., $\mathbb{E}_{k \in [D] \setminus 1}[\tilde{C}(\theta_1 e_1 + \theta_k e_k)]$, which are always used as the vertical axis in the plot. We can see that LCE ensures a better gradient when initializing close to $\theta = 0$, as predicted by Theorem 2. This is due to initializing at the steepest slope of the landscape [green-blue area where $\tilde{C}(\theta) \approx 0$], which ensures a gradient "boost" in the positive LCE-direction (which could be exploited for momentum). The landscape appears to be shifted to the left in the LCE direction by $\pi/2$ (stemming from the operator Q causing a translation), and then slightly stretched in the non-LCE directions (stemming from the operator $\tilde{Q}$ causing a change of basis). In the case of $Z^{\otimes N}$, this transformation guarantees faster convergence to the global minimum value -1 when directly employing gradient descent with sufficiently small learning rates (cf. Fig. 6). On the other hand, the non-LCE loss landscape has a global maximum at $\theta = 0$, showing that using vanilla gradient descent requires more iterations compared to LCE. In fact, using LCE in this situation even results in closer parameter initialization to the solution.

We have explained how to make the gradient constant in one direction. This naturally raises the question of whether multiple gradient components can be controlled simultaneously. In principle, this is possible to some extent: One must solve an over-determined system of operator equations. As the number of components increases, the likelihood of a solution decreases. The following elaborates the details.

Consider a collection $\mathcal{K} \subseteq [D]$ of gradient elements we would like to control. We would like to construct a stronger LCE $(Q, \tilde{Q})$ in the sense that

$$\big(\partial_{\theta_k} \tilde{C}\big)(0) = \pm 1 \quad \text{for all } k \in \mathcal{K}, \tag{F20}$$

after replacing $U(\theta)$ with $\tilde{Q} U(\theta) Q$. To see how this works, for each $k \in \mathcal{K}$ we decompose

$$U\big(\xi_{j,e_k}\big) = W_{a_k} R_{V_k}\Big(\pm \frac{\pi}{2} e_k\Big) W_{b_k}, \tag{F21}$$

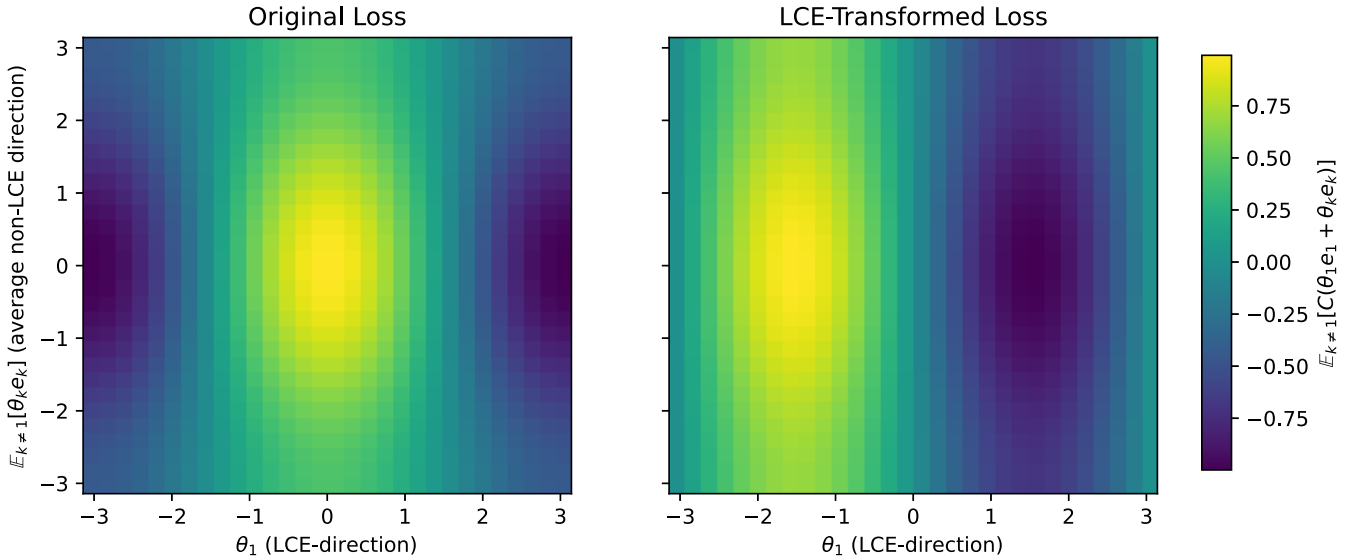


FIG. 11. Illustrating the landscape transformation effect of LCE for an `mHEA` model with $N = 12$ qubits, $L = 6$ layers, and $H = Z^{\otimes N}$ as the observable. We fix $k = 1$ as the LCE direction, and average the grids $\{\tilde{C}(\theta_1 e_1 + \theta_k e_k) : \theta_1, \theta_k \in [-\pi, \pi]\}$ over non-LCE directions $k \neq 1$ (always used as the vertical axis).

for some Clifford operators $W_{a_k}, W_{b_k} \in \langle H, S, \mathrm{CX}\rangle_N$. The solution (if it exists) is determined by the system of $2\#\mathcal{K}$ operator equations:

$$\left\{ \begin{array}{l} \left(W_{a_k}^\dagger \tilde{Q}^\dagger P\, \tilde{Q}\, W_{a_k}\right)_k \notin \{I, V_k\} \\ Q^\dagger W_{b_k}^\dagger R_{V_k}\left(\pm\frac{\pi}{2} e_k\right)^\dagger W_{a_k}^\dagger\, \tilde{Q}^\dagger P\, \tilde{Q}\, W_{a_k}\, R_{V_k}\left(\pm\frac{\pi}{2} e_k\right) W_{b_k}\, Q \in \langle Z\rangle_N \end{array} : k \in \mathcal{K} \right\}. \tag{F22}$$

The difficulty of finding a solution $(Q, \tilde{Q})$ in Eq. (F22) is influenced by the circuit ansatz first-order Clifford architecture $\{(W_{a_k}, W_{b_k}) : k \in \mathcal{K}\}$, which is obtained from the parameter shifts in Eq. (F21).

If $U(\cdot)$ exhibits additional structure, for example, like a hardware-efficient, alternating ansatz (such as `EfficientSU2` [59]), and if $\mathcal{K}$ chosen such that all V_k are identical and in the same layer, then Eq. (F22) reduces to a system of two equations [since $W_{a_k} = W_a$ and $W_{b_k} = W_b$ for all $k \in \mathcal{K}$ for some fixed $W_a, W_b \in \mathrm{Clifford}(N)$ independent of k], and we may use the algorithm established in Theorem 1 to construct an explicit solution $(Q, \tilde{Q})$. That is, for alternating ansätze of this form, we have $\#\mathcal{K} \geqslant N$ since there are N-many identical Pauli rotations in a fixed layer.

If we impose more conditions on $U(\cdot)$, it is possible to achieve even larger initial gradients: Suppose the PQC is of the form

$$U(\theta) = \prod_{\ell=L}^{1} W_\ell U_\ell(\theta_\ell) W_\ell^\dagger, \tag{F23}$$

for some $W_\ell \in \mathrm{Clifford}(N)$ and $U_\ell(\theta_\ell) := R_{P^{\otimes N}}(\theta_\ell) R_{S^{\otimes N}}(\theta_\ell)$, a product of two single layers of Pauli rotations generated by $P, S \in \{X, Y, Z\}$ with $P \neq S$ [e.g., a layer of Y-Z rotations as in Fig. 7(a)], where $\theta_\ell \in [-\pi, \pi]^{2N}$ are parameter subvectors of θ independent over $\ell \in [L]$. In that case, the number of parameters $D = 2NL$ is even, and $\theta = 0$ without LCE corresponds to the identity initialization since $W_\ell W_\ell^\dagger = I$ for all $\ell \in [L]$. On the other hand, $\theta = 0$ with LCE generally yields a Clifford initialization $U(0) = Q\tilde{Q}$.

As a result, given any gradient component $k \in [D]$, we know that there are exactly N Pauli rotation gates (with generator either P or S) living in the same fixed layer $\ell_0 \in [L]$, and crucially, we again have $W_{a_k} = W_a$ and $W_{b_k} = W_b$ for all these rotations because of $W_\ell W_\ell^\dagger = I$ for all $\ell \neq \ell_0$. Therefore, we find that $\#\mathcal{K} \geqslant NL = \frac{D}{2}$.

2. Generalization to arbitrary observables

We showed how to render the gradient constant along one direction for $H = P$. We now extend this result to an arbitrary linear combinations of Pauli strings. Consider a general observable of the form in Eq. (F1), i.e.,

$$H = \sum_{i\in\mathcal{I}} c_i P_i. \tag{F24}$$

The linear Clifford encoder constructed in Theorem 1 may still be used in that case.

Theorem 1. Consider an expectation of the form Eq. (6) with a variational ansatz $U(\theta)$ consisting of Clifford gates and single-qubit Pauli-rotation gates involving $D \geqslant 1$ parameters. Let H be an arbitrary observable as in Eq. (5). Then, for every $k \in [D]$ and every $j \in \mathcal{I}$ there exist two explicitly determined Clifford operators $Q \in \langle H, S, \mathrm{CX}\rangle_N$ and $\tilde{Q} \in \langle H, S\rangle_N$ such that

$$\left(\partial_{\theta_k}\tilde{C}\right)(0) = \pm\, c_j + \frac{1}{2}\sum_{i\in\mathcal{I}\setminus\{j\}} c_i \big\{\langle 0|Q^\dagger \tilde{P}_i^{0,e_k} Q|0\rangle - \langle 0|Q^\dagger \tilde{P}_i^{e_k,e_k} Q|0\rangle\big\}, \tag{F25}$$

after replacing $U(\theta)$ with $\tilde{Q}\,U(\theta)\,Q$, where, for each $i \in \mathcal{I}$, $\tilde{P}_i^{j,\alpha} := U(\xi_{j,\alpha})^\dagger \tilde{Q}^\dagger P_i\, \tilde{Q}\, U(\xi_{j,\alpha})$, and $\xi_{j,\alpha} := \frac{\pi}{2}(\alpha - 2j)$ for multi-indices $\alpha \in \mathbb{N}_0^D$ and $0 \leqslant j \leqslant \alpha$.

Proof. Let $j \in \mathcal{I}$ be arbitrary. By Theorem F1, there exists a pair $(Q, \tilde{Q})$ of Clifford operators with the desired properties such that

$$\frac{c_j}{2}\left\{\langle 0|Q^\dagger \tilde{P}_j^{0,e_k}\, Q|0\rangle - \langle 0|Q^\dagger \tilde{P}_j^{e_k,e_k}\, Q|0\rangle\right\} = \pm c_j. \tag{F26}$$

Hence, after replacing $U(\theta)$ with $\tilde{Q}\,U(\theta)\,Q$, applying Lemma E2 with $\xi_{0,e_k} = \frac{\pi}{2}e_k$ and $\xi_{e_k,e_k} = -\frac{\pi}{2}e_k$ yields

$$\begin{aligned}
\big(\partial_{\theta_k}\tilde{C}\big)(0) &= \frac{1}{2}\left\{\tilde{C}\big(\xi_{0,e_k}\big) - \tilde{C}\big(\xi_{e_k,e_k}\big)\right\} \\
&= \frac{1}{2}\sum_{i\in\mathcal{I}} c_i\Big\{\langle 0|Q^\dagger U\big(\xi_{0,e_k}\big)^\dagger \tilde{Q}^\dagger P_i\, \tilde{Q}\, U\big(\xi_{0,e_k}\big)\, Q|0\rangle \\
&\quad - \langle 0|Q^\dagger U\big(\xi_{e_k,e_k}\big)^\dagger \tilde{Q}^\dagger P_i\, \tilde{Q}\, U\big(\xi_{e_k,e_k}\big)\, Q|0\rangle\Big\} \\
&= \frac{c_j}{2}\left\{\langle 0|Q^\dagger \tilde{P}_j^{0,e_k}\, Q|0\rangle - \langle 0|Q^\dagger \tilde{P}_j^{e_k,e_k}\, Q|0\rangle\right\} \\
&\quad + \frac{1}{2}\sum_{i\in\mathcal{I}\setminus\{j\}} c_i\Big\{\langle 0|Q^\dagger \tilde{P}_i^{0,e_k}\, Q|0\rangle - \langle 0|Q^\dagger \tilde{P}_i^{e_k,e_k}\, Q|0\rangle\Big\} \\
&\overset{(F26)}{=} \pm c_j + \frac{1}{2}\sum_{i\in\mathcal{I}\setminus\{j\}} c_i\Big\{\langle 0|Q^\dagger \tilde{P}_i^{0,e_k}\, Q|0\rangle \\
&\quad - \langle 0|Q^\dagger \tilde{P}_i^{e_k,e_k}\, Q|0\rangle\Big\},
\end{aligned} \tag{F27}$$

as claimed. ∎

APPENDIX G: ESCAPING BARREN PLATEAUS

In the previous Appendix, we introduced the LCE to ensure constant gradient directions at zero initialization. We employ this construction now to mitigate the barren-plateau problem [11]. In what follows, $\|H\|$ denotes the operator norm of the observable in Eq. (F1).

1. Taylor perturbation around Clifford points

The proposition below characterizes the perturbation of higher-order derivatives $D_\theta^\alpha C(\theta)$ around their zero evaluation $(D_\theta^\alpha C)(0)$—interpreted here as Taylor coefficients—by an error term $\mathcal{O}(\mathbb{P}_\theta)$ that depends on the initial distribution $\mathbb{P}_\theta$.

Proposition G1. Consider an expectation of the form Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters. Then, for each multi-index $\alpha \in \mathbb{N}_0^D$, we have

$$\mathbb{E}_\theta\big[D_\theta^\alpha C(\theta)\big] = \big(D_\theta^\alpha C\big)(0) + \|H\|\,\mathcal{O}(\mathbb{P}_\theta),$$

where the error term $\mathcal{O}(\mathbb{P}_\theta)$ scaling depends on the probability distribution $\mathbb{P}_\theta$.

Proof. Let $\alpha \in \mathbb{N}_0^D$ be an arbitrary multi-index. Since $C(\cdot)$ is analytic, we may use a Taylor expansion combined with the uniform bound in higher-order parameter-shift rule (Lemma E2), finding

$$\begin{aligned}
\mathbb{E}_\theta\big[D_\theta^\alpha C(\theta)\big] &= \sum_{\|\beta\|_1\geqslant 0} \underbrace{\big(D_\theta^{\alpha+\beta}C\big)(0)}_{=\mathcal{O}(\|H\|)} \frac{\mathbb{E}_\theta[\theta^{\alpha+\beta}]}{\alpha!\beta!} \\
&= \big(D_\theta^\alpha C\big)(0) + \mathcal{O}(\|H\|) \sum_{\|\beta\|_1>0} \frac{\mathbb{E}_\theta[\theta^{\alpha+\beta}]}{\alpha!\beta!} \\
&= \big(D_\theta^\alpha C\big)(0) + \|H\|\,\mathcal{O}(\mathbb{P}_\theta),
\end{aligned} \tag{G1}$$

where we separated the zero term $(D_\theta^\alpha C)(0)$ (i.e., $\{\|\beta\|_1 = 0\}$) from the remaining terms in the summation over $\{\|\beta\|_1 > 0\}$, for which we applied Lemma E2. Note that the bound $(D_\theta^{\alpha+\beta}C)(0) = \mathcal{O}(\|H\|)$ also follows from Lemma E2. We defined the asymptotic error term

$$\mathcal{O}(\mathbb{P}_\theta) := \mathcal{O}\left(\sum_{\|\beta\|_1>0} \frac{\mathbb{E}_\theta[\theta^{\alpha+\beta}]}{\alpha!\beta!}\right), \tag{G2}$$

which has a scaling depending on the probability distribution $\mathbb{P}_\theta$.

Similar to Proposition G1, the following theorem lower bounds the scaling of the expected squared gradient norm $\nabla_\theta C(\theta)$ under local Gaussian/uniform initialization $\mathbb{P}_\theta$ of the parameters. We generally assume that $\sigma < 1$.

Theorem G1. Consider an expectation of the form [Eq. (E1)] with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters. Let $\mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ be a probability distribution with i.i.d. components $\theta_1, \ldots, \theta_D$. Then, for every $k \in [D]$,

$$\begin{aligned}
\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] &\geqslant \mathbb{E}_\theta\big[\big(\partial_{\theta_k} C(\theta)\big)^2\big] \\
&= \big(\partial_{\theta_k} C\big)(0)^2 + \|H\|^2\,\mathcal{O}(D\sigma^2).
\end{aligned}$$

Proof. Let $k \in [D]$ be arbitrary. We may expand the gradient using the Taylor series [Eq. (10)],

$$\begin{aligned}
\partial_{\theta_k} C(\theta) &= \big(\partial_{\theta_k} C\big)(0) + \sum_{\|\alpha\|_1>0} \underbrace{\big(D_\theta^{\alpha+e_k} C\big)(0)}_{=\mathcal{O}(\|H\|)} \frac{\theta^\alpha}{\alpha!} \\
&= \big(\partial_{\theta_k} C\big)(0) + \mathcal{O}(\|H\|) \sum_{\|\alpha\|_1>0} \frac{\theta^\alpha}{\alpha!},
\end{aligned} \tag{G3}$$

where we made use of the strong regularity property in Lemma E2, and using the observation that $\partial_{\theta_k} D_\theta^\alpha = D_\theta^{\alpha+e_k}$ for all $\alpha \in \mathbb{N}_0^D$. Taking the expectation then involves the computation of the well-known moments (δ denoting the Dirac delta function):

$$\begin{aligned}
\mathbb{E}_\theta[\theta^\alpha] &= \delta_{\{\alpha_k \in 2\mathbb{N}\ \forall k\}}(\alpha) \\
&\quad \times \begin{cases} (\alpha-1)!!\,\sigma^\alpha, & \theta \sim \mathcal{N}(0, \sigma^2 I_D), \\ \sigma^\alpha \prod_{k=1}^D (\alpha_k+1)^{-1}, & \theta \sim \mathrm{Unif}([-\sigma,\sigma]^D), \end{cases}
\end{aligned} \tag{G4}$$

which uses the assumption that $\theta_1, \ldots, \theta_D$ are independent. Again, we use the canonical notations for $\sigma^\alpha := \sigma^{\alpha_1}\cdots\sigma^{\alpha_D}$ and $(\alpha-1)!! := \prod_{k\in[D]}(\alpha_k - 1)!!$. This particularly implies that

$$\sum_{\|\alpha\|_1>0} \frac{1}{\alpha!}\mathbb{E}_\theta[\theta^\alpha] = \mathcal{O}(D\sigma^2) = \sum_{\|\alpha\|_1,\|\beta\|_1>0} \frac{1}{\alpha!\,\beta!}\mathbb{E}_\theta[\theta^{\alpha+\beta}], \tag{G5}$$

noting that there are exactly D many multi-indices $\alpha \in \mathbb{N}_0^D$ for which the error bound $\mathbb{E}_\theta[\theta^\alpha] \overset{(G4)}{=} \mathcal{O}(\sigma^2)$ holds true (namely,

$\{\alpha = 2e_k : k \in [D]\}$). Note that the higher-order terms $\mathcal{O}(\sigma^r)$ with $r > 2$ are dominated by $\mathcal{O}(\sigma^2)$ since $\sigma < 1$. As a result, we find

$$\begin{aligned}\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] &= \sum_{j=1}^{D} \mathbb{E}_\theta\big[\big(\partial_{\theta_j} C(\theta)\big)^2\big] \\ &\geqslant \mathbb{E}_\theta\big[\big(\partial_{\theta_k} C(\theta)\big)^2\big] \\ &\overset{(G3)}{=} \big(\partial_{\theta_k} C\big)(0)^2 + 2\underbrace{\big(\partial_{\theta_k} C\big)(0)}_{=\mathcal{O}(\|H\|)}\mathcal{O}(\|H\|) \\ &\quad \times \sum_{\|\alpha\|_1>0} \frac{\mathbb{E}_\theta[\theta^\alpha]}{\alpha!} + \mathcal{O}(\|H\|^2) \\ &\quad \times \sum_{\|\alpha\|_1,\|\beta\|_1>0} \frac{\mathbb{E}_\theta[\theta^{\alpha+\beta}]}{\alpha!\,\beta!} \\ &\overset{(G5)}{=} \big(\partial_{\theta_k} C\big)(0)^2 + \|H\|^2\,\mathcal{O}(D\sigma^2), \qquad \text{(G6)}\end{aligned}$$

as desired. ■

2. Constant gradient lower bound in small patches

Having derived a lower bound for the expected squared gradient norm with an error term perturbing a given component $(\partial_{\theta_k} C)(0)$, we combine this result with the LCE (cf. Theorem 1) to establish a constant lower bound. In particular, this guarantees a tight escape from barren plateaus as per Eq. (9).

Corollary G1. Consider an expectation of the form [Eq. (E1)] with a variational ansatz $U(\theta)$ consisting of Clifford gates and single-qubit Pauli-rotation gates involving $D \geqslant 1$ parameters. Assume that $H = P_i$ is a single Pauli string. For any arbitrarily small $\delta > 0$, initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components $\theta_1, \ldots, \theta_D$, where $\sigma = \mathcal{O}(D^{-\frac{1+\delta}{2}})$. Then, there exists a pair of Clifford operators $(Q, \tilde{Q})$ (cf. Theorem 1) such that

$$\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] \geqslant \mathbb{E}_\theta\big[\big(\partial_{\theta_k} C(\theta)\big)^2\big] = 1 + \mathcal{O}(D^{-\delta}),$$

after replacing $U(\theta)$ by $\tilde{Q}\,U(\theta)\,Q$.

Proof. Let $k \in [D]$ be arbitrary. By Theorem G1, we know that

$$\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] \geqslant \mathbb{E}_\theta\big[\big(\partial_{\theta_k} C(\theta)\big)^2\big] = \big(\partial_{\theta_k} C\big)(0)^2 + \mathcal{O}(D\sigma^2), \qquad \text{(G7)}$$

noting that $\|H\| = 1$ if $H = P$. On the other hand, Theorem 1 shows how to find the desired pair $Q, \tilde{Q}$ of Clifford operators such that

$$\big(\partial_{\theta_k} C\big)(0) = \pm 1, \qquad \text{(G8)}$$

after replacing $U(\theta)$ with $\tilde{Q}\,U(\theta)\,Q$. Therefore, observing that

$$\sigma = \mathcal{O}\big(D^{-\frac{1+\delta}{2}}\big) \Rightarrow \mathcal{O}(D\sigma^2) = \mathcal{O}(D^{-\delta}) \qquad \text{(G9)}$$

concludes the proof. ■

As in Theorem 1, this result may be generalized to arbitrary observables H as defined in Eq. (F1) with a lower bound depending on the distribution of the Pauli coefficients $\{c_i\}_{i\in\mathcal{I}}$.

Theorem 2. Let H be an arbitrary observable as in Eq. (5) with Pauli terms indexed by $\mathcal{I}$. Let $\delta > 0$ and initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components and $\sigma = \mathcal{O}(\|H\|^{-1} D^{-\frac{1+\delta}{2}})$. Then, for each $i_0 \in \mathcal{I}$, there exists a pair of Clifford operators $(Q, \tilde{Q})$ such that

$$\mathbb{E}_\theta[\|\nabla_\theta \tilde{C}(\theta)\|^2] \geqslant \mathbb{E}_\theta\big[\big(\partial_{\theta_k} \tilde{C}(\theta)\big)^2\big] = \beta_{i_0}(H)^2 + \mathcal{O}(D^{-\delta}),$$

where

$$\beta_{i_0}(H) := \pm c_{i_0} + \frac{1}{2}\sum_{i\in\mathcal{I}\setminus\{i_0\}} c_i\big\{\langle 0|\tilde{P}_i^{0,e_k}|0\rangle - \langle 0|\tilde{P}_i^{e_k,e_k}|0\rangle\big\},$$

defining $\tilde{P}_i^{j,\alpha} := Q^\dagger U(\xi_{j,\alpha})^\dagger \tilde{Q}^\dagger P_i \tilde{Q} U(\xi_{j,\alpha}) Q$ and $\xi_{j,\alpha} := \frac{\pi}{2}(\alpha - 2j)$ for $0 \leqslant j \leqslant \alpha$, and $\tilde{C}(\cdot)$ denoting the LCE-transformed cost. In particular, $\beta_{i_0}(H) = 1$ if $H = P$ is a single Pauli string. In particular, $\mathbb{E}_\theta[\|\nabla_\theta \tilde{C}(\theta)\|^2] = \Omega(1)$ with probability at least $1 - \mathcal{O}(|\mathcal{I}|2^{-N})$ if $c_{i_0} = \Omega(1)$ (conditioned on P_{i_0}).

Proof. Let $k \in [D]$ and $i_0 \in \mathcal{I}$ be arbitrary. By Theorem G1, we know that

$$\begin{aligned}\mathbb{E}_\theta[\|\nabla_\theta C(\theta)\|^2] &\geqslant \mathbb{E}_\theta\big[\big(\partial_{\theta_k} C(\theta)\big)^2\big] \\ &= \big(\partial_{\theta_k} C\big)(0)^2 + \|H\|^2\,\mathcal{O}(D\sigma^2). \qquad \text{(G10)}\end{aligned}$$

On the other hand, Theorem 1 yields the desired pair $(Q, \tilde{Q})$ of Clifford operators such that

$$\begin{aligned}\big(\partial_{\theta_k} \tilde{C}\big)(0) &= \pm c_{i_0} + \frac{1}{2}\sum_{i\in\mathcal{I}\setminus\{i_0\}} c_i\big\{\langle 0|\tilde{P}_i^{0,e_k}|0\rangle - \langle 0|\tilde{P}_i^{e_k,e_k}|0\rangle\big\} \\ &= \beta_{i_0}(H), \qquad \text{(G11)}\end{aligned}$$

after replacing $U(\theta)$ with $\tilde{Q}\,U(\theta)\,Q$ via the LCE transformation. Therefore, observing that

$$\sigma = \mathcal{O}\big(\|H\|^{-1} D^{-\frac{1+\delta}{2}}\big) \Rightarrow \|H\|^2\,\mathcal{O}(D\sigma^2) = \mathcal{O}(D^{-\delta}). \qquad \text{(G12)}$$

The second statement follows from the assumption that $c_{i_0} = \Omega(1)$, which dominates the decaying term $\mathcal{O}(D^{-\delta})$, and Lemma 1 shown independently below. This concludes the proof. ■

The observable-dependent lower bound $\beta_{i_0}(H)^2$ in Theorem 2 is only meaningful if it behaves like $\Omega(1)$, i.e., when it dominates error term $\mathcal{O}(D^{-\delta})$. The scaling behavior of $\beta_{i_0}(H)$ is less straightforward to analyze due to the cancellations between Pauli terms. The following lemma shows that the probability of total cancellation [i.e., $\beta_{i_0}(H) = 0$], conditioned on P_{i_0}, which is given for the LCE construction, is vanishing exponentially in the number of qubits N.

Lemma 1. Let $H = \sum_{i\in\mathcal{I}} c_i P_i$ be a random observable with $P_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathcal{P}_N \setminus \{I\})$. Then, for each $i_0 \in \mathcal{I}$, we have the conditional probability that

$$\mathbb{P}\big(\beta_{i_0}(H)^2 = c_{i_0}^2 \big| P_{i_0}\big) \geqslant 1 - \mathcal{O}(|\mathcal{I}|\,2^{-N}),$$

as $N \to \infty$.

Proof. Recall the definition of the N-qubit Pauli group,

$$\mathcal{P}_N := \{\pm P, \pm iP : P \in \{I, X, Y, Z\}^{\otimes N}\}, \qquad \text{(G13)}$$

which contains $4 \times 4^N = 4^{N+1}$ elements. In the construction of LCE, we were particularly interested in the subgroup

$$\langle I, Z\rangle_N = \{\pm P, \pm iP : P \in \{I, Z\}^{\otimes N}\} \qquad \text{(G14)}$$

of diagonal Pauli strings, containing $4 \times 2^N = 2^{N+2}$ elements.

First, we note that a uniformly random (nonidentity) Pauli string P satisfies that

$$\mathbb{P}(\langle 0|P|0\rangle = 1) = \frac{\#\langle I, Z\rangle_N \setminus \{I\}}{\#\mathcal{P}_N \setminus \{I\}} = \frac{2^{N+2}-1}{4^{N+1}-1} = \mathcal{O}(2^{-N}), \tag{G15}$$

since the only nonzero Pauli expectations are achieved by diagonal operators in $\langle I, Z\rangle_N$. In other words, the probability of having a nonzero expectation is determined by the total number of nonidentity diagonal Pauli matrices divided by the total number of nonidentity Pauli matrices. That is, for any uniformly random (nonidentity) Pauli string P we have

$$\mathbb{P}(\langle 0|P|0\rangle = 0) = 1 - \mathcal{O}(2^{-N}), \tag{G16}$$

which we will use to conclude with the lower bound.

Recall from the proof of Theorem F1 that for each (j, α), and $i \in \mathcal{I}$,

$$\tilde{P}_i^{j,\alpha} := Q^\dagger U_{j,\alpha}^\dagger \tilde{Q}^\dagger P_i \tilde{Q} U_{j,\alpha} Q, \tag{G17}$$

for some deterministic $U_{j,\alpha} \in \text{Clifford}(N)$ [i.e., $U(\cdot)$ evaluated at certain parameter points] depending on the LCE direction $k \in [D]$ (since $\alpha = e_k$ and $j \in \{0, e_k\}$), which is deterministically fixed by assumption. On the other hand, the LCE pair $(Q, \tilde{Q})$ depends on both the (deterministic) LCE direction $k \in [D]$, and Pauli term $i_0 \in \mathcal{I}$.

Crucially, the LCE pair $(Q, \tilde{Q})$ is explicitly constructed from the random Pauli string P_{i_0} in order to isolate the Pauli coefficient c_{i_0}, thus causing $(Q, \tilde{Q})$ to be random as well. Hence, given the choice of the Pauli string P_{i_0} for the LCE construction, the pair $(Q, \tilde{Q})$ is also given, which in turn results in $\tilde{P}_i^{j,\alpha}$ being pairwise independent in $i \in \mathcal{I}$ with respect to the conditional probability $\mathbb{P}(\cdot \mid P_{i_0})$.

Defining $d_i := \langle 0|\tilde{P}_i^{0,e_k}|0\rangle - \langle 0|\tilde{P}_i^{e_k,e_k}|0\rangle$, we thus obtain

$$\begin{aligned}\mathbb{P}\big(d_i = 0\,\forall i \in \mathcal{I}\setminus\{i_0\}\big|P_{i_0}\big) &\overset{\text{i.i.d.}}{=} \prod_{i\in\mathcal{I}\setminus\{i_0\}} \mathbb{P}\big(d_i = 0\big|P_{i_0}\big) \\ &\overset{\text{(G16)}}{\geqslant} (1-\mathcal{O}(2^{-N}))^{|\mathcal{I}|} \\ &= 1 - \mathcal{O}(|\mathcal{I}|\,2^{-N}).\end{aligned} \tag{G18}$$

Consequently, by the previous estimation,

$$\begin{aligned}\mathbb{P}\left(\sum_{i\in\mathcal{I}\setminus\{i_0\}} c_i d_i = 0 \,\middle|\, P_{i_0}\right) &\geqslant \mathbb{P}\big(d_i = 0\,\forall i \in \mathcal{I}\setminus\{i_0\}\big|P_{i_0}\big) \\ &\geqslant 1 - \mathcal{O}(|\mathcal{I}|\,2^{-N}),\end{aligned} \tag{G19}$$

which directly implies the desired claim since by definition,

$$\beta_{i_0}(H) = \pm c_{i_0} + \frac{1}{2}\sum_{i\in\mathcal{I}\setminus\{i_0\}} c_i d_i. \tag{G20}$$

■

APPENDIX H: STATE EXPRESSIVITY OF PARAMETRIZED QUANTUM CIRCUITS

It is essential to verify that the LCE transformation $C(\cdot) \mapsto \tilde{C}(\cdot)$ preserves desirable properties of the optimization landscape. Recall that $\tilde{C}(\cdot)$ denotes the LCE-transformed cost obtained by replacing $U(\cdot)$ with $\tilde{U}(\cdot) := \tilde{Q}U(\cdot)Q$ as in Theorem 1. We show that both expressivity and global optimality remain intact, using the following geometric notion of expressivity.

Definition H1 (ε-state-expressivity). Let $\varepsilon > 0$. A PQC $U(\cdot)$ is said to be ε-state expressive if for every state $|\psi\rangle \in \mathcal{H}_N$ and every stabilizer state $|\psi_0\rangle \in \text{Stab}(N)$, there exists a parameter configuration $\theta \in [-\pi, \pi]^D$ such that

$$\|U(\theta)|\psi_0\rangle - |\psi\rangle\|_2 < \varepsilon. \tag{H1}$$

Here, it suffices to consider the metric

$$\||\psi\rangle - |\phi\rangle\|_2 := \sqrt{2(1 - \text{Re}\langle\psi|\phi\rangle)}, \tag{H2}$$

induced by the ℓ_2 norm, to measure the distance between quantum states $|\psi\rangle, |\phi\rangle \in \mathcal{H}_N$. The reason for considering arbitrary stabilizer inputs $|\psi_0\rangle \in \text{Stab}(N)$ is that the expressivity of the model $U(\cdot)$ should not depend on a finite set of classically tractable input states.

This definition captures the geometric capacity of the ansatz $U(\cdot)$ to explore the Hilbert space $\mathcal{H}_N$. Intuitively, if $U(\cdot)$ explores $\mathcal{H}_N$ in an unbiased fashion, then the Kullback-Leibler (KL) divergence $D_{\text{KL}}(\hat{P}_{\text{PQC}}(F;\theta)\,\|\,P_{\text{Haar}}(F))$ in Ref. [29] is small if and only if $U(\cdot)$ is ε-state expressive for sufficiently small $\varepsilon > 0$.

Recall that we consider a VQE [2] problem of the form

$$\theta_* = \arg\min_{\theta\in[-\pi,\pi]^D} C(\theta), \tag{H3}$$

where $C(\theta_*)$ approximates the ground state energy $\lambda_{\min}(H)$ of the observable in Eq. (F1). We first show that sufficient ε-state expressivity guarantees the existence of a good solution to Eq. (H3).

Lemma H1. Assume that the quantum model $U(\cdot)$ is $\frac{\varepsilon}{2\|H\|}$-state expressive. Then, there exists a configuration $\theta_* \in [-\pi, \pi]^D$ such that

$$|C(\theta_*) - \lambda_{\min}(H)| < \varepsilon,$$

where $\lambda_{\min}(H)$ denotes the smallest eigenvalue of the observable in Eq. (F1).

Proof. Let $|\psi\rangle \in \mathcal{H}_N$ be the eigenstate corresponding to the smallest eigenvalue $\lambda_{\min}(H)$ of H. That is, $H|\psi\rangle = \lambda_{\min}(H)|\psi\rangle$. Then, since $U(\cdot)$ is $\frac{\varepsilon}{2\|H\|}$-state expressive, there exists a configuration $\theta_* \in [-\pi, \pi]^D$ such that

$$\|U(\theta_*)|0\rangle - |\psi\rangle\| < \frac{\varepsilon}{2\|H\|}. \tag{H4}$$

As a consequence, defining $|\phi\rangle := U(\theta_*)|0\rangle$,

$$\begin{aligned}&|C(\theta_*) - \lambda_{\min}(H)| \\ &= |\langle\phi|H|\phi\rangle - \langle\psi|H|\psi\rangle| \\ &= |(\langle\phi| - \langle\psi|)H|\phi\rangle + \langle\psi|H(|\phi\rangle - |\psi\rangle)| \\ &\leqslant |(\langle\phi| - \langle\psi|)H|\phi\rangle| + |\langle\psi|H(|\phi\rangle - |\psi\rangle)| \\ &\leqslant \|H\|\,\underbrace{\||\phi\rangle\|}_{=1}\,\||\phi\rangle - |\psi\rangle\| + \|H\|\,\underbrace{\||\psi\rangle\|}_{=1}\,\||\phi\rangle - |\psi\rangle\| \\ &= 2\|H\|\,\||\phi\rangle - |\psi\rangle\| \overset{\text{(H4)}}{<} \varepsilon,\end{aligned} \tag{H5}$$

where we employed the triangle inequality and Cauchy Schwarz, respectively. ■

Lemma H1 proves that, under its assumptions, $C(\theta_*)$ provides a good approximation of the ground energy of H, and thus solves the VQE problem in Eq. (H3). The next step is to establish that the LCE transformation $C(\cdot) \mapsto \tilde{C}(\cdot)$ does not compromise ε-state expressivity. That is, the quantum model is just as ε-state expressive after replacing $U(\cdot)$ with $\tilde{Q}\,U(\cdot)\,Q$ as per Theorem 1.

Lemma H2. If the quantum model $U(\cdot)$ is ε-state expressive, then so is $\tilde{Q}\,U(\cdot)\,Q$ for each $Q, \tilde{Q} \in \text{Clifford}(N)$. In particular, the LCE transformation $C(\cdot) \mapsto \tilde{C}(\cdot)$ leaves the ε-state expressivity of the quantum model $U(\cdot)$ invariant.

Proof. Let $|\psi_0\rangle \in \text{Stab}(N)$, and $|\psi\rangle \in \mathcal{H}_N$ be arbitrary. By assumption that $U(\cdot)$ is ε-state expressive, we find

$$\begin{aligned}\|\tilde{Q}U(\theta)Q|\psi_0\rangle - |\psi\rangle\| &= \|\tilde{Q}U(\theta)|\tilde{\psi}_0\rangle - |\psi\rangle\| \\ &= \|U(\theta)|\tilde{\psi}_0\rangle - |\tilde{\psi}\rangle\| < \varepsilon, \end{aligned} \tag{H6}$$

using that $|\tilde{\psi}_0\rangle := Q|\psi_0\rangle \in \text{Stab}(N)$ defines a stabilizer state, and that $\tilde{Q}$ is unitary, leaving the norm invariant in the second equality with $|\tilde{\psi}\rangle := \tilde{Q}^\dagger|\psi\rangle$. ■

It remains to show that the global quality of the optimization landscape is preserved under the LCE transformation $C(\cdot) \mapsto \tilde{C}(\cdot)$. Specifically, we prove that the global minimum value satisfies $C(\theta_*) \approx \tilde{C}(\theta_*)$, with an error controllable by assuming stronger ε-state expressivity. The proof relies on ensuring that a change of basis leaves the eigenspectrum invariant.

Lemma H3. Let $\tilde{Q}$ be an arbitrary unitary operator, and define $\tilde{H} := \tilde{Q}^\dagger H\,\tilde{Q}$, where H denotes the observable in Eq. (5). Then, H and $\tilde{H}$ share the same eigenspectrum.

Proof. Let $\lambda(H)$ be an arbitrary eigenvalue of H, and take the corresponding eigenstate $|\psi\rangle \in \mathcal{H}_N$. In other words, $H|\psi\rangle = \lambda(H)|\psi\rangle$. Define the state $|\tilde{\psi}\rangle := \tilde{Q}^\dagger|\psi\rangle$. Then,

$$\begin{aligned}\tilde{H}|\tilde{\psi}\rangle = \tilde{Q}^\dagger H \underbrace{\tilde{Q}\tilde{Q}^\dagger}_{=I}|\psi\rangle = \tilde{Q}^\dagger \underbrace{H|\psi\rangle}_{=\lambda(H)|\psi\rangle} \\ = \lambda(H)\tilde{Q}^\dagger|\psi\rangle = \lambda(H)|\tilde{\psi}\rangle, \end{aligned} \tag{H7}$$

implying that $\lambda(H) = \lambda(\tilde{H})$. The claim follows since $\lambda(H)$ was arbitrary. ■

We are now ready to prove the remaining statement.

Proposition H1. Assume that the quantum model $U(\cdot)$ is $\frac{\varepsilon}{4\|H\|}$-state expressive. Then, we have that

$$\left|\min_{\theta\in[-\pi,\pi]^D} C(\theta) - \min_{\theta\in[-\pi,\pi]^D} \tilde{C}(\theta)\right| < \varepsilon,$$

where $\tilde{C}(\cdot)$ denotes the cost after replacing $U(\cdot)$ by $\tilde{Q}\,U(\cdot)\,Q$. This ensures that the LCE does not compromise the quality of the global minimum.

Proof. By definition, the LCE-transformed cost equals

$$\tilde{C}(\theta) := \langle 0|Q^\dagger U(\theta)^\dagger \tilde{H} U(\theta) Q|0\rangle, \tag{H8}$$

where we introduced the transformed observable $\tilde{H} = \tilde{Q}^\dagger H \tilde{Q}$. The goal is to prove that $C(\cdot)$ and $\tilde{C}(\cdot)$ have approximately the same global minimum. The idea is to first make use of state expressivity in order to ensure that $C(\cdot)$ is able to approximate the smallest eigenvalue $\lambda_{\min}(H)$. This step is achieved with the help of Lemma H1. Next, we use Lemma H2 to establish an identical bound for the LCE-transformed cost $\tilde{C}(\cdot)$ approximating $\lambda_{\min}(\tilde{H})$. To conclude with the proof, we employ Lemma H3 stating that the change of basis $H \mapsto \tilde{Q}^\dagger H \tilde{Q} = \tilde{H}$ leaves the eigenspectrum invariant, i.e., $\lambda_{\min}(H) = \lambda_{\min}(\tilde{H})$, allowing us to combine both previous bounds for the desired result.

We now elaborate the above proof strategy more formally: Since $U(\cdot)$ is $\frac{\varepsilon}{4\|H\|}$-state expressive by assumption, it follows from Lemma H1 that we can approximate the lowest eigenvalue for the original cost:

$$\left|\min_{\theta\in[-\pi,\pi]^D} C(\theta) - \lambda_{\min}(H)\right| < \frac{\varepsilon}{2}. \tag{H9}$$

By Lemma H2, we know that the LCE transformation leaves ε-expressivity invariant, thus establishing the exact same bound for the LCE-transformed cost:

$$\left|\min_{\theta\in[-\pi,\pi]^D} \tilde{C}(\theta) - \lambda_{\min}(\tilde{H})\right| < \frac{\varepsilon}{2}. \tag{H10}$$

As a result, by the triangle inequality,

$$\begin{aligned}&\left|\min_{\theta\in[-\pi,\pi]^D} C(\theta) - \min_{\theta\in[-\pi,\pi]^D} \tilde{C}(\theta)\right| \\ &\leqslant \left|\min_{\theta\in[-\pi,\pi]^D} C(\theta) - \lambda_{\min}(H)\right| + \left|\min_{\theta\in[-\pi,\pi]^D} \tilde{C}(\theta) - \lambda_{\min}(H)\right| \\ &= \left|\min_{\theta\in[-\pi,\pi]^D} C(\theta) - \lambda_{\min}(H)\right| + \left|\min_{\theta\in[-\pi,\pi]^D} \tilde{C}(\theta) - \lambda_{\min}(\tilde{H})\right| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned} \tag{H11}$$

where we made use of Lemma H3, which states that $\lambda_{\min}(H) = \lambda_{\min}(\tilde{H})$. ■

Proposition H1 confirms that the LCE transformation preserves the quality of the global minimum. Notably, we make no structural assumptions on the PQC ansatz $U(\cdot)$ (e.g., no light cone or locality constraints), allowing us to assume ε-state expressivity for arbitrarily small $\varepsilon > 0$.

APPENDIX I: MINIMAL TRUNCATION THRESHOLD OF THE TAYLOR SURROGATE

We approximate the PQC using the Taylor surrogate $C_m(\theta)$ from Eq. (10), where the truncation threshold $m = m(\theta)$ determines computational complexity. To meet a prescribed surrogate error tolerance, we analyze the minimal necessary m via two approaches:

(1) *Worst-case error*: $m = m(\theta)$ is a deterministic function of a fixed configuration $\theta \in [-\pi, \pi]^D$.

(2) *Mean-squared error*: $m = m(\mathbb{P}_\theta)$ is defined probabilistically based on an initialization distribution $\mathbb{P}_\theta$.

1. Worst-case error analysis

The following lemma quantifies the (deterministic) worst-case error of the approximation $C_m(\theta) \approx C(\theta)$ as a function of θ.

Lemma I1. Consider an expectation of the form Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series

$C_m(\theta)$ defined in Eq. (D9). Let $\theta \in [-\pi, \pi]^D$ and $m \in \mathbb{N}$ be arbitrary. Then, for every $\theta \in [-\pi, \pi]^D$ and $m \in \mathbb{N}$, we have

$$|C(\theta) - C_m(\theta)| \leqslant \frac{\|H\|}{m!} \|\theta\|_1^m, \tag{I1}$$

where $\|H\|$ is the operator norm of the observable in Eq. (F1).

Proof. Let $\theta \in [-\pi, \pi]^D$ and $m \in \mathbb{N}$ be arbitrary. Using Lagrange's form of the remainder in the Taylor series, we find an element $\xi \in [0, \theta]$ in the line segment from 0 to θ, such that

$$\begin{aligned} C(\theta) - C_m(\theta) &= \sum_{\|\alpha\|_1 \geqslant m} \left(D_\theta^\alpha C\right)(0) \frac{\theta^\alpha}{\alpha!} = \sum_{\|\alpha\|_1 = m} \underbrace{\left(D_\theta^\alpha C\right)(\xi)}_{\leqslant \|H\|} \frac{\theta^\alpha}{\alpha!} \\ &\leqslant \|H\| \sum_{\|\alpha\|_1 = m} \left|\frac{\theta^\alpha}{\alpha!}\right| \leqslant \|H\| \frac{1}{m!} \left(\sum_{k=1}^D |\theta_k|\right)^m \\ &= \frac{\|H\|}{m!} \|\theta\|_1^m, \end{aligned} \tag{I2}$$

where in the first estimate we made use of the uniform bound in the higher-order parameter-shift rule (Lemma E2), and the multinomial theorem in the second, i.e.,

$$\sum_{\|\alpha\|_1 = m} \binom{m}{\alpha} \theta^\alpha = \left(\sum_{k=1}^D \theta_k\right)^m, \quad \binom{m}{\alpha} := \frac{m!}{\alpha_1! \cdots \alpha_D!}. \tag{I3}$$

■

In order to solve for the truncation threshold m, we make use of asymptotic properties of the Lambert-W function [69].

Lemma I2. Let $W(\cdot)$ denote the main branch of the Lambert-W function [69]. Then,

$$W(x) = x + \mathcal{O}(x^2)$$

as $x \to 0$.

Proof. By definition, $W(x)\, e^{W(x)} = x$. In particular, $W(0) = 0$. We shall use the implicit function theorem to prove the lemma statement. For that purpose, let us define the analytic function $F(w, x) := we^w - x$. By the previous observation, we have $F(0, 0) = 0$. Next, we note that $\partial_w F(w, x) = (w + 1)e^w$, which implies $(\partial_w F)(0, 0) = 1 \neq 0$. By the implicit function theorem, there exists a unique analytic function $W(\cdot)$ satisfying $W(0) = 0$ and $F(W(x), x) = 0$ in a neighborhood of $x = 0$. By uniqueness, $W(\cdot)$ has to coincide with the principal branch of the Lambert-W function [69]. Since $W(\cdot)$ is analytic (in a neighborhood of $x = 0$), we may consider the Taylor expansion,

$$W(x) = W(0) + W'(0)\, x + \mathcal{O}(x^2). \tag{I4}$$

Taking the derivative on both sides of $W(x)e^{W(x)} = x$ yields

$$\begin{aligned} W'(x)e^{W(x)}(1 + W(x)) = 1 \Rightarrow W'(x) = \frac{e^{-W(x)}}{1 + W(x)} \Rightarrow W'(0) \\ = 1, \end{aligned} \tag{I5}$$

making use of $W(0) = 0$. Thus, Eq. (I4) reduces to the desired asymptotics $W(x) = x + \mathcal{O}(x^2)$ as $x \to 0$. ■

We are now ready to deterministically lower bound the truncation threshold $m(\theta)$.

Lemma I3. Consider an expectation in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, simulated classically via the Taylor series $C_m(\theta)$ from Eq. (D9). Assume that the Taylor remainder in Eq. (I1) is bounded by a tolerance $\varepsilon > 0$, and let $m = m(\theta)$ be the truncation order in Eq. (D9). Then, the truncation threshold $m = m(\theta)$ satisfies the lower bound

$$m(\theta) \geqslant e\|\theta\|_1 \exp\left(W\left(\frac{1}{e\|\theta\|_1} \log\left(\frac{\|H\|}{\varepsilon\sqrt{e^3\|\theta\|_1}}\right)\right)\right) - \frac{1}{2},$$

where $e \approx 2.718\,28$ is the natural constant, $W(\cdot)$ denotes the main branch of the Lambert-W function [69], and $\|H\|$ is the operator norm of the observable in Eq. (F1). In particular, if $\|\theta\|_1 \to \infty$ increases as $D \to \infty$, then the asymptotic equivalence $m(\theta) \sim e\|\theta\|_1$ is sufficient for the truncation threshold to ensure the error tolerance ε (where the contributions of $\|H\|$ and ε become negligible).

Proof. Let $\varepsilon > 0$ be a given error tolerance. Without loss of generality $\theta \neq 0$, so that $m(\theta) \geqslant 2$. Otherwise if $\theta = 0$, the Taylor approximation is exact with $m(0) = 1$. Utilizing a Stirling approximation yields the lower bound

$$m! \leqslant e\sqrt{m}\left(\frac{m}{e}\right)^m, \tag{I6}$$

for all $m \geqslant 2$. This bound is derived from the remark of Robbins [70] that for all $m \geqslant 1$,

$$m! \leqslant \sqrt{2\pi m}\left(\frac{m}{e}\right)^m e^{\frac{1}{12m}}, \tag{I7}$$

and verifying that for all $m \geqslant 2$,

$$\sqrt{2\pi}\, e^{\frac{1}{12m}} \leqslant e. \tag{I8}$$

As a result, we require by Lemma I1 that the Taylor approximation error is bounded by

$$\begin{aligned} \log\varepsilon &\overset{!}{\geqslant} \log\left(\frac{\|H\|}{m!}\|\theta\|_1^m\right) \overset{(I6)}{\geqslant} \log\left(\frac{\|H\|}{e\sqrt{m}} \frac{(e\|\theta\|_1)^m}{m^m}\right) \\ &= \log\|H\| - \frac{1}{2}\log(e^2 m) - m\log\left(\frac{m}{e\|\theta\|_1}\right). \end{aligned} \tag{I9}$$

Noting that $\log\varepsilon = -\log(1/\varepsilon)$, we may manipulate the following expression:

$$\begin{aligned} \gamma &:= \frac{1}{e\|\theta\|_1} \log\left(\frac{\|H\|}{\varepsilon\sqrt{e^3\|\theta\|_1}}\right) \\ &= -\frac{1}{e\|\theta\|_1}\log\varepsilon + \frac{1}{e\|\theta\|_1}\log\|H\| - \frac{1}{2e\|\theta\|_1}\log(e^3\|\theta\|_1) \\ &\overset{(I9)}{\leqslant} \frac{1}{2e\|\theta\|_1}\log(e^2 m) + \frac{m}{e\|\theta\|_1}\log\left(\frac{m}{e\|\theta\|_1}\right) \\ &\quad - \frac{1}{2e\|\theta\|_1}\log(e^3\|\theta\|_1) \\ &= \frac{1}{2e\|\theta\|_1} \underbrace{\{\log(e^2 m) - \log(e^3\|\theta\|_1)\}}_{=\log\left(\frac{m}{e\|\theta\|_1}\right)} \\ &\quad + \frac{m}{e\|\theta\|_1}\log\left(\frac{m}{e\|\theta\|_1}\right) \end{aligned} \tag{I10}$$

$$= \frac{2m+1}{2e\|\theta\|_1} \log\left(\frac{m}{e\|\theta\|_1}\right) \leqslant \frac{2m+1}{2e\|\theta\|_1} \log\left(\frac{2m+1}{2e\|\theta\|_1}\right). \tag{I11}$$

We have therefore established the bound

$$\gamma \overset{(I11)}{\leqslant} y \log y =: f(y), \quad \text{substituting } y := \frac{2m+1}{2e\|\theta\|_1}. \tag{I12}$$

The equation $f(y) = \gamma$ is known to be solved by the main branch of the Lambert-W function [69] with solution $y = e^{W(\gamma)}$. Moreover, the function f is invertible on $y > 1$. In that case,

$$\frac{2m+1}{2e\|\theta\|_1} = y \overset{(I12)}{\geqslant} f^{-1}(\gamma) = e^{W(\gamma)} = \frac{1}{e\|\theta\|_1} \log\left(\frac{\|H\|}{\varepsilon\sqrt{e^3\|\theta\|_1}}\right), \tag{I13}$$

or equivalently,

$$m \geqslant e\|\theta\|_1 \exp\left(W\left(\frac{1}{e\|\theta\|_1} \log\left(\frac{\|H\|}{\varepsilon\sqrt{e^3\|\theta\|_1}}\right)\right)\right) - \frac{1}{2} =: b(\theta). \tag{I14}$$

In order to prove that $y > 1$ always holds true in our case, we prove the following: If $\|\theta\|_1 \to \infty$ as $D \to \infty$, then there exists a lower bound $b(\theta) \sim e\|\theta\|_1$ as $D \to \infty$, where $b(\theta)$ is the lower bound defined in Eq. (I14). Let us substitute

$$x := \frac{1}{e\|\theta\|_1} \log\left(\frac{\|H\|}{\varepsilon\sqrt{e^3\|\theta\|_1}}\right). \tag{I15}$$

Observe that $x \to 0$ as $D \to \infty$ by assumption on the asymptotics of $\|\theta\|_1$ [even if $\|H\|$ and/or $\varepsilon > 0$ depend on D, provided that their ratio is not exponential in D, which is ensured by $|\mathcal{I}| = \mathcal{O}(\mathrm{Poly}(D))$]. Moreover, by definition of the Lambert-W function [69], we have $W(x)e^{W(x)} = x$. Therefore, applying Lemma I2 [which tells us that $x/W(x) \sim 1$ as $x \to 0$],

$$m(x) \overset{(I14)}{\geqslant} b(\theta) = e\|\theta\|_1 \exp(W(x)) - \frac{1}{2} = e\|\theta\|_1 \frac{x}{W(x)} - \frac{1}{2} \sim e\|\theta\|_1 \tag{I16}$$

as $D \to \infty$ (since $x \to 0$). In particular, recalling Eq. (I12), $y > 1$ equivalently means that

$$m > e\|\theta\|_1 - 1, \tag{I17}$$

which always holds true by Eq. (I16). Moreover, Eq. (I16) also shows that $m(\theta) \sim e\|\theta\|_1$ as $D \to \infty$ is sufficient for the truncation threshold to ensure the error tolerance $\varepsilon > 0$. ■

The asymptotics of the lower bound in Lemma I3 may be used to establish useful statistical insights about the expected truncation threshold, given a certain initial distribution $\mathbb{P}_\theta$.

Proposition I1. Consider an expectation of the form in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series $C_m(\theta)$ defined in Eq. (D9) within some error tolerance ε. Let $m = m(\theta)$ be the truncation threshold in the Taylor approximation in Eq. (D9). Initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components $\theta_1, \ldots, \theta_D$, where $\sigma = D^{-r}$ for some $r \in [0, 1)$. Then, as $D \to \infty$, the expected truncation threshold behaves according to

$$\mathbb{E}_\theta[m(\theta)] = \Theta(D^{1-r}).$$

Proof. Let $r \in [0, 1)$ be arbitrary. By Lemma I3, we know that $m(\theta) \sim e\|\theta\|_1$ as $D \to \infty$ is sufficient to simulate $C_m(\theta)$ within some error tolerance $\varepsilon > 0$. By the law of large numbers, we may simply take the expectation value of the ℓ_1 norm, which gives

$$\mathbb{E}_\theta[\|\theta\|_1] = \sum_{k=1}^{D} \mathbb{E}_\theta[|\theta_k|] = D \underbrace{\mathbb{E}_\theta[|\theta_1|]}_{=\Theta(\sigma)} = D\,\Theta(D^{-r}) = \Theta(D^{1-r}), \tag{I18}$$

where the second equality uses that the components of θ are i.i.d. ■

2. Mean-squared error analysis

Next, we quantify the mean-squared error of the approximation $C_m(\theta) \approx C(\theta)$ with respect to the local distribution $\mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ for some $\sigma > 0$. However, we first require the following counting argument to conclude with the result.

Lemma I4. Let $m = \mathcal{O}(1)$ be of constant order, i.e., independent of D. Then,

$$\sum_{\|\alpha\|_1 < m} 1 = \mathcal{O}(D^{m-1}), \tag{I19}$$

where $\{\|\alpha\|_1 < m\} := \{\alpha \in \mathbb{N}_0^D : |\alpha_1| + \cdots + |\alpha_D| < m\}$. On the other hand, if $m = o(D)$ strictly increases sublinearly as $D \to \infty$, then

$$\sum_{\|\alpha\|_1 < m} 1 = \Omega\left(\mathrm{Poly}(D)^{-1}\left(\frac{eD}{m}\right)^m\right) \tag{I20}$$

is superpolynomially lower bounded. In particular, if $m = \Theta(D^\delta)$ for some $\delta \in (0, 1)$, then $\#\{\|\alpha\|_1 < m\}$ is at least superpolynomial in D.

Proof. We begin by proving Eq. (I19). By the "balls and cells" formula [Eq. (5.2) in Ref. [71]], we count

$$\sum_{\|\alpha\|_1 < m} 1 = \sum_{i=0}^{m-1} \left(\sum_{\|\alpha\|_1 = i} 1\right) = \sum_{i=0}^{m-1} \binom{D+i-1}{i} = \binom{D+m-1}{m-1}, \tag{I21}$$

where the last identity follows from the hockey stick identity [72] from combinatorics, i.e.,

$$\binom{n+1}{k+1} = \sum_{j=k}^{n} \binom{j}{k} \quad \text{for all } n, k \in \mathbb{N} \text{ with } n \geqslant k. \tag{I22}$$

Indeed, by symmetry of the binomial coefficient, which is $\binom{n}{k} = \binom{n}{n-k}$ for all $n, k \in \mathbb{N}_0$, we find

$$\sum_{i=0}^{m-1} \binom{D+i-1}{i} = \sum_{i=0}^{m-1} \binom{D+i-1}{D-1} = \sum_{j=D-1}^{D+m-2} \binom{j}{D-1} \overset{(I22)}{=} \binom{D+m-1}{D} = \binom{D+m-1}{m-1}, \tag{I23}$$

where we applied symmetry in the first and last identity, respectively.

We are now ready to prove the polynomial upper bound in the case where $m = \mathcal{O}(1)$. For that purpose, note that if $D \geqslant m-1$ [which holds true as $D \to \infty$ the assumption that $m = \mathcal{O}(1)$], then

$$\begin{aligned}\binom{D+m-1}{m-1} &= \frac{(D+m-1)(D+m-2)\cdots(D+1)D}{(m-1)!} \\ &\leqslant \frac{(D+m-1)^{m-1}}{(m-1)!} \leqslant \frac{(2D)^{m-1}}{(m-1)!} \\ &= \mathcal{O}(D^{m-1}),\end{aligned} \tag{I24}$$

where we used $m = \mathcal{O}(1)$ in the last step.

It remains to prove the superpolynomial lower bound if $m \to \infty$ as $D \to \infty$. With the help of Stirling's inequality, $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$, we may lower bound

$$\begin{aligned}\binom{D+m-1}{m-1} &= \frac{(D+m-1)(D+m-2)\cdots(D+1)\times D}{(m-1)!} \\ &\geqslant \frac{D^{m-1}}{(m-1)!} \sim \frac{(eD)^{m-1}}{e\sqrt{m-1}(m-1)^{m-1}} \\ &\geqslant \underbrace{\frac{1}{e^2 D\sqrt{m-1}}}_{=\Theta(\mathrm{Poly}(D)^{-1})} \left(\frac{eD}{m}\right)^m,\end{aligned} \tag{I25}$$

using that $m = o(D)$ strictly increases sub-linearly as $D \to \infty$, which directly proves that Eq. (I25) is superpolynomial and subexponential. ∎

We are now equipped with all necessary tools to bound the mean-squared error.

Lemma I5. Consider an expectation of the form in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series $C_m(\theta)$ defined in Eq. (D9). Initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma, \sigma]^D)\}$ with i.i.d. components $\theta_1, \ldots, \theta_D$. Then, if $m = \mathcal{O}(1)$, the mean-squared error satisfies the bound

$$\mathbb{E}_\theta[(C(\theta) - C_m(\theta))^2] = \mathcal{O}\left(\|H\|^2 \frac{(D\sigma^2)^m}{m^m}\right), \tag{I26}$$

where $\|H\|$ is the operator norm of the observable in Eq. (F1). In particular, if $\sigma \leqslant q\sqrt{m}\,\|H\|^{-\frac{1}{m}} D^{-\frac{1}{2}}$ for some $0 < q < 1$, then the mean-squared error decays like $\mathcal{O}(q^{2m})$.

Proof. We use the Lagrange remainder to quantify the error,

$$\begin{aligned}C(\theta) - C_m(\theta) &= \sum_{\|\alpha\|_1 \geqslant m} \left(D_\theta^\alpha C\right)(0)\, \frac{\theta^\alpha}{\alpha!} \\ &= \sum_{\|\alpha\|_1 = m} \underbrace{\left(D_\theta^\alpha C\right)(\xi)}_{\leqslant \mathcal{O}(\|H\|)} \frac{\theta^\alpha}{\alpha!} \\ &= \mathcal{O}(\|H\|) \sum_{\|\alpha\|_1 = m} \frac{\theta^\alpha}{\alpha!},\end{aligned} \tag{I27}$$

employing the uniform bound $(D_\theta^\alpha C)(\xi) \leqslant \mathcal{O}(\|H\|)$ in the higher-order parameter-shift rule (Lemma E2) in the last estimation. As a result,

$$\begin{aligned}\mathbb{E}_\theta[(C(\theta) - C_m(\theta))^2] &= \mathcal{O}(\|H\|^2)\, \mathbb{E}_\theta\left[\left(\sum_{\|\alpha\|_1 = m} \frac{\theta^\alpha}{\alpha!}\right)^2\right] \\ &= \mathcal{O}(\|H\|^2) \frac{1}{m!^2} \mathbb{E}_\theta\left[\left(\sum_{k=1}^D \theta_k\right)^{2m}\right],\end{aligned} \tag{I28}$$

where the last identity follows from the multinomial theorem. In order to compute the last expectation, we first focus on the case $\mathbb{P}_\theta = \mathcal{N}(0, \sigma^2)$. In that case, we recall from basic probability theory that the sum $\sum_{k=1}^D \theta_k \sim \mathcal{N}(0, D\sigma^2)$ of independent Gaussian random variables remains a Gaussian with well-known moments explicitly given by

$$\mathbb{E}_\theta\left[\left(\sum_{k=1}^D \theta_k\right)^{2m}\right] = D^m \sigma^{2m} (2m-1)!! = D^m \sigma^{2m} \frac{(2m)!}{2^m m!}, \tag{I29}$$

noting that the double factorial $(2m-1)!! := (2m-1)(2m-3)\cdots 3 \times 1$ satisfies the identity $(2m-1)!! = \frac{(2m)!}{2^m m!}$. Indeed,

$$\begin{aligned}(2m)! = &\underbrace{(2m)(2m-2)\cdots 4 \times 2}_{=(2m)!!} \\ &\times \underbrace{(2m-1)(2m-2)\cdots 3 \times 1}_{=(2m-1)!!}\end{aligned} \tag{I30}$$

and

$$\begin{aligned}(2m)!! &= (2m)(2m-2)\cdots 4 \times 2 = 2^m m \times (m-1)\cdots 2 \times 1 \\ &= 2^m m!,\end{aligned} \tag{I31}$$

so that $(2m)! = 2^m m! \cdot (2m-1)!!$ as claimed. This shows that

$$\begin{aligned}\mathbb{E}_\theta[(C(\theta) - C_m(\theta))^2] &\overset{(I28)}{=} \mathcal{O}(\|H\|^2) \frac{1}{m!^2} \mathbb{E}_\theta\left[\left(\sum_{k=1}^D \theta_k\right)^{2m}\right] \\ &\overset{(I29)}{=} \mathcal{O}(\|H\|^2) \frac{(2m)!}{2^m m!^3} (D\sigma^2)^m.\end{aligned} \tag{I32}$$

Next, by utilizing Stirling's formula, $m! \sim \sqrt{2\pi m}(\frac{m}{e})^m$, we asymptotically establish

$$\frac{(2m)!}{2^m m!^3} \sim \frac{\sqrt{2\pi(2m)}\left(\frac{2m}{e}\right)^{2m}}{2^m(\sqrt{2\pi m})^3\left(\frac{m}{e}\right)^{3m}} = \frac{\sqrt{4\pi m}\,2^{2m}\cancel{\left(\frac{m}{e}\right)^{2m}}}{2\pi m\sqrt{2\pi m}\,\cancel{2^m}\left(\frac{m}{e}\right)^m\cancel{\left(\frac{m}{e}\right)^{2m}}} = \mathcal{O}\left(\frac{(2e)^m}{m^{m+1}}\right), \tag{I33}$$

which results in the desired bound,

$$\begin{aligned}\mathbb{E}_\theta[(C(\theta)-C_m(\theta))^2] &= \mathcal{O}(\|H\|^2)\frac{(2m)!}{2^m m!^3}(D\sigma^2)^m \\ &= \mathcal{O}\left(\|H\|^2\frac{(2eD\sigma^2)^m}{m^{m+1}}\right) \\ &= \mathcal{O}\left(\|H\|^2\frac{(D\sigma^2)^m}{m^m}\right),\end{aligned} \tag{I34}$$

using $m = \mathcal{O}(1)$ in the last identity.

In the case where $\mathbb{P}_\theta = \mathrm{Unif}([-\sigma,\sigma]^D)$ is i.i.d. uniform in each component, then $\mathbb{E}_\theta[\theta_k] = 0$ and $\mathrm{Var}_\theta[\theta_k] = \frac{\sigma^2}{3}$ for all $k \in [D]$. By the central limit theorem, the sum

$$\sum_{k=1}^{D}\theta_k \xrightarrow{d} \mathcal{N}\left(0, \frac{D\sigma^2}{3}\right) \tag{I35}$$

converges in distribution to a Gaussian as $D \to \infty$. This allows us to reduce this case asymptotically to the previous from Eq. (I28).

The second statement is immediately obtained by solving for σ in $\|H\|^2\frac{D^m\sigma^{2m}}{m^m} \overset{!}{\leqslant} c_0 q^{2m}$, where $c_0 > 0$ is some constant. Indeed, rearranging the inequality gives

$$\begin{aligned}\sigma^{2m} \leqslant \frac{c_0 q^{2m} m^m}{\|H\|^2 D^m} \iff \sigma \leqslant \underbrace{c_0^{-1/2m}}_{=\Theta_m(1)}\frac{q\sqrt{m}}{\|H\|^{1/m}\sqrt{D}} \\ = \mathcal{O}\big(q\sqrt{m}\,\|H\|^{-1/m}D^{-\frac{1}{2}}\big).\end{aligned} \tag{I36}$$

■

The next step is to establish a lower bound for the necessary truncation threshold $m = m(\mathbb{P}_\theta)$ in order to Taylor surrogate the cost in Eq. (E1); i.e., evaluate $C_m(\theta)$ with high probability and error of at most $\varepsilon > 0$.

Lemma I6. Consider an expectation of the form in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series $C_m(\theta)$ defined in Eq. (D9). Assume that $\|H\| = \mathcal{O}(1)$ and that the mean squared error in Eq. (I26) is bounded by a given tolerance $\varepsilon > 0$. Initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0,\sigma^2 I_D), \mathrm{Unif}([-\sigma,\sigma]^D)\}$ with i.i.d. components $\theta_1, \ldots, \theta_D$, where $\sigma = \mathcal{O}(q\,D^{-\frac{1}{2}})$ for some $0 < q < 1$. Then, there exists a constant $c > 0$ such that, with probability at least $1-\rho$, a constant truncation threshold $m = m(\mathbb{P}_\theta)$ satisfying

$$m(\mathbb{P}_\theta) \geqslant \frac{1}{2}\frac{\log\left(\frac{c}{\rho\,\varepsilon^2}\right)}{\log\left(\frac{1}{q}\right)}$$

is sufficient to simulate $C_m(\theta)$ with error at most $\varepsilon > 0$.

Proof. Let $\delta, \varepsilon, \rho > 0$ be arbitrary. Assume that $\|H\| = \mathcal{O}(1)$. If $\sigma = \mathcal{O}(q\,D^{-\frac{1}{2}})$, Lemma I5 yields

$$\mathbb{E}_\theta[(C(\theta)-C_m(\theta))^2] \leqslant c\,q^{2m}, \tag{I37}$$

for some constant $c > 0$. In that case, Markov's inequality bounds the probability of failure,

$$\begin{aligned}\mathbb{P}_\theta(|C(\theta)-C_m(\theta)| \geqslant \varepsilon) &\leqslant \frac{\mathbb{E}_\theta[(C(\theta)-C_m(\theta))^2]}{\varepsilon^2} \\ &\overset{(I37)}{\leqslant} \frac{c\,q^{2m}}{\varepsilon^2} \overset{!}{\leqslant} \rho.\end{aligned} \tag{I38}$$

Solving for the truncation threshold m in $\varepsilon^{-2}c\,q^{2m} \leqslant \rho$ implies

$$2m\log q \leqslant \log\left(\frac{\rho\,\varepsilon^2}{c}\right), \tag{I39}$$

or equivalently, since $\log q = -\log(\frac{1}{q})$ because of $0 < q < 1$,

$$m \geqslant \frac{1}{2}\frac{\log\left(\frac{c}{\rho\,\varepsilon^2}\right)}{\log\left(\frac{1}{q}\right)}. \tag{I40}$$

■

APPENDIX J: COMPUTATIONAL COMPLEXITY

The first part of this Appendix is dedicated to results in complexity theory showing that the Hamiltonian ground energy decision problem for observables of the form in Eq. (16) is classically infeasible in general, unless BPP = QMA.

1. The complexity class of LCE Hamiltonians

In order to successfully mitigate barren plateaus of arbitrarily expressive PQC unitaries $U(\theta)$ on parameter landscape patches with LCE, we are restricted to the class of observables

$$\begin{aligned}H = \sum_{i\in\mathcal{I}} c_i P_i, \quad \text{with} \quad \|H\| = \mathcal{O}(1) \\ \text{and} \quad c_{i_0} = \Omega(1) \quad \text{for some } i_0 \in \mathcal{I},\end{aligned} \tag{J1}$$

where $\mathcal{I} \subset \{I,X,Y,Z\}^{\otimes N} \setminus \{I_{2^N}\}$ is an arbitrary Pauli indexing set with $|\mathcal{I}| = \mathcal{O}(\mathrm{Poly}(N))$ and $c_i = \mathcal{O}(1)$ for all $i \in \mathcal{I}$. Here, we rigorously prove that, in general, finding the ground energy of such observables is classically hard in the sense that there exists no classical algorithm able to solve the problem under standard complexity theory assumptions. Since the Pauli strings define an $\mathbb{R}$-linear basis of the 4^N-dimensional Hilbert-Schmidt vector space $\mathrm{Herm}(2^N)$ of Hermitian matrices, we are particularly interested in the following class:

$$\begin{aligned}\mathscr{H}_{\mathrm{LCE}} := \Bigg\{ H + \sum_{j\in\mathcal{J}} c_j P_j \ \text{traceless} \Bigg| \mathrm{H} \in \mathrm{Herm}(2^{\mathrm{N}}) : \|\mathrm{H}\| \\ = \mathcal{O}(1),\ \mathcal{J} \neq \emptyset : |\mathcal{J}| = \mathcal{O}(1),\ \forall j \in \mathcal{J} : c_j = \Theta(1) \Bigg\}\end{aligned} \tag{J2}$$

of (arbitrarily local or global) traceless, normalized N-qubit Hamiltonians perturbed by an $\mathcal{O}(1)$-sparse $\Theta(1)$-linear com-

bination of Pauli strings, or for short, the class $\mathscr{H}_{\mathrm{LCE}}$ of LCE Hamiltonians. First, we note that every Hamiltonian $H \in \mathscr{H}_{\mathrm{LCE}}$ automatically satisfies Eq. (J1). Indeed, by the triangle inequality,

$$\left\| H + \sum_{j\in\mathcal{J}} c_j P_j \right\| \leqslant \underbrace{\|H\|}_{=\mathcal{O}(1)} + \sum_{j\in\mathcal{J}} \underbrace{|c_j|}_{\Theta(1)} \cdot \underbrace{\|P_j\|}_{=1} = \mathcal{O}(1) + \Theta(1) \cdot \underbrace{\sum_{j\in\mathcal{J}} 1}_{=|\mathcal{J}|=\mathcal{O}(1)} = \mathcal{O}(1), \quad \text{(J3)}$$

noting that $\|P_i\| = 1$ for every Pauli string P_i. This shows that LCE Hamiltonians are normalized. By definition, the LCE perturbation itself has constant-scaling coefficients $c_j = \Omega(1)$ for all $j \in \mathcal{J}$ thus addressing that condition in Eq. (J1). Finally, it remains to see (noting $H \neq 0$) that $\mathrm{Tr}\, H = 0$ in Eq. (J1) if and only if $I_{2^N} \notin \mathcal{I}$ since all Pauli strings P_i are traceless with $\mathrm{Tr}\, I_{2^N} = 2^N$ being the only exception. This shows that every LCE Hamiltonian in Eq. (J2) indeed satisfies the LCE condition in Eq. (J1).

In the following, we present two proofs showing that there exists no "one-size-fits-all" (deterministic or probabilistic) classical algorithm that can find the ground energy of an observable in the class $\mathscr{H}_{\mathrm{LCE}}$. We do this by reducing the problem to well-known results from complexity theory: The complexity class QMA is the quantum analog of NP, which contains both bounded-error probabilistic polynomial time (BPP) and P [73]. It is well known that the k local Hamiltonian problem defines a QMA-complete problem, even for $k = 2$ [54]. In that case, "QMA completeness" means that any QMA problem can be reduced to the k-local Hamiltonian problem. Note that BPP $\subseteq$ QMA [73]. In fact, it is a standard assumption in complexity theory that BPP $\neq$ QMA. Otherwise, any QMA problem (in particular, every k-local Hamiltonian problem) is guaranteed to be solvable by some probabilistic classical algorithm [54]. As a result, if BPP $\neq$ QMA, then there exists no classical algorithm that generally finds the ground energy of any given 2-local Hamiltonian. Increasing the locality k only makes the Hamiltonian problem more difficult.

Here, we mainly focus on the complexity of the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem. As a disclaimer, we do not claim that there always exists an LCE initialization strategy leading to training dynamics that converges to the unperturbed solution $\lambda_{\min}(H)$. We further emphasize that this work focuses solely on the theory of initialization and we thus leave convergence results as open problems for future efforts.

Let us briefly motivate the strategy behind our probabilistic proof first: Given an arbitrary (local or global) Hamiltonian $H \in \mathrm{Herm}(2^N)$ with $\|H\| = \mathcal{O}(1)$, we assume by contradiction that the ground energy of any LCE perturbation $H + V_{\mathcal{J}} \in \mathscr{H}_{\mathrm{LCE}}$ of H can be solved in BPP. We are then going to use the output $\lambda \approx \lambda_{\min}(\tilde{H})$ of such a classical BPP algorithm in combination with a certain, sufficiently small and randomly sampled LCE perturbation $V_{\mathcal{J}}$ to find the original solution $\lambda_{\min}(H)$. In particular, if H is chosen to be an arbitrary 2-local Hamiltonian, this would thus imply that QMA = BPP by QMA completeness of the 2-local Hamiltonian problem, which is absurd under standard complexity assumptions. Therefore, if BPP $\neq$ QMA, no such algorithm exists, hence implying nonclassicality of the class $\mathscr{H}_{\mathrm{LCE}}$ of LCE Hamiltonians. That is, in practice we are completely free to choose *any* LCE perturbation $V_{\mathcal{J}}$ strategy with the goal of finding the ground energy $\lambda_{\min}(H)$ of an arbitrary normalized Hamiltonian H. Finding the ground energy of the corresponding LCE perturbation $H + V_{\mathcal{J}} \in \mathscr{H}_{\mathrm{LCE}}$ remains classically hard in general.

We first focus on a simpler, deterministic proof to illustrate the idea behind our proof. In doing so, we make use of a major conjecture in complexity theory, namely, the *quantum probabilistically checkable proofs* (QPCPs) conjecture [74,75] and the standard assumption that P $\neq$ QMA. The deterministic proof only guarantees nonexistence of deterministic algorithms that can solve the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problems. Indeed, at the time of writing, it remains an open question whether randomized algorithms are strictly more powerful than deterministic ones, i.e., whether P $\neq$ BPP. Note that this is insufficient to prove total nonexistence of classical algorithms since there could still exist a randomized algorithm solving the problem. Nevertheless, we provide a slightly more sophisticated probabilistic proof (inspired by the technique of the deterministic proof) that does not make use of the QPCP conjecture, and only requires the standard assumption that BPP $\neq$ QMA.

The classical PCP theorem states that every problem in NP has a probabilistically verifiable, classically feasible proof requiring $\mathcal{O}(\log n)$ random bits and only $\mathcal{O}(1)$ bits of the proof to be read [76]. The quantum analog, i.e., QPCP, is believed to hold true but no proof has been found at the time of writing. Most notably, the QPCP conjecture is mathematically equivalent to the question of whether the local Hamiltonian problem with *at least constant* promise gaps (rather than at least inversely polynomial gaps [54]) is QMA hard [74] or not.

a. The local Hamiltonian problem and classical simulability

Before we prove these claims, we set the context by formally introducing the necessary concepts. First, we clarify what we mean by the *local Hamiltonian problem* in the sense of Ref. [54].

Definition J1 (k-local Hamiltonian problem). We are given a decomposition

$$H = \sum_{j=1}^{r} H_j$$

of an N-qubit Hamiltonian $H \in \mathrm{Herm}(2^N)$ with k-local Hermitian terms $H_j \in \mathrm{Herm}(2^N)$ (i.e., each H_j acts on at most k qubits) with $\Theta(\mathrm{Poly}(N))$-bit description of its entries such that $\|H_j\| = \mathcal{O}(\mathrm{Poly}(N))$ for all $j \in [r]$, where $r = \Theta(\mathrm{Poly}(N))$. Let $\lambda_{\min}(H) \in \mathbb{R}$ denote the smallest eigenvalue of H. The k-local Hamiltonian problem is the task of deciding whether

$$\texttt{YES:}\ \lambda_{\min}(H) < a \quad \text{or} \quad \texttt{NO:}\ \lambda_{\min}(H) > b, \quad \text{(J4)}$$

for some given rational numbers $a, b \in \mathbb{Q}$ of $\Theta(\mathrm{Poly}(N))$-bit description with $a < b$ and promise gap satisfying

$$b - a = \Omega(\mathrm{Poly}(N)^{-1}). \quad \text{(J5)}$$

The most interesting condition of the k-local Hamiltonian problem is the (asymptotic) size of the promise gap $b - a > 0$

as $N \to \infty$. For larger gaps, the problem of deciding YES or NO generally becomes easier, which is the main difficulty in the QPCP conjecture, namely, whether a constant promise gap $b - a = \Omega(1)$ is already sufficient for QMA hardness [74]. On the other hand, decreasing the promise gap $b - a$ only makes the decision problem harder. For the purpose of our analysis, since the promise gap is at least inversely polynomial in N, we may assume without loss of generality that $\|H\| = \mathcal{O}(1)$ since one may rescale $H \mapsto \frac{1}{r}H$, and due to $r = \Theta(\mathrm{Poly}(N))$, this transformation asymptotically leaves the promise gap invariant as $N \to \infty$.

Note that a sufficiently good approximation of $\lambda_{\min}(H)$ can be used to decide Eq. (J4). This specifically motivates what we mean by a BPP and P algorithms, respectively, which can solve the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem.

Definition J2. A BPP algorithm $\mathcal{A}$ is said to solve the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem if for all $H \in \mathscr{H}_{\mathrm{LCE}}$ and all $\varepsilon > 0$, the algorithm $\mathcal{A}$ running on a probabilistic Turing machine outputs $\lambda \in \mathbb{R}$ such that

$$|\lambda - \lambda_{\min}(H)| < \varepsilon,$$

with success probability strictly larger than $\frac{1}{2}$, and does so in polynomial time, $\mathcal{O}(\mathrm{Poly}(N, \frac{1}{\varepsilon}))$. If $\mathcal{A}$ solves the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem on a deterministic Turing machine (i.e., with success probability 1), then $\mathcal{A}$ is called a P algorithm that solves the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem.

By definition, it follows that if such an algorithm $\mathcal{A}$ exists, then the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem lives in the complexity class BPP, and analogously in P if $\mathcal{A}$ defines a P algorithm.

b. Deterministic proof using P ≠ QMA and the QPCP conjecture

We begin with the more straightforward, deterministic proof that relies on the standard assumption that P ≠ QMA and additionally assumes the QPCP conjecture to be true. First, we establish the following approximation argument, which intuitively establishes that any normalized Hamiltonian can be perturbed with small deviation from the unperturbed ground energy.

Lemma J1. Let $H \in \mathrm{Herm}(2^N)$ be an arbitrary Hamiltonian with $\|H\| = \mathcal{O}(1)$. Then, for all $\varepsilon > 0$ independent of N [i.e., $\varepsilon = \Theta(1)$] there exists $\tilde{H} \in \mathscr{H}_{\mathrm{LCE}}$ such that

$$|\lambda_{\min}(\tilde{H}) - \lambda_{\min}(H)| < \varepsilon.$$

Proof. Let $\varepsilon > 0$ be arbitrary but fixed, which is $\varepsilon = \Theta(1)$. Let $H \in \mathrm{Herm}(2^N)$ be arbitrary with $\|H\| = \mathcal{O}(1)$. The central idea is to perturb H with a deterministic LCE perturbation $V_{\mathcal{J}}$ with operator norm smaller than $\varepsilon > 0$. More precisely, for some nonempty $\mathcal{J} \subseteq \{0, 1, 2, 3\}^N \setminus \{0\}$ with $|\mathcal{J}| = \mathcal{O}(1)$ we define

$$\tilde{H} := H + V_{\mathcal{J}}, \quad \text{where} \quad V_{\mathcal{J}} := \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} a_j Q_j, \tag{J6}$$

with $a_j \in (-\varepsilon, \varepsilon) \setminus \{0\}$ [in particular, $a_j = \Theta(1)$ since $\varepsilon = \Theta(1)$ by assumption], and arbitrary nonidentity Pauli strings Q_j indexed by $j \in \mathcal{J}$. By construction, we have that $V_{\mathcal{J}}$ defines an LCE perturbation in the sense of Eq. (J1). Indeed, this follows from $a_j \in \Theta(1)$ for all $j \in \mathcal{J}$, and $|\mathcal{J}| = \Theta(1)$ due to $\mathcal{J} \neq \emptyset$ being a discrete set. In particular, $\tilde{H} \in \mathscr{H}_{\mathrm{LCE}}$ by using the assumption that $\|H\| = \mathcal{O}(1)$. Furthermore, by the triangle inequality,

$$\sup_{|\psi\rangle} |\langle\psi|V_{\mathcal{J}}|\psi\rangle| \leqslant \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \underbrace{|a_j|}_{<\varepsilon} \underbrace{\sup_{|\psi\rangle} |\langle\psi|Q_j|\psi\rangle|}_{\leqslant 1} < \frac{\varepsilon}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} 1 = \varepsilon, \tag{J7}$$

using the fact that Pauli strings have eigenvalues in ± 1, and thus uniformly bounded operator norm $\|Q_j\| = 1$ for all $j \in \mathcal{J}$. As a result, by the Weyl inequality,

$$|\lambda_{\min}(\tilde{H}) - \lambda_{\min}(H)| \overset{(J6)}{=} |\lambda_{\min}(H + V_{\mathcal{J}}) - \lambda_{\min}(H)| \overset{\text{Weyl}}{\leqslant} \sup_{|\psi\rangle} |\langle\psi|V_{\mathcal{J}}|\psi\rangle| \overset{(J7)}{<} \varepsilon.$$

■

We are now going to show that there exists no P algorithm that can solve the $\mathscr{H}_{\mathrm{LCE}}$-class problem, unless the QPCP conjecture is false or P = QMA.

Proposition J1. If the QPCP conjecture holds true and if P ≠ QMA, then there exists no deterministic classical algorithm that can solve the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem. In particular, the problem of finding the ground energy of an arbitrary Hamiltonian $H \in \mathscr{H}_{\mathrm{LCE}}$ is guaranteed to be classically infeasible on deterministic machines in general.

Proof. Assume that P ≠ QMA and the QPCP conjecture to be true. Suppose by contradiction that there exists a P algorithm $\mathcal{A}$ that can solve the $\mathscr{H}_{\mathrm{LCE}}$-class Hamiltonian problem. Let $\varepsilon > 0$ to be determined and $H \in \mathrm{Herm}(2^N)$ an arbitrary Hamiltonian with $\|H\| = \mathcal{O}(1)$. Let $a, b \in \mathbb{Q}$ be given boundaries of a promise gap for $\lambda_{\min}(H)$ according to Eq. (J4). Crucially, the QPCP conjecture allows us to assume that $b - a = \Theta(1)$. By Lemma J1, we know that there exists $\tilde{H} \in \mathscr{H}_{\mathrm{LCE}}$ such that

$$|\lambda_{\min}(\tilde{H}) - \lambda_{\min}(H)| < \frac{\varepsilon}{2}. \tag{J8}$$

Furthermore, by the assumption of classicality of $\mathscr{H}_{\mathrm{LCE}}$, we know that the P algorithm $\mathcal{A}$ outputs $\lambda \in \mathbb{R}$ such that

$$|\lambda_{\min}(\tilde{H}) - \lambda| < \frac{\varepsilon}{2}. \tag{J9}$$

Therefore, by combining the two inequalities and choosing $\varepsilon := \frac{b-a}{2} = \Theta(1)$, we find that

$$\begin{cases} \lambda < \lambda_{\min}(\tilde{H}) + \frac{\varepsilon}{2} < \lambda_{\min}(H) + \varepsilon < a + \varepsilon = \frac{a+b}{2}, & \text{if } \lambda_{\min}(H) < a, \\ \lambda \overset{(J9)}{>} \lambda_{\min}(\tilde{H}) - \frac{\varepsilon}{2} \overset{(J8)}{>} \lambda_{\min}(H) - \varepsilon > b - \varepsilon = \frac{a+b}{2}, & \text{if } \lambda_{\min}(H) > b. \end{cases} \tag{J10}$$

As a result, the P algorithm $\mathcal{A}$ successfully solves the decision problem in Eq. (J4) by deciding

$$\begin{cases} \texttt{YES}, & \text{if } \lambda < \frac{a+b}{2}, \\ \texttt{NO}, & \text{if } \lambda > \frac{a+b}{2}. \end{cases} \tag{J11}$$

Since H with $\|H\| = \mathcal{O}(1)$ and $a, b \in \mathbb{Q}$ with $b - a = \Theta(1)$ were both arbitrary, it especially follows that the k-local Hamiltonian problem with constant promise gaps is in P for every $k \geqslant 2$. On the other hand, the point of this argument is using the QPCP conjecture, which states that the local Hamiltonian problem with constant promise gaps $b - a = \Omega(1)$ is QMA hard [74,75]; i.e., it is at least as hard as all QMA problems. Since $\texttt{P} \subseteq \texttt{QMA}$ [73], we thus must have that $\texttt{P} = \texttt{QMA}$. This is the desired contradiction to our assumption that $\texttt{P} \neq \texttt{QMA}$. Consequently, no such classical algorithm $\mathcal{A}$ can exist. ■

c. *Probabilistic proof using BPP ⊊ QMA only*

Here, we prove that one does not require the QPCP conjecture in order to show that there exists no classical (probabilistic) BPP algorithm that can solve the $\mathscr{H}_{\text{LCE}}$-class Hamiltonian problem, provided the standard assumption that $\texttt{BPP} \neq \texttt{QMA}$. First, we need to make the following technical reduction for the probabilistic proof to work flawlessly:

Lemma J2. Let $\mathscr{H} \subseteq \text{Herm}(2^N)$ be an arbitrary class of Hamiltonians and $\eta = \Theta(1)$ a constant. Suppose that solving the Hamiltonian decision problem of the form in Eq. (J4) with promise gap $b - a > 0$ for the class $\mathscr{H}$ has some complexity $\mathcal{C} \in \{\texttt{P}, \texttt{BPP}, \texttt{NP}, \texttt{BQP}, \texttt{QMA}, \dots\}$. Then, solving the decision problem for the affine class $\mathscr{H}' := \{H + \eta I_{2^N} : H \in \mathscr{H}\}$ leaves the promise gap invariant and thus has exactly the same complexity $\mathcal{C}$, where I_{2^N} denotes the $2^N \times 2^N$ identity matrix. In particular, the 2-local Hamiltonian problem remains QMA complete after such an affine transformation.

Proof. Let $\mathscr{H} \subseteq \text{Herm}(2^N)$ be an arbitrary Hamiltonian class and let $\mathcal{C}$ denote the complexity of the Hamiltonian decision problem with promise gap $b - a > 0$ as stated in the lemma above. Let $\eta = \Theta(1)$ and $H \in \mathscr{H}$ be arbitrary and define $\tilde{H} := H + \eta I_{2^N}$. Recall from linear algebra that the eigenvalues of H are defined as the roots of the characteristic polynomial $P_H(X) = \det(H - XI_{2^N}) \in \mathbb{C}[X]$ in the variable X. Therefore, the affine transformation $H \mapsto \tilde{H} = H + \eta I_{2^N}$ has the effect of shifting the eigenspectrum by η. In particular,

$$\lambda_{\min}(\tilde{H}) = \lambda_{\min}(H) + \eta, \tag{J12}$$

with an accordingly transformed promise gap for the transformed decision problem given by $\tilde{a} := a + \eta$ and $\tilde{b} := b + \eta$. Notably, this affine transformation leaves the promise gap invariant:

$$\tilde{b} - \tilde{a} = (b + \eta) - (a + \eta) = b - a. \tag{J13}$$

Therefore, the transformed problem, i.e., the Hamiltonian problem for the class $\mathscr{H}'$ has exactly the same complexity as the original problem, namely, $\mathcal{C}$. ■

Lemma J2 allows us to assume without loss of generality that either $a = \Theta(1)$ or $b = \Theta(1)$ is constant in the formulation of the 2-local Hamiltonian problem in Eq. (J4). Now, the idea of our probabilistic proof is to construct a randomly sampled LCE perturbation $V_{\mathcal{J}}$ by using the promise gap boundaries a and b, which are *a priori* given by the problem statement. As detailed below, since at least one of them is constant, this ensures that $V_{\mathcal{J}}$ indeed defines an LCE perturbation. We are now ready to prove nonexistence of BPP algorithms that can solve the $\mathscr{H}_{\text{LCE}}$-class Hamiltonian problem.

Theorem J1. If $\texttt{BPP} \neq \texttt{QMA}$, then there exists no BPP algorithm that can solve the $\mathscr{H}_{\text{LCE}}$-class Hamiltonian problem. In particular, the problem of finding the ground energy of an arbitrary Hamiltonian $H \in \mathscr{H}_{\text{LCE}}$ is guaranteed to be classically infeasible in general.

Proof. Assume that $\texttt{BPP} \neq \texttt{QMA}$ and suppose by contradiction that there exists a BPP algorithm $\mathcal{A}$ that can solve the $\mathscr{H}_{\text{LCE}}$-class Hamiltonian problem. Let $\varepsilon > 0$ to be determined and let $H \in \text{Herm}(2^N)$ be an arbitrary Hamiltonian with $\|H\| = \mathcal{O}(1)$. Let $a, b \in \mathbb{Q}$ be the given rational boundaries of the promise gap $b - a = \Omega(\text{Poly}(N)^{-1})$. Without loss of generality, we may assume that $b - a \to 0$ as $N \to \infty$; otherwise, $b - a = \Theta(1)$ is constant and the theorem follows directly from $\texttt{P} \subseteq \texttt{BPP}$ and Proposition J1 by an analogous deterministic LCE perturbation $V_{\mathcal{J}}$ without having to assume the QPCP conjecture (since the promise gap is already constant in that case). Thus, it suffices to handle the case where the promise gap $b - a$ vanishes as $N \to \infty$.

First, we construct a randomly sampled LCE perturbation $V_{\mathcal{J}}$ to establish a probabilistic analog of Lemma J1 with sufficiently high probability for compatibility with our BPP algorithm $\mathcal{A}$. Here, we employ the reduction argument of Lemma J2, allowing us to assume that $a = \Theta(1)$ or $b = \Theta(1)$. Therefore,

$$V_{\mathcal{J}} := c_0 \sum_{j \in \mathcal{J}} \underbrace{\left(a_j - \frac{a+b}{2}\right)}_{=\Theta(1)} Q_j$$

almost surely defines an LCE perturbation for $a_j \overset{\text{i.i.d.}}{\sim} \text{Unif}([a, b])$, where $Q_j \in \mathcal{P}_N \setminus \{I\}$ are arbitrary nonidentity Pauli strings indexed by $\mathcal{J}$ such that $|\mathcal{J}| = \mathcal{O}(1)$, where $c_0 = \Theta(1)$ is an arbitrarily small constant that is chosen toward the end of the proof to control the probability. In particular

$$\tilde{H} := H + V_{\mathcal{J}} \in \mathscr{H}_{\text{LCE}}. \tag{J14}$$

The next step is to compute the mean and variance of the largest quantum measurement outcome of the LCE perturbation $V_{\mathcal{J}}$ with respect to the randomness of a_j in order concentrate the probability of the desired event that $|\lambda_{\min}(\tilde{H}) - \lambda_{\min}(H)|$ is sufficiently small. For that purpose, let $|\psi\rangle \in \mathcal{H}_N$ be an arbitrary N-qubit state. Then, the expected measurement outcome of the LCE perturbation $V_{\mathcal{J}}$ has mean

$$\mathbb{E}[\langle\psi|V_{\mathcal{J}}|\psi\rangle] = c_0 \sum_{j \in \mathcal{J}} \underbrace{\left(\mathbb{E}[a_j] - \frac{a+b}{2}\right)}_{=0} Q_j = 0, \tag{J15}$$

by construction since

$$\begin{aligned} \mathbb{E}[a_j] &= \frac{1}{b-a} \int_a^b x\, dx = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} \\ &= \frac{1}{b-a} \cdot \frac{(b-a)(b+a)}{2} = \frac{a+b}{2}, \end{aligned} \tag{J16}$$

for all $j \in \mathcal{J}$. Therefore, since the operator norm of Pauli matrices is equal to 1, which is $\langle\psi|Q_j|\psi\rangle \leqslant 1$ for all $j \in \mathcal{J}$, the expected measurement outcome of the LCE perturbation $V_{\mathcal{J}}$ has variance

$$\begin{aligned}\mathrm{Var}[\langle\psi|V_{\mathcal{J}}|\psi\rangle] &= \mathbb{E}[\langle\psi|V_{\mathcal{J}}|\psi\rangle^2] = c_0^2 \sum_{j,j'\in\mathcal{J}} \mathbb{E}\left[\left(a_j - \frac{a+b}{2}\right)\left(a_{j'} - \frac{a+b}{2}\right)\right]\langle\psi|Q_jQ_{j'}|\psi\rangle \\ &= c_0^2 \sum_{j\in\mathcal{J}} \mathbb{E}\left[\left(a_j - \frac{a+b}{2}\right)^2\right]\underbrace{\langle\psi|Q_j|\psi\rangle^2}_{\leqslant 1} + c_0^2 \sum_{\substack{j,j'\in\mathcal{J}\\ j\neq j'}} \underbrace{\left(\mathbb{E}[a_j] - \frac{a+b}{2}\right)}_{\overset{(J16)}{=}0}\underbrace{\left(\mathbb{E}[a_{j'}] - \frac{a+b}{2}\right)}_{\overset{(J16)}{=}0}\langle\psi|Q_j|\psi\rangle\langle\psi|Q_{j'}|\psi\rangle \\ &\leqslant c_0^2 \sum_{j\in\mathcal{J}} \mathbb{E}\left[\left(a_j - \frac{a+b}{2}\right)^2\right] = \frac{c_0^2(b-a)^2}{12}\sum_{j\in\mathcal{J}} 1 = \frac{c_0^2|\mathcal{J}|}{12}(b-a)^2 \overset{N\to\infty}{\to} 0, \end{aligned} \tag{J17}$$

recalling that the promise gap $b - a > 0$ vanishes as $N \to \infty$, where in the second-to-last equation we computed

$$\begin{aligned}\mathbb{E}\left[\left(a_j - \frac{a+b}{2}\right)^2\right] &= \mathbb{E}\big[a_j^2\big] - 2\underbrace{\mathbb{E}[a_j]}_{\overset{(J16)}{=}\frac{a+b}{2}}\frac{a+b}{2} + \frac{(a+b)^2}{4} = \mathbb{E}\big[a_j^2\big] - \frac{(a+b)^2}{4} \\ &= \underbrace{\frac{a^2+ab+b^2}{3}}_{=\mathbb{E}[a_j^2]} + \frac{a^2+2ab+b^2}{4} = \frac{1}{12}(4a^2+4ab+4b^2-3a^2-6ab-3b^2) \\ &= \frac{1}{12}(a^2-2ab+b^2) = \frac{(b-a)^2}{12} \end{aligned} \tag{J18}$$

with second moments

$$\mathbb{E}\big[a_j^2\big] = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\frac{b^3-a^3}{3} = \frac{1}{b-a}\frac{(b-a)(b^2+ab+a^2)}{3} = \frac{a^2+ab+b^2}{3} \tag{J19}$$

for all $j \in \mathcal{J}$.

This means that both the mean and variance with respect to a_j of the expected quantum measurement outcome of $V_{\mathcal{J}}$ are independent of $|\psi\rangle$ by construction. Now, let us again take $\varepsilon := \frac{b-a}{2} \overset{N\to\infty}{\to} 0$ so that by the Weyl and Chebyshev inequalities,

$$\begin{aligned}\mathbb{P}_a\left(|\lambda_{\min}(\tilde{H}) - \lambda_{\min}(H)| < \frac{\varepsilon}{2}\right) &\overset{(J14)}{=} \mathbb{P}_a\left(|\lambda_{\min}(H+V_{\mathcal{J}}) - \lambda_{\min}(H)| < \frac{\varepsilon}{2}\right) \overset{\text{Weyl}}{\geqslant} \mathbb{P}_a\left(\sup_{|\psi\rangle}|\langle\psi|V_{\mathcal{J}}|\psi\rangle| < \frac{\varepsilon}{2}\right) \\ &\geqslant 1 - \frac{4}{\varepsilon^2}\,\mathrm{Var}\left[\sup_{|\psi\rangle}\langle\psi|V_{\mathcal{J}}|\psi\rangle\right] \overset{(J17)}{=} 1 - \frac{(b-a)^2}{3\varepsilon^2}c_0^2|\mathcal{J}| = 1 - \underbrace{\frac{4}{3}c_0^2|\mathcal{J}|}_{=:\delta_0}, \end{aligned} \tag{J20}$$

by our choice of $\varepsilon > 0$. In particular, since $|\mathcal{J}| = \mathcal{O}(1)$, the error between the perturbed ground energy $\lambda_{\min}(\tilde{H})$ and the exact solution $\lambda_{\min}(H)$ is bounded by a fraction of the vanishing promise gap $\mathcal{O}(b-a)$ with arbitrarily high probability by choosing the constant $c_0 > 0$ to be sufficiently small. Note that this argument only works due to the fact that $|\mathcal{J}| = \mathcal{O}(1)$ since we require that $c_0 = \Theta(1)$ for a valid perturbation $V_{\mathcal{J}}$.

The final step is to show that the BPP algorithm $\mathcal{A}$ successfully solves the decision problem in Eq. (J4) and conclude with a contradiction. By the assumption on the BPP algorithm $\mathcal{A}$, we know that

$$|\lambda_{\min}(\tilde{H}) - \lambda| < \frac{\varepsilon}{2}, \tag{J21}$$

with probability p strictly larger than $\frac{1}{2}$ in feasible time, using $\varepsilon = \frac{b-a}{2} = \Omega(\mathrm{Poly}(N)^{-1})$. That is, Eq. (J21) occurs with probability at least $p = \frac{1}{2} + \delta$ for some fixed $\delta \in (0, \frac{1}{2}]$. Hence, combining both inequalities results into

$$\begin{cases} \lambda < \lambda_{\min}(\tilde{H}) + \frac{\varepsilon}{2} < \lambda_{\min}(H) + \varepsilon < a + \varepsilon = \frac{a+b}{2}, & \text{if } \lambda_{\min}(H) < a, \\ \lambda \overset{(J21)}{>} \lambda_{\min}(\tilde{H}) - \frac{\varepsilon}{2} \overset{(J20)}{>} \lambda_{\min}(H) - \varepsilon > b - \varepsilon = \frac{a+b}{2}, & \text{if } \lambda_{\min}(H) > b, \end{cases} \tag{J22}$$

with probability at least $\frac{1}{2} + (\delta - \delta_0)$, which is arbitrarily close to $\frac{1}{2}$ by choosing $c_0 > 0$ to be sufficiently small as shown below. Indeed, to justify the lower bound of the probability of Eq. (J22) occurring, let us denote the probabilistic events

$$\mathcal{E}_1 := \left\{|\lambda_{\min}(\tilde{H}) - \lambda| < \frac{\varepsilon}{2}\right\} \quad \text{and} \quad \mathcal{E}_2 := \left\{|\lambda_{\min}(\tilde{\mathcal{H}}) - \lambda_{\min}(H)| < \frac{\varepsilon}{2}\right\}. \tag{J23}$$

Then, by elementary probability theory,

$$\mathbb{P}(\mathcal{E}_1 \text{ and } \mathcal{E}_2) = \underbrace{\mathbb{P}(\mathcal{E}_1)}_{\geqslant \frac{1}{2}+\delta} + \underbrace{\mathbb{P}(\mathcal{E}_2)}_{\overset{(J20)}{\geqslant} 1-\delta_0} - \underbrace{\mathbb{P}(\mathcal{E}_1 \text{ or } \mathcal{E}_2)}_{\leqslant 1}$$
$$\geqslant \frac{1}{2} + \underbrace{(\delta - \delta_0)}_{\overset{!}{>}0},$$

with respect to the joint randomness of the BPP algorithm $\mathcal{A}$ and the uniformly sampled coefficients a_j, as claimed. Crucially, the difference $\delta - \delta_0 > 0$ can be made arbitrarily small by choosing a suitable constant $c_0 > 0$ in the definition of δ_0. This ensures that the probability of Eq. (J22) occurs with a probability strictly larger than $\frac{1}{2}$ in accordance with Definition J2.

As a consequence of Eq. (J22), the BPP algorithm $\mathcal{A}$ successfully solves the decision problem in Eq. (J4) by deciding

$$\begin{cases} \texttt{YES}, & \text{if } \lambda < \frac{a+b}{2}, \\ \texttt{NO}, & \text{if } \lambda > \frac{a+b}{2}, \end{cases} \tag{J24}$$

with probability at least $\frac{1}{2} + \delta'$ for $\delta' := \delta - \delta_0 > 0$, which can be made arbitrarily small by choosing the constant $c_0 > 0$ accordingly.

Since H with $\|H\| = \mathcal{O}(1)$ and $a, b \in \mathbb{Q}$ with $b - a = \Omega(\text{Poly}(N)^{-1})$ were both arbitrary, it especially follows that the 2-local Hamiltonian problem with inversely polynomial promise gaps is in BPP. Due to the fact that the 2-local Hamiltonian problem is QMA complete [54], and that $\texttt{BPP} \subseteq \texttt{QMA}$ [73], we thus find $\texttt{BPP} = \texttt{QMA}$, which is the desired contradiction to our assumption that $\texttt{BPP} \neq \texttt{QMA}$. Therefore, there cannot exist such a classical algorithm $\mathcal{A}$. ■

We have thus rigorously proven that there exists no classical algorithm that can find the ground energy of observables in the class $\mathscr{H}_{\text{LCE}}$ of LCE Hamiltonians, either provided that $\texttt{BPP} \neq \texttt{QMA}$. Note that we only made a statement about nonclassicality and gradient scalability at initialization via LCE perturbations. In particular, there is no guarantee that a VQA can generally find the ground energy of a Hamiltonian in $\mathscr{H}_{\text{LCE}}$ due to the possibility of solutions θ_* existing outside landscape patches on which LCE fails to exhibit scalable gradient signals. It remains an open question of whether there exist Hamiltonians in $\mathscr{H}_{\text{LCE}}$ with VQA solutions θ_* lying inside LCE-effective landscape patches for which no classical algorithm exists. Since this work focuses on the theory of initialization only, we leave this question as an open problem.

2. Deterministic worst-case simulation complexity

The next parts leverage the results of the previous Appendix, and focuse on the computational complexity of simulating the cost function $C(\theta)$ in Eq. (E1) with the help of the Taylor surrogate $C_m(\theta)$ in Eq. (D9). We shall use $\text{Time}\, C_m(\theta)$ to denote the time complexity of evaluating $C_m(\theta)$, which is, by definition, the number of operations required for evaluating the expression for $C_m(\theta)$.

To quantify the time $\text{Time}\, C_m(\theta)$ required to evaluate $C_m(\theta)$ classically, we begin with the deterministic error analysis, where we are able to highlight a computational phase transition from polynomial to superpolynomial complexity, which is characterized by the ℓ_1 norm $\|\theta\|_1$ of the parameter configuration.

Theorem 3 (Deterministic complexity). Consider an expectation of the form Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series $C_m(\theta)$ defined in Eq. (D9) with truncation threshold $m = m(\theta)$ given by Lemma I3. Then, if $\|\theta\|_1 = \mathcal{O}(1)$, the run-time of computing $C_m(\theta)$ is polynomial, namely, $\mathcal{O}(|\mathcal{I}|D^m)$. On the other hand, if $\|\theta\|_1 = o(D)$ strictly increases sublinearly as $D \to \infty$, and $\#\{k \in [D] : \exists \xi \in \frac{\pi}{2}\mathbb{Z} : |\theta_k - \xi| > \Omega(D^{-1})\} = \Omega(D)$, then the run-time becomes superpolynomial,

$$\text{Time}\, C_m(\theta) = \Omega\left(\text{Poly}(D)^{-1}\left(\frac{D}{\|\theta\|_1}\right)^{e\|\theta\|_1}\right),$$

and is at most $\mathcal{O}(|\mathcal{I}|D\, 2^{2D})$ in the worst case.

Proof. By the higher-order parameter-shift rule (Lemma E2), we may write

$$\begin{aligned} C_m(\theta) &= \sum_{\|\alpha\|_1 < m} \left(D_\theta^\alpha C\right)(0)\, \frac{\theta^\alpha}{\alpha!} \\ &= \sum_{\|\alpha\|_1 < m} \frac{\theta^\alpha}{\alpha!} \sum_{\substack{j \in \mathbb{N}_0^D \\ j_k \leqslant \alpha_k\, \forall k}} (-1)^{|j|} \binom{\alpha}{j} C(\xi_{j,\alpha}) \\ &= \sum_{(j,\alpha) \in \Lambda_m} \frac{\theta^\alpha}{\alpha!} (-1)^{|j|} \binom{\alpha}{j} C(\xi_{j,\alpha}), \end{aligned} \tag{J25}$$

where $\xi_{j,\alpha} := \frac{\pi}{2} \sum_{k=1}^D (\alpha_k - 2j_k) e_k$, with e_k denoting the kth unit vector, and indexing set

$$\Lambda_m := \big\{(j,\alpha) \in \mathbb{N}_0^D \times \mathbb{N}_0^D : \|\alpha\|_1 < m,\ \forall k \in [D] : 0 \leqslant j_k \leqslant \alpha_k\big\}. \tag{J26}$$

The computational complexity for each of the evaluation of $C(\xi_{j,\alpha})$ using Gottesman-Knill is of order $\mathcal{O}(D\,|\mathcal{I}|)$ [18,19], which is the most expensive computation in our simulation strategy. At the end of the proof, we will justify why Gottesman-Knill may be applied for each $\xi_{j,\alpha}$. As a result, the time resources required to compute $C_m(\theta)$ scales like

$$\begin{aligned} \text{Time}\, C_m(\theta) &= \mathcal{O}(D\,|\mathcal{I}|) \cdot \#\{C(\xi_{j,\alpha}) : (j,\alpha) \in \Lambda_m\} \\ &\leqslant \mathcal{O}(D\,|\mathcal{I}|)\, \#\Lambda_m = \mathcal{O}(D\,|\mathcal{I}|) \sum_{\|\alpha\|_1 < m} \underbrace{\prod_{k=1}^D (\alpha_k + 1)}_{\leqslant 2^{\|\alpha\|_1} \leqslant 2^{m-1}} \\ &\leqslant \mathcal{O}\left(D\,|\mathcal{I}|\, 2^m \sum_{\|\alpha\|_1 < m} 1\right). \end{aligned} \tag{J27}$$

If $\|\theta\|_1 = \mathcal{O}(1)$, then we know that $m = m(\theta) = \mathcal{O}(1)$ by Lemma I3. Therefore, we may employ Eq. (I19) in Lemma I4 to establish the polynomial complexity,

$$\text{Time}\, C_m(\theta) = \mathcal{O}\left(D\,|\mathcal{I}| \sum_{\|\alpha\|_1 < m} 1\right) = \mathcal{O}(|\mathcal{I}|D^m), \tag{J28}$$

as claimed.

Next, we will prove the superpolynomial case where $\|\theta\|_1 = o(D)$ strictly increases sublinearly as $D \to \infty$. Here is where we will use the assumption that $\#\{k \in [D] : \exists \xi \in \frac{\pi}{2}\mathbb{Z} : |\theta_k - \xi| > \Omega(D^{-1})\} = \Omega(D)$; otherwise, since $U(\xi) \in \mathcal{C}_N$ is Clifford for all $\frac{\pi}{2}\mathbb{Z}$ (as shown at the end of the proof), one may adapt the Clifford architecture of $U(\cdot)$ to obtain a PQC $\tilde{U}(\cdot)$ such that $\tilde{U}(0) = U(\xi)$, and as a result, find a translational change of coordinates $\theta \mapsto \tilde{\theta}$ such that $\|\tilde{\theta}\|_1 = \mathcal{O}(1)$, giving rise to a polynomial time Taylor surrogate by the previous case. Therefore, this assumption is made in order to rule out classical simulability beyond the case $\|\theta\|_1 = \mathcal{O}(1)$ by ensuring that most of the rotation gates $R_{V_k}(\theta_k)$ are sufficiently far from a single-qubit Clifford gate.

We continue by proving the superpolynomial lower bound. The computational cost of evaluating the Taylor surrogate in Eq. (J25) is lower bounded by

$$\begin{aligned}\mathrm{Time}C_m(\theta) &\geqslant \min_{(j,\alpha)\in\Lambda_m} \underbrace{\mathrm{Time}C(\xi_{j,\alpha})}_{\geqslant 1} \cdot \sum_{(j,\alpha)\in\Lambda_m} 1 \\ &\geqslant \sum_{\|\alpha\|_1<m} \{k \in [D] : 0 \leqslant j_k \leqslant \alpha_k\} \\ &= \sum_{\|\alpha\|_1<m} \prod_{k=1}^{D} \underbrace{(\alpha_k+1)}_{\geqslant 1} \geqslant \sum_{\|\alpha\|_1<m} 1 \\ &= \Omega\left(\mathrm{Poly}(D)^{-1}\left(\frac{D}{\|\theta\|_1}\right)^{e\|\theta\|_1}\right), \end{aligned} \tag{J29}$$

where the last equality follows from Lemma I4 and applying Lemma I3, which tells us that $m = m(\theta) \sim e\|\theta\|_1$ as $D \to \infty$.

What one could think about is whether it is possible to significantly reduce the cardinality of Λ_m by using the fact that $C(\cdot)$ is 2π periodic in each direction in parameter space.

Indeed, note that $(\xi_{j,\alpha})_k$ corresponds to a shift of 2π in the kth coordinate if and only if $\alpha_k - 2j_k = 4$. Therefore, $C(\xi_{j,\alpha})$ depends only on $\alpha_k - 2j_k \bmod 4$ for all $k \in [D]$. In other words,

$$\begin{aligned}&\text{for each } (j,\alpha) \in \Lambda_m, \text{ there exists } \beta_{j,\alpha} \in \{0,1,2,3\}^D \\ &\text{such that } C(\xi_{j,\alpha}) = C\left(\frac{\pi}{2}\beta_{j,\alpha}\right). \end{aligned} \tag{J30}$$

This particularly means that in the worst-case scenario, we require at most $4^D = 2^{2D}$ different evaluations of $C(\xi_{j,\alpha})$. As a result, it is possible to optimize the worst-case run-time to achieve

$$\mathrm{Time}\, C_m(\theta) = \mathcal{O}(|\mathcal{I}|\, D\, 2^{2D}), \tag{J31}$$

as claimed. Note that this scales worse compared to the computational complexity $\mathcal{O}(|\mathcal{I}|\, 2^{2N})$ of naively evaluating the exponential quadratic form in Eq. (6), since high expressivity requires that $N \ll D$.

It remains to justify the application of Gottesman-Knill's theorem for each $\xi_{j,\alpha}$. It suffices to show that $U(\xi_{j,\alpha}) \in \mathrm{Clifford}(N)$ because of $C(\xi_{j,\alpha}) = \langle 0|U(\xi_{j,\alpha})^\dagger \hat{O} U(\xi_{j,\alpha})|0\rangle$ [where $\hat{O}$ denotes the observable in Eq. (5) here in order to avoid confusion with the Hadamard gate]. Indeed, note that $e^{-i\frac{\pi}{4}} R_X(\frac{\pi}{2}) = HSH \in \mathrm{Clifford}(N)$, $R_Y(\frac{\pi}{2}) = XH \in \mathrm{Clifford}(N)$, and $e^{-i\frac{\pi}{4}} R_Z(\frac{\pi}{2}) = S \in \mathrm{Clifford}(N)$ (global phases are irrelevant). This particularly means for multiples of $\frac{\pi}{2}$ that $R_X(\frac{\pi}{2}\mathbb{Z}), R_Y(\frac{\pi}{2}\mathbb{Z}), R_Z(\frac{\pi}{2}\mathbb{Z}) \subset \mathrm{Clifford}(N)$, respectively. This follows from the fact that $\mathrm{Clifford}(N)$ defines a group. As a result, since $\frac{\pi}{2}(\alpha_k - 2j_k) \in \frac{\pi}{2}\mathbb{Z}$ is also multiple of $\frac{\pi}{2}$, we find

$$\begin{aligned}U(\xi_{j,\alpha}) \in \left\langle R_X\left(\frac{\pi}{2}\mathbb{Z}\right), R_Y\left(\frac{\pi}{2}\mathbb{Z}\right), R_Z\left(\frac{\pi}{2}\mathbb{Z}\right), \mathrm{Clifford}(N)\right\rangle_N \\ \subset \mathrm{Clifford}(N). \end{aligned} \tag{J32}$$

By the Gottesman-Knill theorem, we thus know that the cost evaluation $C(\xi_{j,\alpha})$ is easy to simulate classically; i.e., it can be computed in polynomial time $\mathcal{O}(D\,|\mathcal{I}|)$ on a conventional computer [18,19]. ■

3. Probabilistic mean-squared simulation complexity

Next, we use the mean-squared error analysis in order to determine the Gaussian/uniform initialization strategy $\mathbb{P}_\theta$ that covers the largest landscape patch where Taylor surrogating the cost $C(\cdot)$ remains polynomially efficient.

Theorem 4. Consider an expectation of the form in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series $C_m(\theta)$ defined in Eq. (D9). Initialize the parameters $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0, \sigma^2 I_D), \mathrm{Unif}([-\sigma,\sigma]^D)\}$ with i.i.d. components $\theta_1, \ldots, \theta_D$ according to Eq. (19) with constant $c > 0$. Assume further that $\|H\| = \mathcal{O}(1)$. Then, with probability at least $1-\rho$, the computational cost $\mathrm{Time}\, C_m(\theta \sim \mathbb{P}_\theta)$ of simulating $C(\theta)$ up to an error $\varepsilon > 0$, is of polynomial order

$$\mathcal{O}\left(|\mathcal{I}|\, D^{\frac{1}{2\log(\frac{1}{q})}\log(\frac{c}{\rho\,\varepsilon^2})}\right),$$

for nonasymptotic local initialization, i.e., $\sigma = \mathcal{O}(qD^{-\frac{1}{2}})$ for some $0 < q < 1$.

Proof. By the higher-order parameter-shift rule (Lemma E2), we may write

$$\begin{aligned}C_m(\theta) &= \sum_{|\alpha|<m} \left(D_\theta^\alpha C\right)(0)\, \frac{\theta^\alpha}{\alpha!} \\ &= \sum_{|\alpha|<m} \frac{\theta^\alpha}{\alpha!} \sum_{\substack{j\in\mathbb{N}_0^D \\ j_k\leqslant\alpha_k\,\forall k}} (-1)^{|j|} \binom{\alpha}{j} C(\xi_{j,\alpha}) \\ &= \sum_{(j,\alpha)\in\Lambda_m} \frac{\theta^\alpha}{\alpha!} (-1)^{|j|} \binom{\alpha}{j} C(\xi_{j,\alpha}), \end{aligned} \tag{J33}$$

where $\xi_{j,\alpha} := \frac{\pi}{2}\sum_{k=1}^D (\alpha_k - 2j_k) e_k$, with e_k denoting the kth unit vector, and indexing set

$$\begin{aligned}\Lambda_m := \big\{(j,\alpha) \in \mathbb{N}_0^D \times \mathbb{N}_0^D : \|\alpha\|_1 < m,\ \forall k \in [D] : 0 \\ \leqslant j_k \leqslant \alpha_k\big\}. \end{aligned} \tag{J34}$$

The computational complexity for each of the evaluation of $C(\xi_{j,\alpha})$ using Gottesman-Knill is of order $\mathcal{O}(D\,|\mathcal{I}|)$, which is the most expensive computation in our simulation strategy. We have already justified in the proof of Theorem 3 why Gottesman-Knill may be applied for each $\xi_{j,\alpha}$. As a result,

the time resources require to compue $C_m(\theta)$ scales like

$$\begin{aligned}\text{Time}\, C_m(\theta) &= \mathcal{O}(D\,|\mathcal{I}|)\cdot \#\{C(\xi_{j,\alpha}) : (j,\alpha)\in\Lambda_m\}\\ &\leqslant \mathcal{O}(D\,|\mathcal{I}|)\,\#\Lambda_m = \mathcal{O}(D\,|\mathcal{I}|)\sum_{|\alpha|<m}\underbrace{\prod_{k=1}^{D}(\alpha_k+1)}_{\leqslant 2^{\|\alpha\|_1}\leqslant 2^{m-1}}\\ &\leqslant \mathcal{O}\left(D\,|\mathcal{I}|\,2^m\sum_{\|\alpha\|_1<m}1\right). \end{aligned} \tag{J35}$$

For $\|\theta\|_1 = \mathcal{O}(1)$, we have already established polynomial complexity in Eq. (J28), i.e.,

$$\text{Time}\, C_m(\theta) = \mathcal{O}\left(D\,|\mathcal{I}|\sum_{\|\alpha\|_1<m}1\right) = \mathcal{O}(|\mathcal{I}|D^m). \tag{J36}$$

By Lemma I6 we know that, with probability at least $1-\rho$, a constant truncation threshold satisfying

$$m \geqslant \frac{1}{2}\frac{\log\left(\frac{c}{\rho\,\varepsilon^2}\right)}{\log\left(\frac{1}{q}\right)} \tag{J37}$$

is sufficient to simulate $C_m(\theta)$ with error at most $\varepsilon > 0$. ■

4. Complexity comparison with the Pauli path surrogate

We conclude our complexity analysis by showing that our Taylor surrogate outperforms the Pauli path surrogate employed by Lerch *et al.* [21], as long as the patch is sufficiently small.

Corollary J1. Consider an expectation of the form in Eq. (E1) with a variational ansatz $U(\theta)$ involving $D \geqslant 1$ parameters, and suppose we simulate it classically using the Taylor series $C_m(\theta)$ defined in Eq. (D9). Assume further that $\|H\| = \mathcal{O}(1)$. Initialize $\theta \sim \mathbb{P}_\theta \in \{\mathcal{N}(0,\sigma^2 I_D), \text{Unif}([-\sigma,\sigma]^D)\}$ with i.i.d. components $\theta_1, \ldots, \theta_D$, where $\sigma = \mathcal{O}(qD^{-1/2})$ for some $0 < q < 1$, so that

$$\mathbb{E}_\theta[(C(\theta) - C_m(\theta))^2] \leqslant c\,q^{2m},$$

for some constant $c > 0$ as per Lemma I5.

As shown by Lerch *et al.* [Theorem 2 and Eq. (13) in Ref. [21]], with probability at least $1-\rho$, the Pauli-propagation algorithm $\hat{C}(\theta)$ can simulate the cost $C(\theta)$ up to error $\varepsilon > 0$ using

$$\text{Time}\,\hat{C}(\theta\sim\mathbb{P}_\theta) = \gamma_1\,|\mathcal{I}|\,D^{\log(\frac{1}{\rho\varepsilon^2})},$$

many operations for some constant $\gamma_1 > 0$. On the other hand, by Theorem 4, we know that, with probability at least $1-\rho$, the Taylor-simulation algorithm $C_m(\theta)$ can simulate the cost $C(\theta)$ up to error $\varepsilon > 0$ using

$$\text{Time}\, C_m(\theta\sim\mathbb{P}_\theta) = \gamma_2\,|\mathcal{I}|\,D^{\frac{1}{2\log(\frac{1}{q})}\log(\frac{c}{\rho\varepsilon^2})},$$

many operations for some constant $\gamma_2 > 0$.

Then, setting $\gamma := \frac{\gamma_1}{\gamma_2} > 0$, our algorithm outperforms the Pauli-propagation technique, i.e., $\text{Time}\, C_m(\theta) < \text{Time}\,\hat{C}(\theta)$, if and only if

$$q < \exp\left(-\frac{\log\left(\frac{c}{\rho\varepsilon^2}\right)}{2\log\left(\frac{1}{\rho\varepsilon^2}\right) + \frac{2\log\gamma}{\log D}}\right), \tag{J38}$$

provided that

$$D > \exp\left(\frac{\log\gamma}{\log\rho + 2\log\varepsilon}\right) \tag{J39}$$

in the case where $\gamma < 1$.

Proof. As shown by Lerch *et al.* [Theorem 2 and Eq. (13) in Ref. [21]], with probability at least $1-\rho$, the Pauli-propagation algorithm $\hat{C}(\theta)$ can simulate the cost $C(\theta)$ up to error $\varepsilon > 0$ using

$$\text{Time}\,\hat{C}(\theta) = \gamma_1\,|\mathcal{I}|\,D^{\log(\frac{1}{\rho\varepsilon^2})} \tag{J40}$$

many operations for some constant γ_1. On the other hand, by Theorem 4, we know that, with probability at least $1-\rho$, the Taylor surrogate $C_m(\theta)$ can simulate the cost $C(\theta)$ up to error $\varepsilon > 0$ using

$$\text{Time}\, C_m(\theta) = \gamma_2\,|\mathcal{I}|\,D^{\frac{1}{2\log(\frac{1}{q})}\log(\frac{c}{\rho\varepsilon^2})} \tag{J41}$$

many operations for some constant $\gamma_2 > 0$. Setting $\gamma := \frac{\gamma_1}{\gamma_2} > 0$, we see that $\text{Time}\, C_m(\theta) < \text{Time}\,\hat{C}(\theta)$ if and only if

$$\gamma\, D^{\log(\frac{1}{\rho\varepsilon^2})} = D^{\log(\frac{1}{\rho\varepsilon^2})+\frac{\log\gamma}{\log D}} > D^{\frac{1}{2\log(\frac{1}{q})}\log(\frac{c}{\rho\varepsilon^2})}. \tag{J42}$$

Taking the logarithm on both sides yields

$$\log\left(\frac{1}{\rho\varepsilon^2}\right) + \frac{\log\gamma}{\log D} > \frac{1}{2\log\left(\frac{1}{q}\right)}\log\left(\frac{c}{\rho\varepsilon^2}\right), \tag{J43}$$

which may equivalently be rewritten as

$$\log\left(\frac{1}{q}\right) > \frac{\log\left(\frac{c}{\rho\varepsilon^2}\right)}{2\log\left(\frac{1}{\rho\varepsilon^2}\right) + \frac{2\log\gamma}{\log D}}, \tag{J44}$$

if and only if the inequality

$$\log\left(\frac{1}{\rho\varepsilon^2}\right) + \frac{\log\gamma}{\log D} > 0 \tag{J45}$$

holds true. Equation (J45) holds true trivially if $\gamma > 1$ because of $\rho, \varepsilon < 1$. On the other hand, if $\gamma < 1$, then Eq. (J45) is equivalent to

$$\begin{aligned}\log\left(\frac{1}{\rho\varepsilon^2}\right) > \frac{\log\left(\frac{1}{\gamma}\right)}{\log D} &\iff D > \exp\left(\frac{\log\left(\frac{1}{\gamma}\right)}{\log\left(\frac{1}{\rho\varepsilon^2}\right)}\right)\\ &= \exp\left(\frac{\log\gamma}{\log\rho + 2\log\varepsilon}\right), \end{aligned} \tag{J46}$$

as assumed by the statement of the corollary.

Returning to Eq. (J44), we find that it equivalently reads

$$q < \exp\left(-\frac{\log\left(\frac{c}{\rho\varepsilon^2}\right)}{2\log\left(\frac{1}{\rho\varepsilon^2}\right) + \frac{2\log\gamma}{\log D}}\right), \tag{J47}$$

which concludes the proof. ■

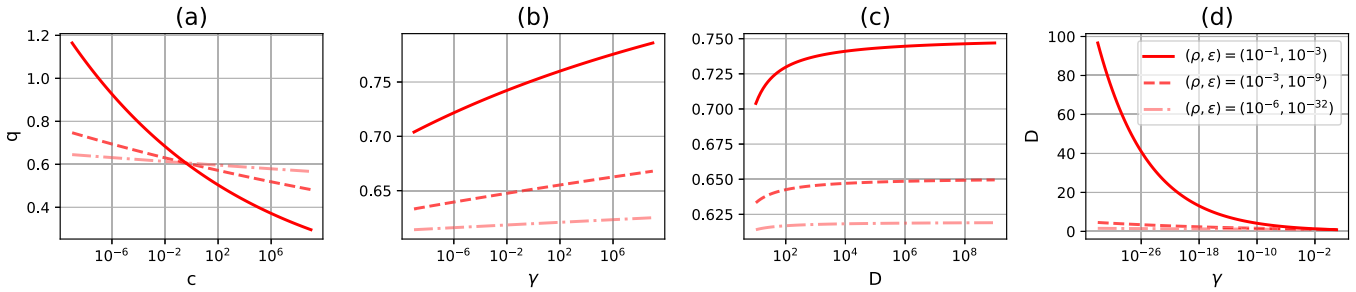


FIG. 12. Behavior of the required quantities q and D, for a polynomial speedup compared to the Pauli path simulation technique [21], as established by Corollary J1. (a)–(c) Upper bound of q. (d) Lower bound of D.

To better understand the practical implications of Corollary J1, we numerically evaluate the bound on the surrogate patch size parameter q and, if $\gamma < 1$, the number of parameters D required for outperforming the Pauli-propagation technique. Figures 12(a) and 12(c) show the upper bounds on q [cf. Eq. (J38)] across various values of the constants c, γ and parameter dimensions D. Panel (d) separately shows the behavior of the lower bound for D [cf. Eq. (J39)] when $\gamma < 1$. The plots clarify that our Taylor-based surrogate becomes advantageous in practically relevant scenarios, especially for large-scale models with high precision and low probability of failure (i.e., small $\varepsilon, \rho > 0$). For constants $c, \gamma \in [10^{-6}, 10^{6}]$ and arbitrary model sizes $D \geqslant 1$, the worst-case scenario requires $q \geqslant 0.35$, which does not limit practicality.

5. Efficient recursive classical implementation

In a practical implementation of our simulation strategy, one has to iterate over the summation $\{\|\alpha\|_1 < m\}$ in Eq. (D9). Concretely, we are interested in finding all multi-index combinations in the set:

$$\mathcal{A}_{m,D} := \left\{\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathbb{N}_0^D : \|\alpha\|_1 < m\right\}. \tag{J48}$$

When using a `for`-loop naively, one will be confronted with unnecessary computational overheads. Alternatively, we propose a recursive method that generates all combinations in the set $\mathcal{A}_{m,D}$, which indeed turns out to avoid the aforementioned overheads (cf. Fig. 9).

For that purpose, let us define $\mathbb{N}_0^0 := \{\emptyset\}$, $\|\emptyset\|_1 := 0$, and $\dim(\emptyset) := -\infty$. Consider the recursion

$$G_{m,D}(\alpha, \sigma) := \begin{cases} \bigcup_{s=0}^{m-\sigma-1} \{G_{m,D}(\alpha \cup (s), s)\}, & \dim(\alpha) < D, \\ (\alpha), & \dim(\alpha) = D, \end{cases} \tag{J49}$$

where for tuples we use $\alpha \cup (s) := (\alpha, s)$ to denote to operation of appending elements to the tuple. Also, $\dim(\alpha) \geqslant 0$ denotes the number of components of nonempty tuples α.

Lemma J3. We have that

$$G_{m,D}(\emptyset, 0) = \mathcal{A}_{m,D},$$

for all $m, D \in \mathbb{N}$.

Proof. Let $m \in \mathbb{N}$ be arbitrary. First, we are going to establish for all $D \geqslant 1$ that

$$G_{m,D}(\emptyset, 0) = \bigcup_{(s_1, \ldots, s_{D-1}) \in \mathcal{A}_{m,D-1}} \bigcup_{s_D=0}^{m-(\sum_{j=1}^{D-1} s_j)-1} \{(s_1, \ldots, s_{D-1}, s_D)\}. \tag{J50}$$

We do this by showing $G_{m,D-1}(\emptyset, 0) = \mathcal{A}_{m,D-1}$ by induction. Indeed, if $D = 1$, we find

$$G_{m,1}(\emptyset, 0) = \bigcup_{s_1=0}^{m-1} \left\{ G_{m,1}\left(\underbrace{(s_1)}_{\dim(\cdot)=1=D}, s_1 \right) \right\} = \bigcup_{s_1=0}^{m-1} \{(s_1)\}$$
$$= \{(0), (1), \ldots, (m-1)\} = \mathcal{A}_{m,1}. \tag{J51}$$

Assume by induction that $G_{m,D-1}(\emptyset, 0) = \mathcal{A}_{m,D-1}$ holds true for some fixed D. In that case,

$$G_{m,D}(\emptyset, 0) = \bigcup_{s_1=0}^{m-1} \{G_{m,D}((s_1), s_1)\} = \bigcup_{s_1=0}^{m-1} \bigcup_{s_2=0}^{m-s_1-1} \{G_{m,D}((s_1, s_2), s_1 + s_2)\}$$

$$= \cdots = \bigcup_{s_1=0}^{m-1} \bigcup_{s_2=0}^{m-s_1-1} \cdots \bigcup_{s_{D-1}=0}^{m-(\sum_{j=1}^{D-2} s_j)-1} \bigcup_{s_D=0}^{m-(\sum_{j=1}^{D-1} s_j)-1} \left\{ G_{m,D}\underbrace{(s_1, \ldots, s_{D-1}, s_D)}_{\dim(\cdot)=D}, \sum_{j=1}^{D} s_j \right\}$$

$$= \bigcup_{s_1=0}^{m-1} \bigcup_{s_2=0}^{m-s_1-1} \cdots \bigcup_{s_{D-1}=0}^{m-(\sum_{j=1}^{D-2} s_j)-1} \left(\bigcup_{s_D=0}^{m-(\sum_{j=1}^{D-1} s_j)-1} \{(s_1, \ldots, s_{D-1}, s_D)\} \right)$$

$$= \bigcup_{(s_1, \ldots, s_{D-1}) \in \underbrace{G_{m,D-1}(\emptyset, 0)}_{=\mathcal{A}_{m,D-1}}} \left(\bigcup_{s_D=0}^{m-(\sum_{j=1}^{D-1} s_j)-1} \{(s_1, \ldots, s_{D-1}, s_D)\} \right), \tag{J52}$$

where in the last equality we simply used that, by definition,

$$\bigcup_{(s_1, \ldots, s_{D-1}) \in G_{m,D-1}(\emptyset,0)} \{(s_1, \ldots, s_{D-1})\} = \underbrace{G_{m,D-1}(\emptyset, 0)}_{=\mathcal{A}_{m,D-1}} = \bigcup_{s_1=0}^{m-1} \bigcup_{s_2=0}^{m-s_1-1} \cdots \bigcup_{s_{D-1}=0}^{m-(\sum_{j=1}^{D-2} s_j)-1} \{(s_1, \ldots, s_{D-1})\}. \tag{J53}$$

Therefore, by induction, we know that Eq. (J50) holds true. Let us now use this identity to conclude with the claim $G_{m,D}(\emptyset, 0) = \mathcal{A}_{m,D}$.

"$\subseteq$": If $\alpha = (\alpha_1, \ldots, \alpha_D) \in G_{m,D}(\emptyset, 0)$, then Eq. (J50) implies that $\alpha_D \leqslant m - (\sum_{j=1}^{D-1} s_j) - 1$. This immediately results in the bound

$$\|\alpha\|_1 = \alpha_D + \sum_{j=1}^{D-1} \alpha_j \leqslant m - 1. \tag{J54}$$

In particular, $\alpha \in \mathcal{A}_{m,D}$.

"$\supseteq$": On the other hand, if $\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathcal{A}_{m,D}$, then $\|\alpha\|_1 = \alpha_D + \sum_{j=1}^{D-1} < m$ yields

$$\alpha_D \leqslant m - \left(\sum_{j=1}^{D-1} \alpha_j \right) - 1. \tag{J55}$$

In particular,

$$\alpha = (\alpha_1, \ldots, \alpha_{D-1}, \alpha_D) \in \bigcup_{s_D=0}^{m-(\sum_{j=1}^{D-1} \alpha_j)-1} \times \{(\alpha_1, \ldots, \alpha_{D-1}, s_D)\}. \tag{J56}$$

Additionally, $\|\alpha\|_1 < m$ means that one necessarily must have $(\alpha_1, \ldots, \alpha_{D-1}) \in \mathcal{A}_{m,D-1}$. Indeed, otherwise we have the contradiction

$$m \leqslant \sum_{j=1}^{D-1} \alpha_j \leqslant \sum_{j=1}^{D} \alpha_j = \|\alpha\|_1 < m. \tag{J57}$$

Therefore,

$$\alpha \in \bigcup_{(s_1, \ldots, s_{D-1}) \in \mathcal{A}_{m,D-1}} \bigcup_{s_D=0}^{m-(\sum_{j=1}^{D-1} \alpha_j)-1} \{(s_1, \ldots, s_{D-1}, s_D)\} = G_{m,D}(\emptyset, 0), \tag{J58}$$

where in the last equality we made use of our previous claim (J50). ■

APPENDIX K: ADDITIONAL NUMERICAL EXPERIMENTS

This Appendix expands upon the numerical experiments in Sec. IV. In particular, we show the scaling of initial gradients for all circuit designs considered in this work, VQE experiments with finite sampling and different learning rates, and ultimately VQE experiments for the Heisenberg model.

1. Gradient scaling experiments

We repeat the initial gradient scaling experiments (cf. Figs. 4 and 5) comparing all models considered in this work: `mHEA`, `fHEA`, and `rPQC` (defined in Appendix C).

Figure 13 illustrates the expected behavior of the initial gradient norm at $\theta = 0$ employing various PQC models with $L = 1$ layer. Gradients are computed analytically via Eq. (14). Results are averaged over ten independent runs, each involving a randomly selected single Pauli string $H = P$ and a random direction $k \in [D]$ used in constructing the LCE transformation. As shown, the initial gradient norm for the non-LCE model decays exponentially and becomes indistinguishable from zero beyond $N = 10$ qubits. In contrast, the LCE model consistently maintains an initial gradient norm of at least 1, in agreement with the theoretical prediction of Theorem 1.

Figure 14 compares initial gradient norms at $\theta = 0$ with and without LCE, employing `mHEA` with $L = 1$ layer. Gradients are computed analytically via Eq. (14). Each run uses a randomly sampled observable H, and a random LCE construction. Panel (a) of Fig. 4 focuses on uniformly random single Pauli strings $H = P$. Without LCE, constructive contributions in Eq. (14) vanish with increasing N, leading to zero expected gradient for sufficiently large N. With LCE, gradient norms remain constant in N. Panels (b) and (c) of Fig. 4 extend results to random observables with $|\mathcal{I}| = \lfloor \sqrt{N} \rfloor$ Pauli terms. For observables with uniformly sampled Pauli strings P_i and $c_i = 1$ for all $i \in \mathcal{I}$ (called random unweighted H), the scaling compares to the single Pauli case. For observables with uniformly sampled Pauli strings P_i and i.i.d. $c_i \sim \mathrm{Unif}([-1, 1])$ for all $i \in \mathcal{I}$ (called random weighted H), gradient norms tend to behave more randomly due to random Pauli coefficients.

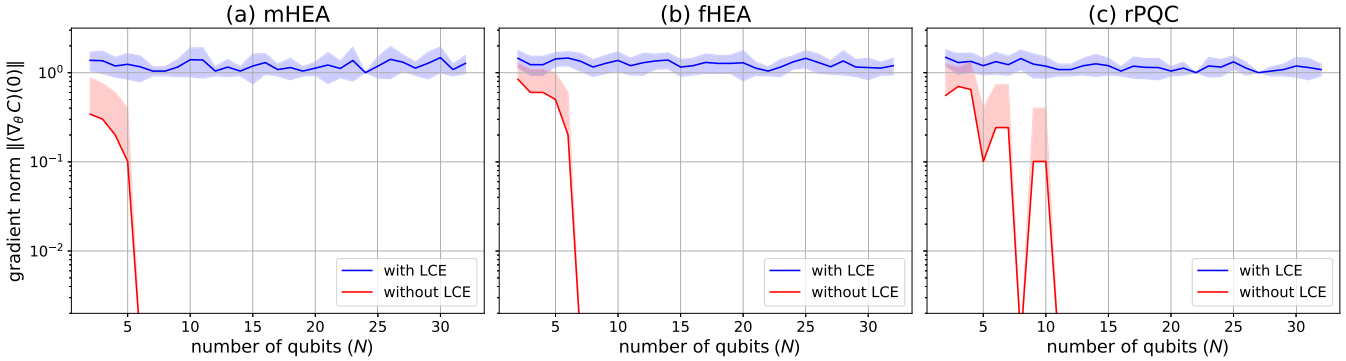


FIG. 13. Scaling behavior of the initial gradient norm at $\theta = 0$ for across various PQCs with $L = 1$ layer. We average over 10 samples of random single Pauli strings $H = P$, and random directions $k \in [D]$ for the LCE transformation.

This is evidence that the LCE transformation remains effective for arbitrary observables.

Figure 15 repeats the experiment of Fig. 13 on near-Clifford patches, comparing various linear-depth PQC models, and varying patch sizes via Gaussian initialization $\sigma = D^{-r}$. Each run uses a randomly sampled parameter, a random single Pauli string observable $H = P$, and a random LCE construction. The red-shaded region corresponds to classically efficient regimes (cf. Fig. 3); blue-shaded region to super-polynomial regimes, where gradients may still decay only polynomially. Within the constant-scaling regime (red-shaded region), LCE ensures gradient scalability. For large σ, gradients decay exponentially. This suggests the existence of a larger patch size where gradients transition from constant decay to polynomial decay [cf. Eq. (9)].

The same experiment is performed using shallow PQCs with a single layer ($L = 1$) as shown in Fig. 16. Comparing the results of linear-depth and single-layer setups reveals the impact on gradient scalability: As the number of parameters increases (i.e., in deeper circuits), the gradient norm tends to diminish more rapidly, especially for larger patch sizes.

2. VQE optimization with different learning rates and finite sampling noise

Here, we demonstrate that LCE is effective for smaller learning rates in vanilla gradient descent. Furthermore, we address the issue of finite sampling noise [60]. We use a small number of finite shots to evaluate the gradients in order to emphasize that LCE remains relevant for near-term practical applications.

Figure 17 shows VQE optimization of the global observable $Z^{\otimes N}$, employing the `mHEA` ansatz ($L = 5$, $N = 15$), averaging over five runs with Gaussian initializations $\theta \sim \mathcal{N}(0, \sigma^2 I_D)$ across varying patch sizes. Optimization is performed with vanilla gradient descent, where we compare various constant learning rates. The gradients are evaluated exactly as per Eq. (14). LCE consistently yields faster convergence, independent of the learning rate. Fastest convergence is achieved with a learning rate of 0.1, and continuously slows down as the learning rate increases. While our theory guarantees trainability up to $\sigma = D^{-1/2}$, numerical results continue to demonstrate successful convergence slightly beyond, even in classically intractable regimes (cf. Theorem 3). LCE advantage is less pronounced as σ increases (cf. Fig. 5).

Figure 18 repeats the same VQE experiment, but this time using finite shots sampling, and a learning rate of 0.01. The improved gradient due to LCE allows fewer sampling shots. In this case, eight shots are already sufficient for fast convergence up to the patch size $\sigma = D^{-1/3}$. Finite shots sampling explaining the increase of noise in the first three rows. Interestingly, using finite shots sampling appears to converge slightly faster compared to the exact dynamics in the last row. Hence, it may be possible that finite sampling noise may be exploited to improve the dynamics of LCE-transformed VQEs. Inspecting the column containing $\sigma = D^{-1/5}$, one no-

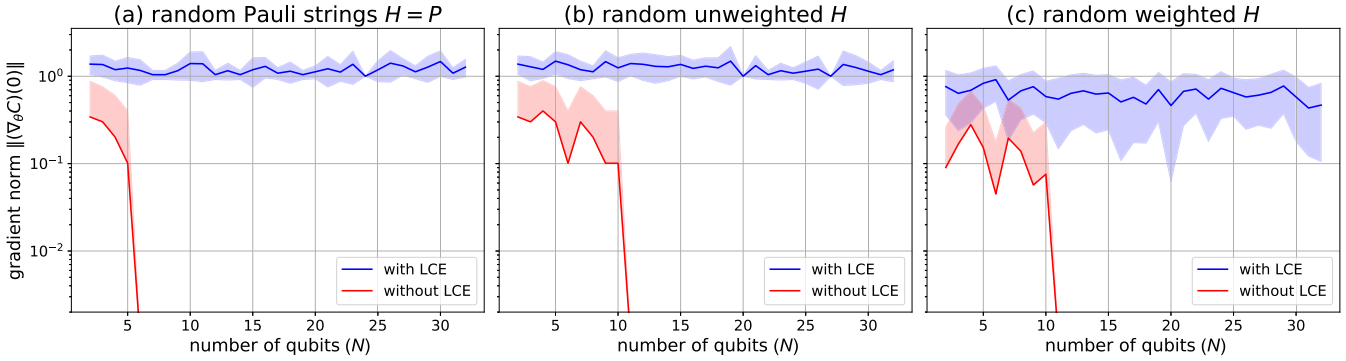


FIG. 14. Scaling behavior of the initial gradient norm at $\theta = 0$ for the `mHEA` model with $L = 1$ layer. The panels vary across types of random observables. The experiment highlights the effect of the LCE transformation. We average over ten independent runs.

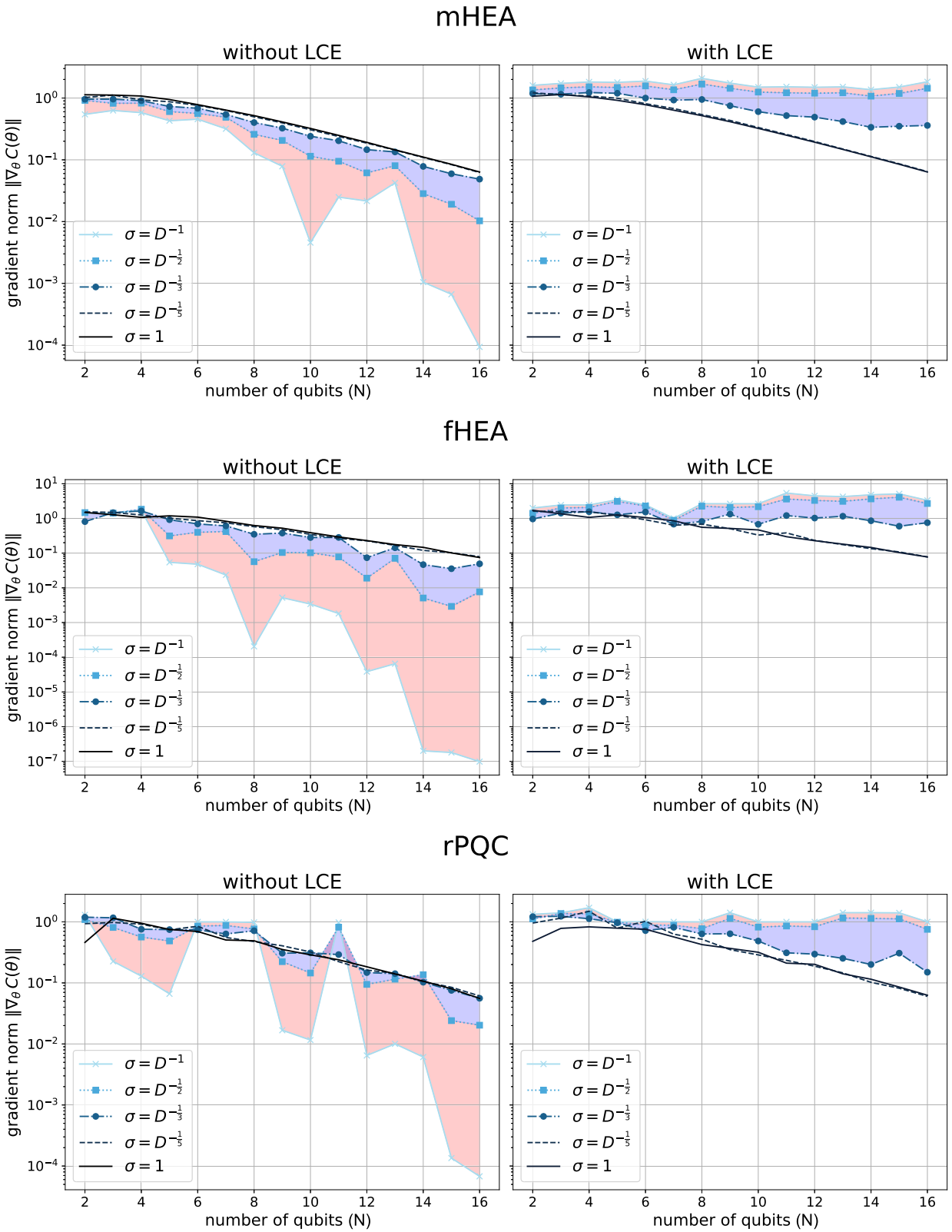


FIG. 15. Scaling behavior of the gradient norm on various surrogate patches for PQC models with $L = N$ layers. We average over 50 independent runs.

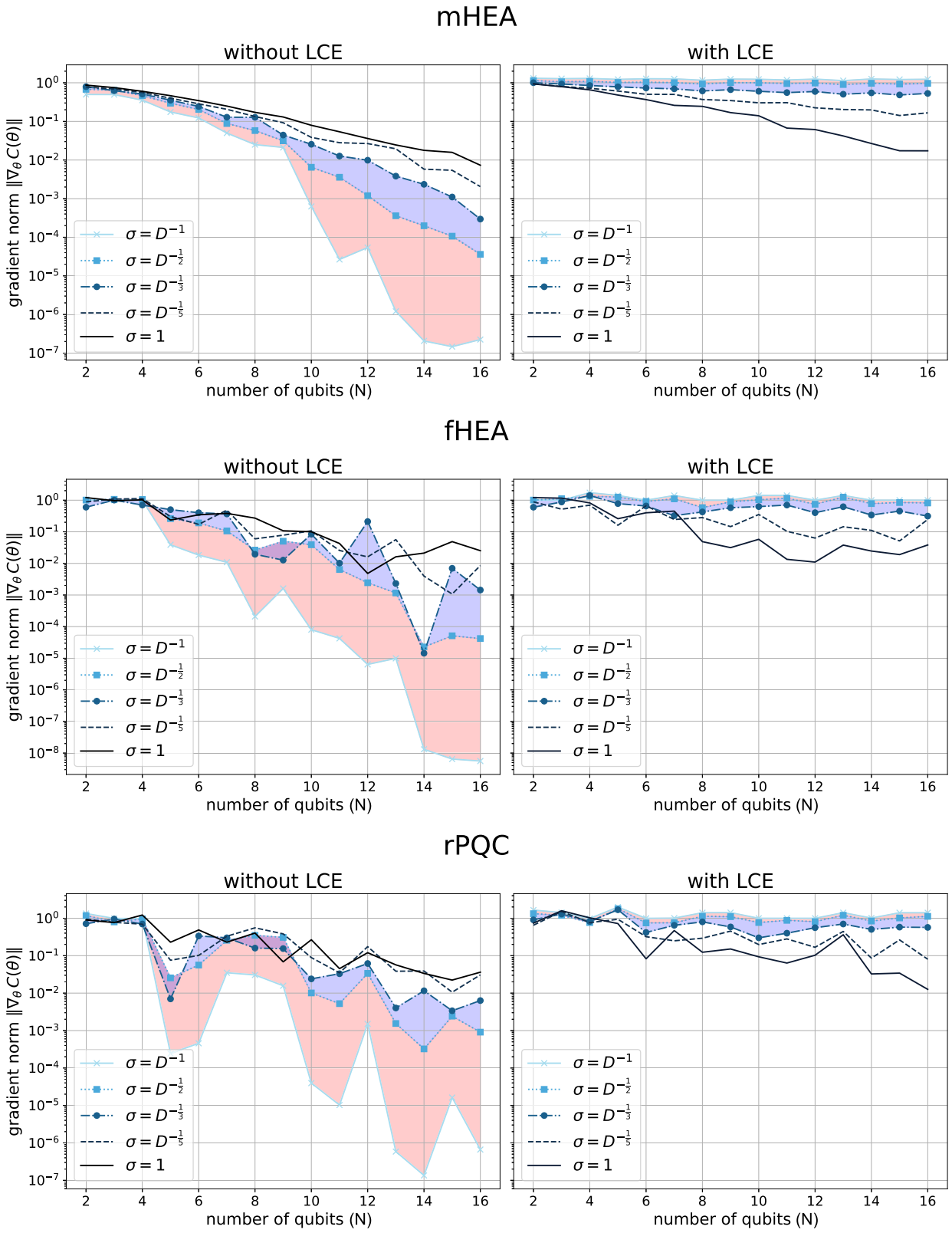


FIG. 16. Scaling behavior of the gradient norm on various surrogate patches for PQC models with $L = 1$ layer. We average over 50 independent runs.

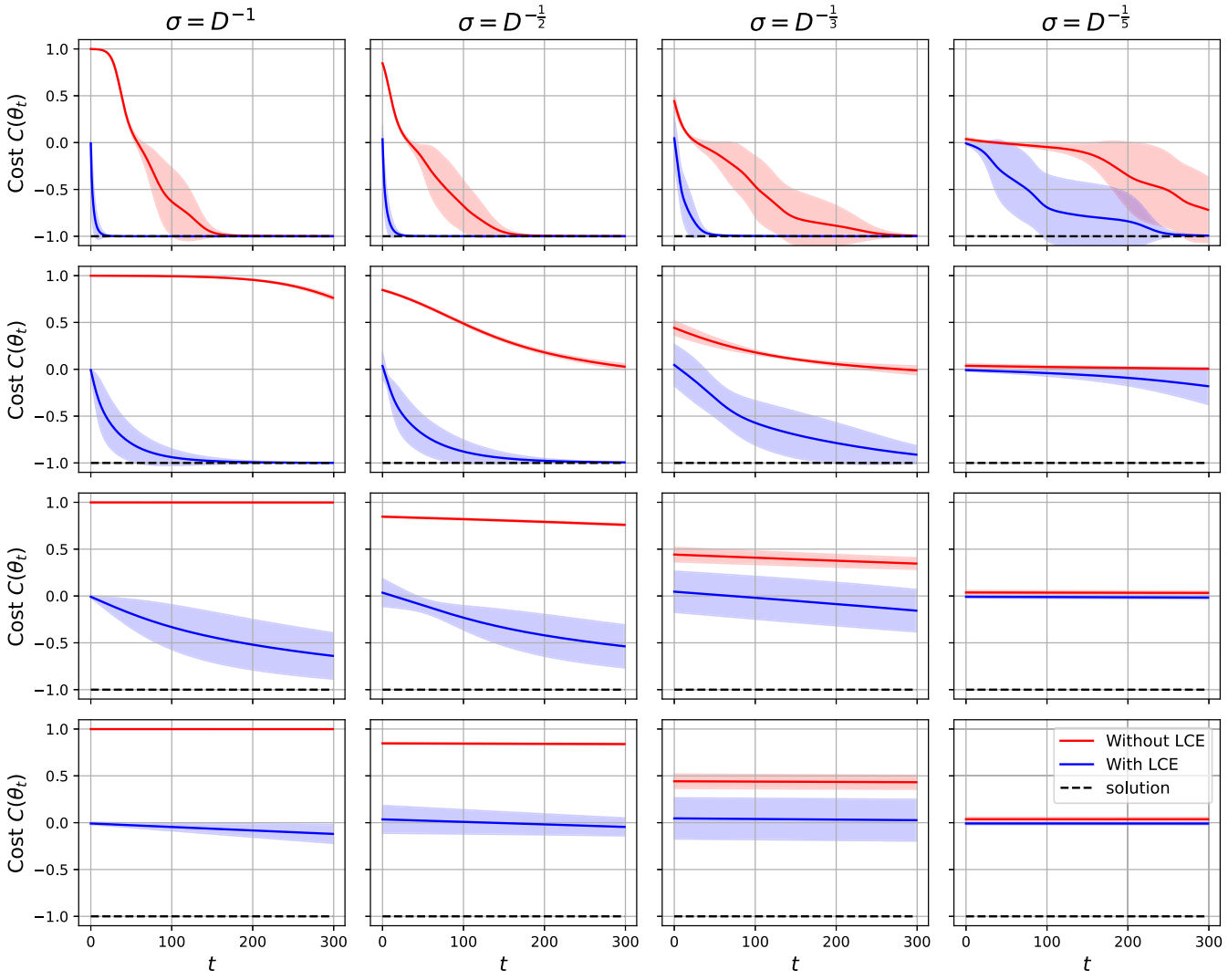


FIG. 17. VQE Optimization using vanilla gradient descent with various learning rates of 0.1 (top row), 0.01 (second row), 0.001 (third row), and 0.0001 (bottom row). We evaluate exact gradients, and average over five runs of random Gaussian initialization across varying patch sizes. We find the ground energy of the global observable $Z^{\otimes N}$ with $N = 15$, employing the `mHEA` ansatz with $L = 5$ layers.

tices that a 15-qubit model suffering from barren plateaus generally requires 2^{15} shots to accurately estimate the gradient in order to eventually achieve convergence, whereas employing exact gradients [cf. Eq. (14)] barely enables slow convergence. These findings can be seen as evidence that LCE has the potential to enhance finite sampling efficiency in VQEs.

3. VQE optimization of the Heisenberg model

In this experiment, we leverage the benchmarked initial gradient scaling statistics to optimize the VQE of the N-qubit Heisenberg model [61],

$$H_{\text{Heisenberg}} := \sum_{j=1}^{N-1} \{X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}\}. \tag{K1}$$

Equation (K1) defines a local observable and thus is less prone to the barren plateau problem [25]. In order to increase the chance of inducing barren plateaus, one may consider deep circuit ansätze, which result in expressivity-induced barren plateaus [30]. Therefore, when employing shallow circuits to find the ground state of a local observable, the effectiveness of LCE is expected to disappear.

We compare the optimization dynamics of the 12-qubit Heisenberg model in Eq. (K1) with and without the LCE transformation. We employ `mHEA` with $L = 36$ layers, and average over five independent runs of random Gaussian initialization of parameters across varying patch sizes. The results are presented in Fig. 19. We can see the advantage of LCE on sufficiently large patch sizes stemming from the improved initial gradient as per Fig. 5, while the non-LCE model remains stuck on a local minimum. If the patch size is too large (e.g., $\sigma = D^{-1/5}$), then the advantage of LCE is less pronounced.

Recalling the VQE experiment with $H_{\text{global}} = Z^{\otimes N}$ (cf. Fig. 6), we note a significant discrepancy between the converge rates for $Z^{\otimes N}$ and the Heisenberg model. This is likely due to the locality of the Heisenberg observable, which

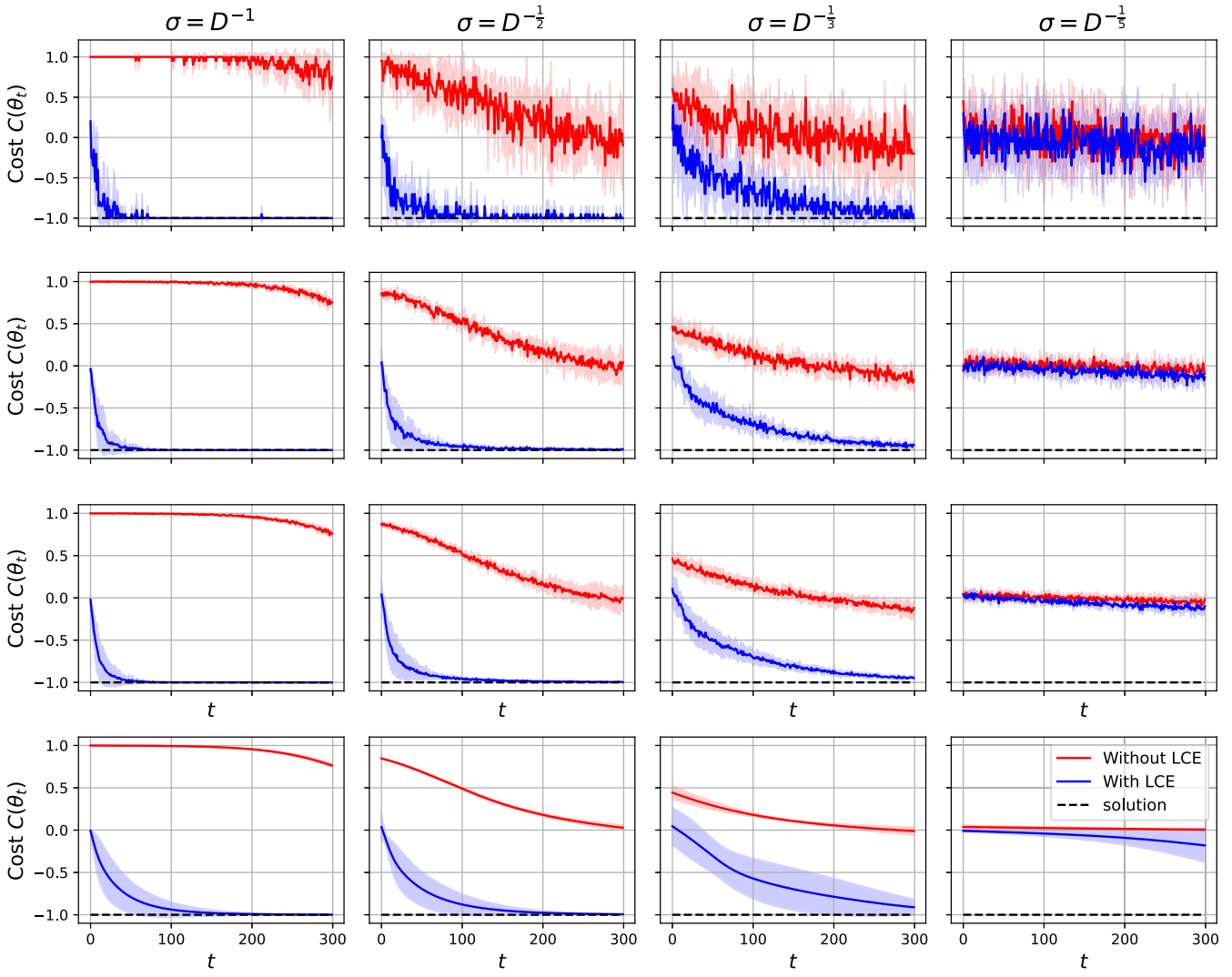


FIG. 18. VQE Optimization using vanilla gradient descent with a learning rate of 0.01, estimating gradients with 8 shots (top row), 64 shots (second row), 512 shots (third row), and exact gradients (bottom row). We average over five runs of random Gaussian initialization across varying patch sizes. We find the ground energy of the global observable $Z^{\otimes N}$ with $N = 15$, employing the `mHEA` ansatz with $L = 5$ layers.

is known to be barren plateaus free for shallow circuits [25]. In fact, by reducing the number of layers from $L = 36$ to $L = 6$ as shown in Fig. 20, we can see that the non-LCE dynamics indeed tends to gets closer to the solution, while the LCE dynamics appears to get trapped in a similar energy level as in the $L = 36$ experiment. These results indicate that the beneficial impact of LCE (being particularly effective for mitigating barren plateaus) is less pronounced compared to the deeper `mHEA` model with $L = 36$ layers. This means that LCE should only be applied for VQEs with observables subject to barren plateaus, such as $H_{\text{global}} = Z^{\otimes N}$ [25,37]. Interestingly, this experiment also suggests that the non-LCE initialization might, in some cases, begin closer to a global minimum, whereas the LCE transformation could potentially lead to entrapment in a local minimum (particularly in underparametrized models). We hypothesize that such behavior is less prevalent in larger-scale models and that alternative optimization strategies, such as the adaptive moment estimation (ADAM) optimizer [77], may offer a way to escape local minima.

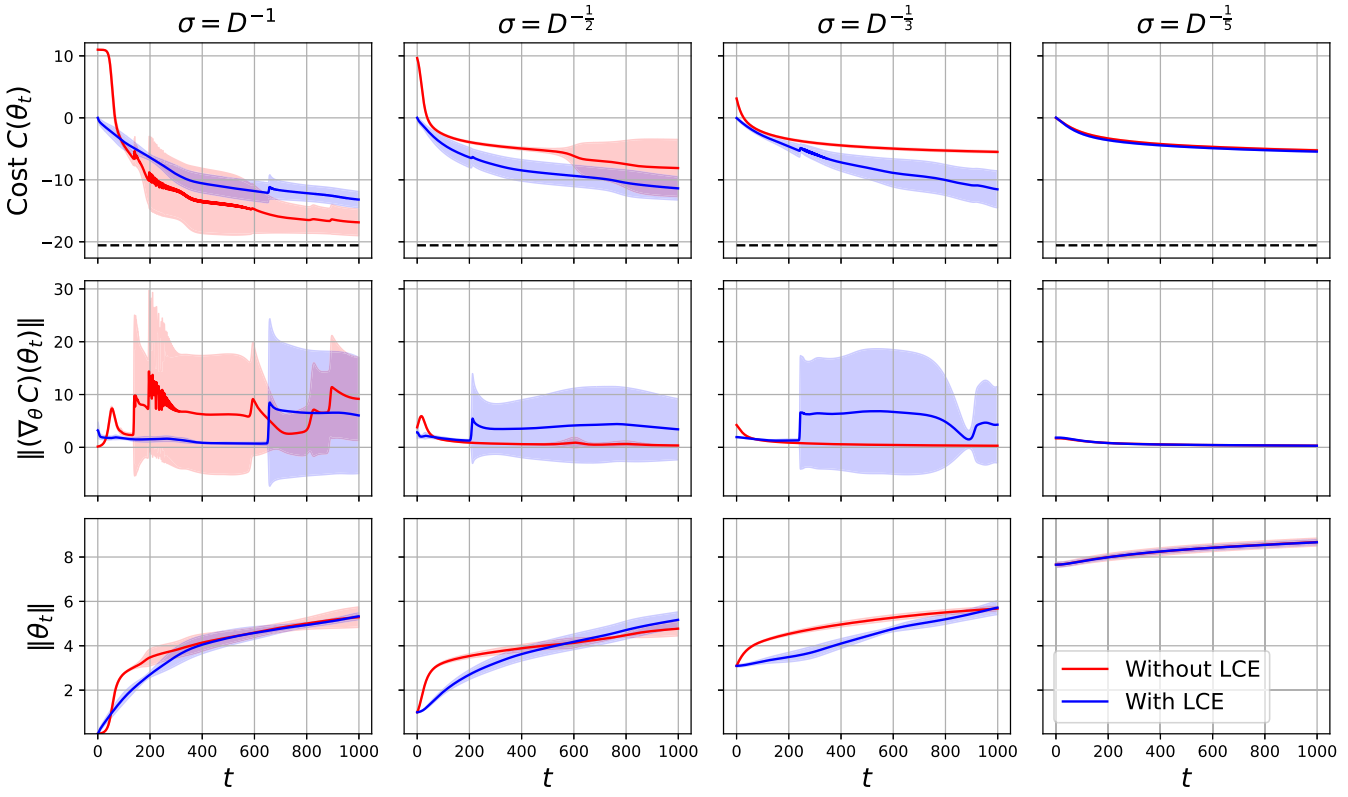


FIG. 19. VQE Optimization using vanilla gradient descent with a learning rate of 0.01, averaging over five runs of random Gaussian initialization across varying patch sizes. We employ the `mHEA` ansatz with $L = 36$ layers to find the ground energy of 12-qubit Heisenberg model.

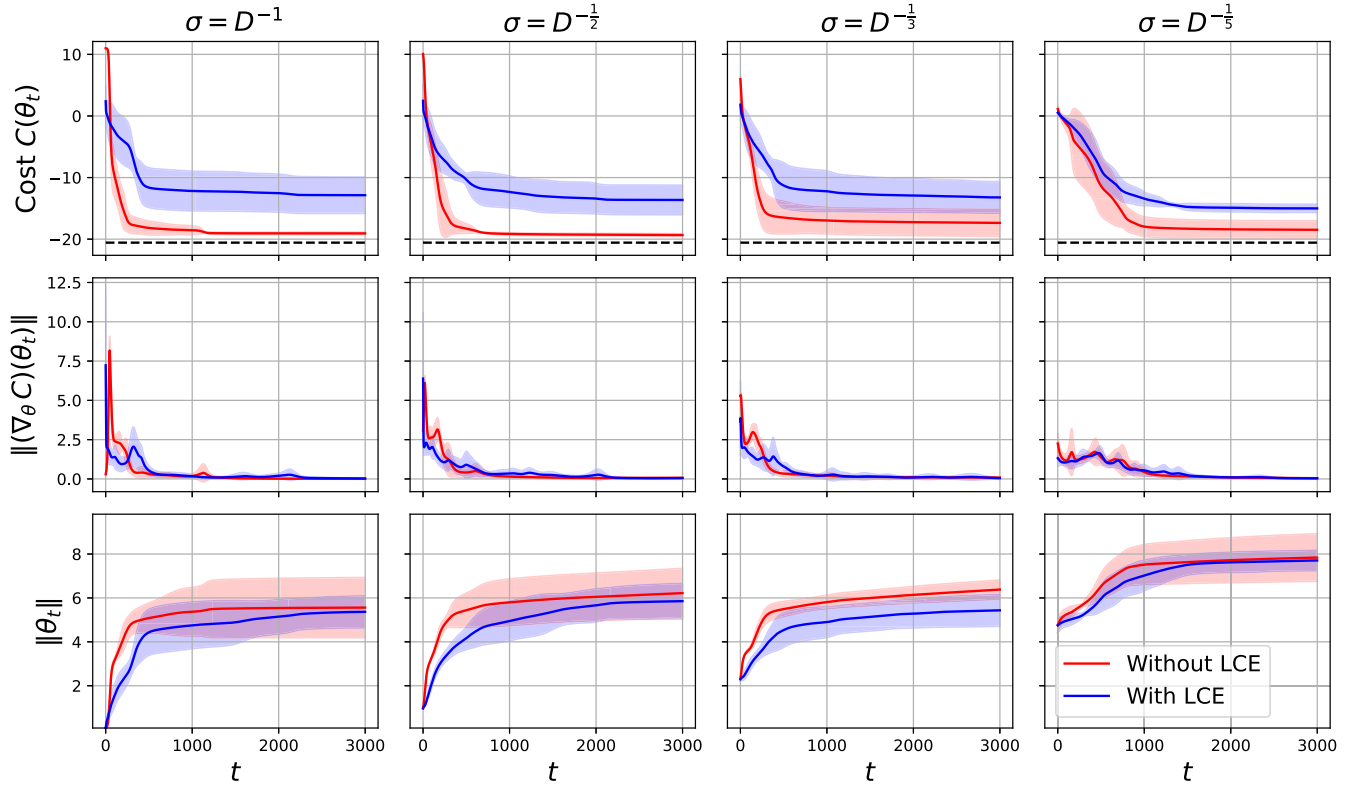


FIG. 20. VQE Optimization using vanilla gradient descent with a learning rate of 0.01, averaging over five runs of random Gaussian initialization across varying patch sizes. We employ the `mHEA` ansatz with $L = 6$ layers to find the ground energy of 12-qubit Heisenberg model.

[1] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[2] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[3] V. V. Kuzmin and P. Silvi, Variational quantum state preparation via quantum data buses, Quantum **4**, 290 (2020).

[4] G. Matos, S. Johri, and Z. Papić, Quantifying the efficiency of state preparation via quantum variational eigensolvers, PRX Quantum **2**, 010309 (2021).

[5] G. Gyawali and M. Lawler, Adaptive variational preparation of the Fermi-Hubbard eigenstates, Phys. Rev. A **105**, 012413 (2022).

[6] R. C. Farrell, M. Illa, A. N. Ciavarella, and M. J. Savage, Scalable circuits for preparing ground states on digital quantum computers: The Schwinger model vacuum on 100 qubits, PRX Quantum **5**, 020315 (2024).

[7] J. S. Kottmann and F. Scala, Quantum algorithmic approach to multiconfigurational valence bond theory: Insights from interpretable circuit design, J. Chem. Theory Comput. **20**, 3514 (2024).

[8] P. Wittek, *Quantum Machine Learning* (Academic Press, Boston, 2014).

[9] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[10] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, Challenges and opportunities in quantum machine learning, Nat. Comput. Sci. **2**, 567 (2022).

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

[12] M. Larocca, S. Thanasilp, S. Wang, K. Sharma, J. Biamonte, P. J. Coles, L. Cincio, J. R. McClean, Z. Holmes, and M. Cerezo, Barren plateaus in variational quantum computing, Nat. Rev. Phys. **7**, 174 (2025).

[13] M. Cerezo, M. Larocca, D. García-Martín, N. L. Diaz, P. Braccia, E. Fontana, M. S. Rudolph, P. Bermejo, A. Ijaz, S. Thanasilp, *et al.*, Does provable absence of barren plateaus imply classical simulability? Nat. Commun. **16**, 7907 (2023).

[14] H.-K. Zhang, S. Liu, and S.-X. Zhang, Absence of barren plateaus in finite local-depth circuits with long-range entanglement, Phys. Rev. Lett. **132**, 150603 (2024).

[15] A. Deshpande, M. Hinsche, K. Najafi, K. Sharma, R. Sweke, and C. Zoufal, Dynamic parameterized quantum circuits: Expressive and barren-plateau free, arXiv:2411.05760.

[16] C.-Y. Park, M. Kang, and J. Huh, Hardware-efficient ansatz without barren plateaus in any depth, arXiv:2403.04844.

[17] K. Mitarai, Y. Suzuki, W. Mizukami, Y. O Nakagawa, and K. Fujii, Quadratic Clifford expansion for efficient benchmarking and initialization of variational quantum algorithms, Phys. Rev. Res. **4**, 033012 (2022).

[18] D. Gottesman, The Heisenberg representation of quantum computers, in *Group22: Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics*, edited by S. P. Corney, R. Delbourgo, and P. D. Jarvis (International Press, Cambridge, MA, 1999), pp. 32–43.

[19] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).

[20] F. J. Schreiber, J. Eisert, and J. J. Meyer, Classical surrogates for quantum learning models, Phys. Rev. Lett. **131**, 100803 (2023).

[21] S. Lerch, R. Puig, M. S. Rudolph, A. Angrisani, T. Jones, M. Cerezo, S. Thanasilp, and Z. Holmes, Efficient quantum-enhanced classical simulation for patches of quantum landscapes, arXiv:2411.19896.

[22] S. Bravyi and D. Gosset, Improved classical simulation of quantum circuits dominated by Clifford gates, Phys. Rev. Lett. **116**, 250501 (2016).

[23] S. Bravyi, D. Browne, P. Calpin, E. Campbell, D. Gosset, and M. Howard, Simulation of quantum circuits by low-rank stabilizer decompositions, Quantum **3**, 181 (2019).

[24] T. Begušić, K. Hejazi, and G. K. Chan, Simulating quantum circuit expectation values by Clifford perturbation theory, J. Chem. Phys. **162**, 15 (2025).

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

[26] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, New York, 2010).

[27] K. Mastel, The Clifford theory of the n-qubit Clifford group, J. Math. Phys. **67**, 021701 (2026).

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

[29] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, Adv. Quantum Technol. **2**, 1900070 (2019).

[30] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

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

[32] K. Zhang, L. Liu, M.-H. Hsieh, and D. Tao, Escaping from the barren plateau via Gaussian initializations in deep variational quantum circuits, Adv. Neural Inf. Process. Syst. **35**, 18612 (2022).

[33] Y. Wang, B. Qi, C. Ferrie, and D. Dong, Trainability enhancement of parameterized quantum circuits via reduced-domain parameter initialization, Phys. Rev. Appl. **22**, 054005 (2024).

[34] M. L. Goh, M. Larocca, L. Cincio, M. Cerezo, and F. Sauvage, Lie-algebraic classical simulations for variational quantum computing, Phys. Rev. Res. **7**, 033266 (2025).

[35] M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, Quantum Sci. Technol. **6**, 035006 (2021).

[36] K. Zhang, M.-H. Hsieh, L. Liu, and D. Tao, Toward trainability of deep quantum neural networks, arXiv:2112.15002.

[37] A. Letcher, S. Woerner, and C. Zoufal, Tight and efficient gradient bounds for parameterized quantum circuits, Quantum, **8**, 1484 (2024).

[38] C.-Y. Park and N. Killoran, Hamiltonian variational ansatz without barren plateaus, Quantum **8**, 1239 (2024).

[39] X. Shi and Y. Shang, Avoiding barren plateaus via Gaussian mixture model, New J. Phys. **27**, 104501 (2025).

[40] A. Skolik, J. R. McClean, M. Mohseni, P. Van Der Smagt, and M. Leib, Layerwise learning for quantum neural networks, Quantum Mach. Intell. **3**, 5 (2021).
[41] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, Phys. Rev. Res. **3**, 033090 (2021).
[42] T. Volkoff and P. J. Coles, Large gradients via correlation in random parameterized quantum circuits, Quantum Sci. Technol. **6**, 025008 (2021).
[43] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, Phys. Rev. X **11**, 041011 (2021).
[44] H. Mhiri, R. Puig, S. Lerch, M. S. Rudolph, T. Chotibut, S. Thanasilp, and Z. Holmes, A unifying account of warm start guarantees for patches of quantum landscapes, arXiv:2502.07889.
[45] A. Angrisani, A. Schmidhuber, M. S. Rudolph, M. Cerezo, Z. Holmes, and H.-Y. Huang, Classically estimating observables of noiseless quantum circuits, Phys. Rev. Lett. **135**, 170602 (2025).
[46] M. S. Rudolph, T. Jones, Y. Teng, A. Angrisani, and Z. Holmes, Pauli propagation: A computational framework for simulating quantum systems, arXiv:2505.21606.
[47] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).
[48] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).
[49] G. Vidal, Efficient classical simulation of slightly entangled quantum computations, Phys. Rev. Lett. **91**, 147902 (2003).
[50] J. Eisert, M. Cramer, and M. B. Plenio, Colloquium: Area laws for the entanglement entropy, Rev. Mod. Phys. **82**, 277 (2010).
[51] J. Eisert, Entanglement and tensor network states, arXiv:1308.3318.
[52] B. Fuller, M. C. Tran, D. Lykov, C. Johnson, M. Rossmannek, K. X. Wei, A. He, Y. Kim, D. Vu, K. Sharma, *et al.*, Improved quantum computation using operator backpropagation, npj Quantum Inf. **12**, 51 (2026).
[53] R. Puig, M. Drudis, S. Thanasilp, and Z. Holmes, Variational quantum simulation: A case study for understanding warm starts, PRX Quantum **6**, 010317 (2025).
[54] J. Kempe, A. Kitaev, and O. Regev, The complexity of the local Hamiltonian problem, SIAM J. Comput. **35**, 1070 (2006).
[55] A. B. Magann, C. Arenz, M. D. Grace, T.-S. Ho, R. L. Kosut, J. R. McClean, H. A. Rabitz, and M. Sarovar, From pulses to circuits and back again: A quantum optimal control perspective on variational quantum algorithms, PRX Quantum **2**, 010101 (2021).
[56] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, Theory of overparametrization in quantum neural networks, Nat. Comput. Sci. **3**, 542 (2023).
[57] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, Quantum optimally controlled transition landscapes, Science **303**, 1998 (2004).
[58] L. Leone, S. F. E. Oliviero, L. Cincio, and M. Cerezo, On the practical usefulness of the hardware efficient ansatz, Quantum **8**, 1395 (2024).
[59] EfficientSU2, IBM Quantum Platform, 2023, https://docs.quantum.ibm.com/api/qiskit/qiskit.circuit.library.EfficientSU2.
[60] D. A. Kreplin and M. Roth, Reduction of finite sampling noise in quantum neural networks, Quantum **8**, 1385 (2024).
[61] F. Bonechi, E. Celeghini, R. Giachetti, E. Sorace, and M. Tarlini, Heisenberg XXZ model and quantum Galilei group, J. Phys. A: Math. Gen. **25**, L939 (1992).
[62] J. Liu, F. Tacchino, J. R. Glick, L. Jiang, and A. Mezzacapo, Representation learning via quantum neural tangent kernels, PRX Quantum **3**, 030323 (2022).
[63] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, Analytic theory for the dynamics of wide quantum neural networks, Phys. Rev. Lett. **130**, 150601 (2023).
[64] E. Abedi, S. Beigi, and L. Taghavi, Quantum lazy training Quantum **7**, 989 (2023).
[65] X. You, S. Chakrabarti, B. Chen, and X. Wu, Analyzing convergence in quantum neural networks: Deviations from neural tangent kernels, in *International Conference on Machine Learning* (PMLR, Honolulu, Hawaii, 2023), pp. 40199–40224.
[66] F. Scala, C. Zoufal, D. Gerace, and F. Tacchino, Towards practical quantum neural network diagnostics with neural tangent kernels, npj Quantum Inf. (2026).
[67] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, Nat. Comput. Sci. **1**, 403 (2021).
[68] F. Tiblias, A. Schroeder, Y. Zhang, M. Gachechiladze, and I. Gurevych, An efficient quantum classifier based on Hamiltonian representations, in *International Conference on Quantum Artifical Intelligence and Natural Language Processing* (Springer Nature, Cham, Switzerland, 2025), pp. 67–89.
[69] R. M. Corless, G. H Gonnet, D. E. G. Hare, D. J Jeffrey, and D. E Knuth, On the Lambert *W* function, Adv. Comput. Math. **5**, 329 (1996).
[70] H. Robbins, A remark on Stirling's formula, Am. Math. Mon. **62**, 26 (1955).
[71] W. Feller, *An Introduction to Probability Theory and Its Applications* (John Wiley & Sons, New York, 1991), Vol. 2.
[72] C. H. Jones, Generalized hockey stick identities and *N*-dimensional blockwalking, Fibonacci Q. **34**, 280 (1996).
[73] D. Aharonov and T. Naveh, Quantum NP—A survey, arXiv:quant-ph/0210077.
[74] D. Aharonov, I. Arad, and T. Vidick, Guest column: The quantum PCP conjecture, ACM SIGACT News **44**, 47 (2013).
[75] A. Natarajan and T. Vidick, Low-degree testing for quantum states, and a quantum entangled games PCP for QMA, in *Proceedings of the 2018 EEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2018), pp. 731–742.
[76] I. Dinur, The PCP theorem by gap amplification, J. ACM **54**, 12 (2007).
[77] D. P. Kingma, and J. Ba, ADAM: A method for stochastic optimization, arXiv:1412.6980.
[78] S. Meyer, sabrimeyer/Taylor-Surrogate-and-LCE, version 1.0, Github, 2026, https://github.com/sabrimeyer/Taylor-Surrogate-and-LCE.