

Efficient Quantum-Enhanced Classical Simulation for Patches of Quantum Landscapes

Sacha Lerch,^{1,2,*} Ricard Puig^{1,2,*}, Manuel S. Rudolph^{1,2,*}, Armando Angrisani^{1,2},
Tyson Jones^{1,2}, M. Cerezo,^{3,4} Supanut Thanasilp^{1,2,5,6,†} and Zoë Holmes^{1,2}

¹*Institute of Physics, [Ecole Polytechnique Fédérale de Lausanne \(EPFL\)](#), CH-1015 Lausanne, Switzerland*

²*Centre for Quantum Science and Engineering, [Ecole Polytechnique Fédérale de Lausanne \(EPFL\)](#), Lausanne, Switzerland*

³*[Los Alamos National Laboratory](#), Los Alamos, New Mexico 87545, USA*

⁴*Quantum Science Center, [Oak Ridge](#), Tennessee 37931, USA*

⁵*Chula Intelligent and Complex Systems Center of Excellence, Department of Physics, Faculty of Science, [Chulalongkorn University](#), Bangkok 10330, Thailand*

⁶*Siam Quantum Square, Faculty of Science, [Chulalongkorn University](#), Bangkok, Thailand*



(Received 24 December 2024; revised 22 December 2025; accepted 13 April 2026; published 12 June 2026)

Understanding the capabilities of classical simulation methods is key to identifying where quantum computers are advantageous. Not only does this ensure that quantum computers are used only where necessary, but also one can potentially identify subroutines that can be offloaded onto a classical device. In this work, we show that it is always possible to generate a *classical surrogate* of a subregion (dubbed a “patch”) of an expectation landscape produced by a parameterized quantum circuit. That is, we provide a quantum-enhanced classical algorithm which, after simple measurements on a quantum device, allows one to classically simulate approximate expectation values of a subregion of a landscape. We provide time and sample complexity guarantees for a range of families of circuits of interest, and further numerically demonstrate our simulation algorithms on an exactly verifiable simulation of a Hamiltonian variational *Ansatz* and long-time dynamics simulation on a 127-qubit heavy-hex topology.

DOI: [10.1103/fhc5-8sm6](https://doi.org/10.1103/fhc5-8sm6)

I. INTRODUCTION

In a well-spirited effort to utilize quantum devices to the best of their abilities, the common approach could be summarized as: *Do as little as possible on the quantum computer by offloading work onto classical machines*. However, this does not relieve quantum computers from the necessary and sufficient conditions to run quantum circuits that are both (i) classically hard and (ii) practically useful. With classical algorithms still outperforming current quantum computers for dynamical simulation [1–7], recent developments in variational quantum computing [8] and no clear trajectory for making use of near-term quantum devices, we are left to wonder: *should we offload more work to classical computers?*

In this work, we provide theoretical guarantees and practical implementations of fully classical and quantum-enhanced classical simulations of *quantum landscape patches*. That is, for simulating a small region in the full expectation landscape spanned by the gate angles of a quantum circuit. We consider a *spectrum of classicality*, where quantum computers are used either (i) not at all, (ii) to provide measurements of the initial states, or (iii) to estimate the most important contributions to a quantum landscape. In the former two cases, our results can be viewed as providing guarantees for the simulation of low-magic or near-Clifford circuits [9]. In the latter two cases, the quantum computer is effectively tasked with collecting data to build a so-called *classical surrogate* [2,10–14] of the patch of a quantum landscape. As such, our work establishes a hybrid quantum-classical simulation framework that attempts to realize the original motivation of utilizing quantum resources only for the smallest possible but crucial component.

Our results are applicable to variational quantum computing where the landscape corresponds to a loss function which one seeks to minimize to solve a given task [15,16]. In this context, prior studies of classical surrogates

*The first three authors contributed equally to this work.

†Contact author: supanut.thanasilp@gmail.com

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.

include those for; shallow hardware-efficient *Ansätze* [17–20], circuits with polynomial dynamical Lie Algebra [21–25], unstructured parameterized quantum circuits [2,26], quantum convolutional neural networks [26,27], quantum kernel-based circuits [28], noisy circuits [14,29], and quantum machine learning [10–12,30]. However, those proposals are to surrogate the entire landscape for relatively restricted families of circuits. Surrogating more general circuits in restricted regions, except for some largely heuristic works [31,32], has so far not been studied. In this paper, we show that it is possible to create a classical surrogate for previously proposed reduced angle range initialization strategies for avoiding barren plateaus [33–42].

We emphasize that although these surrogates superficially have the flavor of dequantizations, they do not in general imply the irrelevance of quantum computers. In the generic setting, one still requires a quantum device to generate the data from which the surrogate is constructed, except in those cases where the relevant data can themselves be obtained by efficient classical simulation. In this sense, the patch-surrogate approach enables a training paradigm of a similar spirit to Ref. [31], in which quantum resources are devoted solely to the construction and updating of the surrogates. We investigate this setting and compare its performance with that of the standard gradient-descent approach. Although our preliminary results do not yield definitive conclusions as there exist regimes in which each method outperforms the other, this observation points to a promising new direction for exploration. In particular, surrogates may offer significant practical benefits given the effect of noise drift and in the context queuing times for cloud-based hardware.

Beyond variational quantum computing, our results have implications in numerous domains including quantum sensing and quantum simulation. For quantum sensing, one could use our surrogated landscape for inference-based techniques [43,44]. In the case of dynamical simulation, the landscape is that of physical expectation values spanned by the system parameters and can be used to analyze hardware data or extend coherence times. Our numerical simulations, and in particular the long-time simulation of a 127-qubit transverse field Ising model on a heavy-hex topology, demonstrate our algorithms' utility beyond the regimes in which we have strict analytic guarantees.

II. BACKGROUND

We consider the task of simulating expectation values of the form

$$f(\alpha) = \text{Tr}[\mathcal{E}_\alpha(\rho)O], \quad (1)$$

where ρ is an n -qubit input state, O is a Hermitian operator, and $\mathcal{E}_\alpha$ is a quantum channel parameterized by an

m -component vector of real parameters α . By varying the choice in ρ , O , and $\mathcal{E}_\alpha$ this expression can be used to capture observables and losses for a wide range of variational quantum algorithms, real-time simulations, and quantum machine learning tasks [45,46].

In the following, we will briefly provide some background on the notions of classical simulation and classical surrogatability. We will then introduce a classical algorithm for classical-quantum surrogation. In particular, these techniques constitute a variant of low-weight simulation algorithm (LOWESA), introduced in Refs. [2,14], and more generally belongs to the family of Pauli propagation algorithms [2,4,14,26,27,47–55].

A. Classical simulation and surrogation

By now, it is well established that a large class of parameterized quantum circuits exhibit the so-called *barren plateau* phenomenon where their parameter landscapes concentrate exponentially with system size [56,57]. On a barren plateau, the probability that any randomly chosen parameter vector α corresponds to an observable expectation value which deviates significantly from its average vanishes exponentially with the number of qubits n upon which the circuit acts [58]. This is widely believed to prohibit the training with random initializations for a broad family of variational quantum algorithms [42, 57–72] and quantum machine learning models [73–82]. Nevertheless, it leaves open the possibility that small subregions of the landscape exhibit substantial landscape variances.

These realizations have prompted recent efforts into variance lower bounds for subregions of quantum landscapes. Early work investigated small angle initializations for the hardware-efficient variational *Ansatz* [33,36,37,83, 84] and Hamiltonian variational *Ansatz* (HVA) [34,85]. Subsequent work derived bounds for iterative learning schemes in the context of variational quantum simulation [38], as well as for ground state preparation [40]. These approaches have recently been unified in Ref. [41] and further extended to generative modeling with a nonlinear loss in Ref. [42]. Common to all these works is the observation that it is possible to guarantee landscape variances that vanish at worst polynomially in the number of trainable parameters in special subregions that have a radius that shrinks polynomially in the number of parameters.

In parallel to this line of work, it has been observed that there is a strong correlation between being able to prove that a variational quantum landscape is barren plateau free (i.e., proving that the variance of expectation values vanishes, at worst, polynomially) and the ability to create a classical surrogate for that landscape [8]. Here we demonstrate that in all the reduced parameter regimes where it is possible to provide polynomial (in n or m) lower bounds on the variance of the expectation value, it is possible

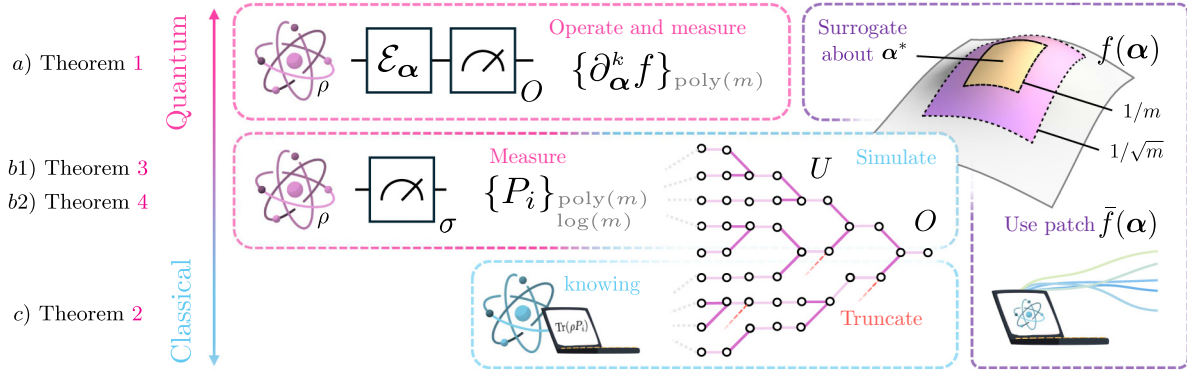


FIG. 1. The process of surrogating the expectation function $f(\alpha)$ obtained as the expectation value of an observable O , measured over a state ρ evolved under the m -parameter circuit $U(\alpha)$. Quantum measurements and/or classical simulations via truncated Pauli propagation are used to surrogate a patch $\bar{f}(\alpha)$ which can be used in lieu of further quantum operations for applications such as variational quantum algorithm (VQA).

to construct a classical surrogate of the expectation value landscape.

To pin down what we mean by a classical surrogate in this context, let us start by defining what it means to simulate an expectation value. We will say that an algorithm can classically simulate an expectation value $f(\alpha)$ with high probability if a polynomial-time classical algorithm can implement a function $\bar{f}(\alpha)$ approximating the expectation value up to error ϵ , i.e.,

$$\Delta f(\alpha) := |\bar{f}(\alpha) - f(\alpha)| \leq \epsilon, \quad (2)$$

with probability $1 - \delta$ for random $\alpha \sim \mathcal{P}$, where $\mathcal{P}$ is some distribution of interest and where $\epsilon \in \Theta(1)$ and $\delta \in \Theta(1)$ are arbitrarily small constants. We will also be interested in computing the function for *any* parameter settings, i.e., the *worst-case* error. Thus, we will also consider a stronger version where an algorithm can compute the expectation function if Eq. (2) holds for all α in a given landscape region.

In contrast, preparing a *classical surrogate* involves using a quantum computer for an initial (polynomial-time) data acquisition phase. Since this data can be utilized by subsequent classical simulation, the combination is more powerful than a purely classical algorithm. More concretely, we say it is possible to efficiently “classically surrogate” a quantum expectation function $f(\alpha)$ over the distribution $\alpha \sim \mathcal{P}$ if, using data from an initial polynomial-time data acquisition phase on a quantum computer, it is subsequently possible to classically simulate expectation values over $\mathcal{P}$.

Crucially, the practical benefits of classically surrogating a quantum landscape depend on the family of channels that need to be run on the quantum computer in the non-adaptive polynomial-time data acquisition phase. For our most general results on the surrogatability of quantum landscapes, we will allow arbitrary polynomial depth and

width circuits to be run. In this case, shown in Fig. 1(a), the main advantage in constructing the classical surrogate landscape is that after an initial overhead, any subsequent algorithm can be run fully classically. Such a scheme transforms a potentially adaptive hybrid quantum-classical algorithm into a nonadaptive one, opening up the possibility of using the power of classical optimization tools such as automatic differentiation, while leaving the total quantum resources requirements unchanged. However, we will also consider more specialized cases where one only needs to perform Pauli measurements on the initial state. In this case, shown in Fig. 1(b), the quantum resource requirements are substantially reduced as compared to running the original quantum channel to evaluate different points on the expectation landscape. This quantum resource reduction is achieved by offloading much of the work of simulating the circuit to a classical simulation algorithm.

It is possible to consider many different possible distributions $\mathcal{P}$ of parameters over subregions of a quantum landscape here. Here, in line with the small-angle initialization literature [34,36,38,39], we will focus on uniform distributions over hypercubes of width $2r$. More concretely, let us define

$$\mathcal{V}_m(\alpha^*, r) := \{\alpha\} \quad \text{such that } \alpha_i \in [\alpha_i^* - r, \alpha_i^* + r] \forall i \quad (3)$$

as the hypercube of parameter space centered around the point α^* , and

$$\mathcal{D}_m(\alpha^*, r) := \text{Unif}[\mathcal{V}_m(\alpha^*, r)] \quad (4)$$

as a uniform distribution over the hypercube $\mathcal{V}_m(\alpha^*, r)$. We expect that our conclusions straightforwardly generalize to other reduced parameter range distributions including Gaussian distributions [33,37] or to allow for independent perturbation radii for each parameter.

B. Pauli propagation

There exist numerous classical techniques to simulate quantum circuits and evaluate the expectation function in Eq. (1). In this work, we employ Pauli propagation methods [2,4,14,26,27,49,50,55,86,87]. In this context, it is most natural (and most efficient) to consider quantum circuits written as parameterized perturbations from Clifford circuits. That is, unitaries of the form

$$U_{\text{CP}}(\alpha) = \prod_{i=1}^m C_i U_i(\alpha_i), \quad (5)$$

where $\{C_i\}_{i=1}^m$ are a set of fixed but arbitrary Clifford unitaries and $\{U_i(\alpha_i) = e^{-i\alpha_i P_i/2}\}_{i=1}^m$ are parameterized rotations generated by Pauli operators $\{P_i\}_{i=1}^m$. This architecture can be universal even when U_i are single-qubit phase rotations, though we allow U_i to be Pauli rotations acting on any number of qubits. In our analytics we will consider both the case where the parameters α_i are uncorrelated and cases where they are correlated. We further emphasize however that this particular circuit form is inessential to our most general complexity bounds in Sec. III which can, for example, also quantify expectation landscapes for parameterized channels, and for circuit which are not near-Clifford.

Pauli propagation methods typically work in the Heisenberg picture where the initial operator (here the observable O) is often sparse in the Pauli basis. In this basis, the operator is backpropagated through the circuit and finally overlapped with the initial state ρ . In particular, since the adjoint action of any Clifford operation on any Pauli observable P_j simply produces another Pauli P'_j ,

$$C_i^\dagger P_j C_i = s_{ij} P'_j, \quad s_{ij} \in \{-1, 1\}, \quad (6)$$

which we will notate as $P_j \xrightarrow{C_i} \pm P'_j$. These transformations can either be computed in $\mathcal{O}(n^2)$ time using the Gottesman-Knill theorem [88] or in $\mathcal{O}(1)$ time for any *a priori* known C_i acting on few qubits by use of look-up tables. Applying a non-Clifford Pauli rotation $e^{-i\alpha_i P_i/2}$ to P_j leaves the operator invariant if $[P_i, P_j] = 0$, but otherwise produces a real-valued weighted sum of two Pauli operators,

$$P_j \xrightarrow{e^{-i\alpha_i P_i/2}} \cos(\alpha_i) P_j + \sin(\alpha_i) P'_j, \quad (7)$$

where $P'_j = i[P_i, P_j]/2$ is a Pauli operator. We note that the sign of the sine term is flipped in the Schrödinger picture. By iterating this procedure, one ends up with a collection of Pauli operators and corresponding trigonometric prefactors.

After Pauli propagation, the expectation function in Eq. (1) takes the form

$$f(\alpha) = \sum_P c_P(\alpha) \text{Tr}[\rho P], \quad (8)$$

where c_P are the coefficients of the backpropagated Pauli operators P . These coefficients capture both the initial weight of each of the relevant Paulis in the target observable O and the sine and cosine coefficients that have been picked up during the propagation.

One approach, embracing the concept of expectation function surrogates [2,14,27], is not to compute the coefficients c_P of the Pauli paths from scratch every time the parameters α are changed. Instead, one can symbolically or graphically represent the coefficients for each final Pauli P . When it becomes clear which P contribute significantly to the expectation in Eq. (8), only their corresponding coefficients c_P can be reevaluated. The approach has a substantial initial time and especially memory overhead, but once performed, enables rapid reevaluation of the expectation value $f(\alpha)$ for any α . The implementation used in this work can be seen as an improved variant of the original proposal of LOWESA. We give a clarifying illustration of the technique in Fig. 2.

In its raw form, the number of Pauli operators in the backpropagated observable (what we call the “*Pauli paths*”), or equivalently the number of terms in the sum in Eq. (8), grows exponentially with the number m of Pauli rotations in the circuit. To make the algorithm efficient

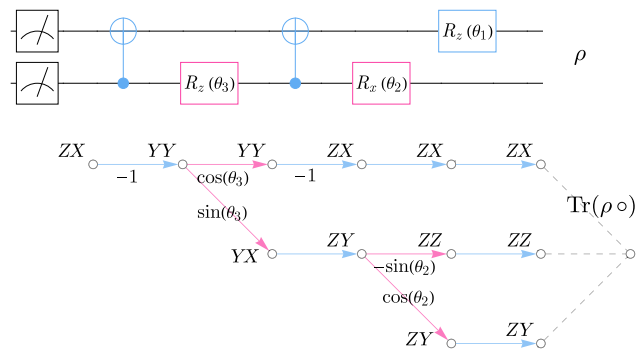


FIG. 2. A visualization of Pauli propagation computing the expectation value of observable $O = ZX$ resulting from the pictured two-qubit circuit upon input state ρ . Evaluation proceeds left-to-right, simulating the circuit in reverse; graph edges indicate the Pauli strings produced by the gate aligned above them, and their labels indicate the induced coefficient; elements of the gate’s Pauli transfer matrix. The Clifford CNOT gates merely multiply Pauli strings by ± 1 (blue edges), while the non-Clifford rotations $R_P(\theta) = e^{-i\theta P/2}$ can sometimes produce *two* Pauli strings (pink edges). Notice the state ρ need not be explicitly instantiated; instead, only its trace with respect to the final Pauli strings are evaluated.

for larger m values, one must therefore apply truncation schemes. Truncating higher weight Pauli operators has been shown to be efficient for both noisy circuits [14,48,52] and random circuits [26]. Here we will provide guarantees for a *small-angle* truncation scheme for patches of expectation landscapes corresponding to perturbations away from a Clifford circuit. We note that this approach is related to the techniques numerically demonstrated in Ref. [50].

III. GENERAL PATCHES

A. Error guarantees

Our first result is a general bound which shows that it is always possible to create a classical surrogate for small subregions of a quantum landscape generated by a general parametrized quantum channel. Concretely, we show that for any function that depends on m independent parameters, it is possible to create a surrogate of a patch of width $\sim 1/\sqrt{m}$, provided that some assumptions on the magnitude of the higher-order derivatives are satisfied. This result is captured by Theorem 1 below, which we prove and state more formally in Appendix C (Supplemental Theorem C1).

Theorem 1 (General Surrogation Guarantee, Informal). Consider an expectation function $f(\alpha)$ of the form in Eq. (1), where the measurement operator satisfies $\|O\|_\infty \in \mathcal{O}(1)$, and where the parameters of the quantum channel are taken from a hypercube region $\mathcal{V}_m(\alpha^*, r)$ around an arbitrary point α^* of width $2r$. Assume that the partial derivatives within this region scale at most exponentially in the derivative order k , i.e.,

$$|\partial_{i_1, i_2, \dots, i_k} f(\alpha)| \leq \gamma^k \|O\|_\infty, \quad \forall \alpha \in \mathcal{V}_m(\alpha^*, r), \quad (9)$$

for some constant $\gamma \in \mathcal{O}(1)$. Furthermore, assume that the partial derivatives for $k \in \mathcal{O}(1)$ can be estimated efficiently on a quantum computer. Then, it is possible to efficiently classically surrogate $f(\alpha)$ up to an arbitrary small error $\epsilon \in \Theta(1)$ with high probability for random $\alpha \sim \mathcal{D}_m(\alpha^*, r)$ for

$$r \in \mathcal{O}\left(\frac{1}{\sqrt{m}}\right) \quad (10)$$

and for all α in $\mathcal{V}_m(\alpha^*, r)$ for

$$r \in \mathcal{O}\left(\frac{1}{m}\right), \quad (11)$$

where in both cases there is an arbitrarily small but constant failure probability from shot noise in the data-collection phase.

The proof of Theorem 1 relies on expressing the expectation value $f(\alpha)$ with Taylor's expansion around the point of interest α^* and then showing that with some mild assumptions regarding the partial derivatives (see Assumptions C1–C3 in Appendix C), a low-order truncation of the expansion gives a sufficiently good surrogate for a small region [89]. These assumptions are in fact met by most parameterized quantum circuits considered currently by the community (see Appendix C 1 f) and could also be applied to parameterized channels. In particular, it includes not just the uncorrelated Pauli rotation *Ansätze* in Eq. (5) but also arbitrarily deep circuits where the m independent parameters can appear in multiple gates. To name just one example, the theorem also holds for gates using Givens Rotations, which are crucial elementary building blocks for quantum chemistry [90].

It is noteworthy that our average-case surrogacy guarantee aligns with the growing body of prior work on absence of barren plateau proofs for patches of a quantum expectation function landscape that scale as $1/\sqrt{m}$. In particular, it has been shown in Refs. [38,39] that a region with $r \in \mathcal{O}(1/\sqrt{m})$ around some minimum α^* has substantial gradients

$$\text{Var}_{\alpha \sim \mathcal{D}_m(\alpha^*, r)} [f(\alpha)] \in \Omega\left(\frac{1}{\text{poly}(n)}\right). \quad (12)$$

Theorem 1 hence implies that, in all of the subregions where it is possible to prove absence of barren plateaus, it is not necessary to run a variational quantum algorithm adaptively on a quantum computer but rather one could first quantumly prepare a surrogate which is then optimized fully classically.

Beyond the average-case surrogatability guarantee, we additionally prove the worst-case error incurred by our algorithm in a smaller patch of width $\mathcal{O}(1/m)$. Inside this smaller patch, there exists no parameter vector α which is classically simulated with an error worse than the constant, but arbitrarily small, chosen error. As discussed further in the Appendixes below Theorem E4, this implies that in the context of dynamical simulation Trotter errors can be reduced with only a polynomial overhead.

Theorem 1 leaves open the nature of the measurements required on the quantum computer, and in general these might require running highly nontrivial circuits. Namely, one will generally need to run circuits to evaluate the loss in order to compute gradients via the generalized parameter-shift rule [91]. Concretely, for general circuits where the patch is expanded around an arbitrary point α^* , the quantum device can be used to estimate the function $f(\alpha)$ at different choices in α and the surrogate is generated by reconstructing a multivariate Taylor expansion of $f(\alpha)$. In this case, the surrogate bears similarities to applying higher-order gradient optimization methods because generating the surrogate requires effectively measuring

higher-order derivatives of the loss. In this sense the distinction between running a variational quantum algorithm explicitly (by say updating by one step per iteration using a higher-order gradient-descent method) and generating the surrogate (by first measuring higher-order gradients and then training classically for a few steps) can become blurred.

Nonetheless, Theorem 1 suggests it may be worthwhile to forego experimentally reevaluating a quantum circuit with varying parameters *at each step*, and instead spend the quantum resources preparing a surrogate. The constructed surrogate can then be trained entirely classically for a few iterations within the patch. Nevertheless, for a whole training process to remain faithful, the surrogate will need to be reconstructed once the training trajectory leaves the patch.

In Appendix D, we present a preliminary exploration of this aspect. Specifically, we investigate the performance of an iterative surrogate training strategy and compare its quantum resource requirements—namely, the number of measurement shots—with those of standard gradient descent. We emphasize that, in this case, the surrogate is constructed directly at the level of the loss gradient, in contrast to the approaches discussed in the main text, where the loss function itself is surrogated. In addition, a shot-noise analysis of this gradient-based surrogate, including a practical approximation of the optimal shot-allocation strategy for reducing shot-noise, is provided in Appendix D4. In the next subsection we will explore one key component of this approach—namely, we investigate numerically how far it is possible to train in practice using the data obtained to build a surrogate.

B. Numerical tests

In this section we numerically investigate the maximum distance we can move in the parameter space before the surrogate becomes unusable. That is, the maximum distance such that training on the surrogate continues to decrease the true loss.

For concreteness, we consider a surrogate of an expectation loss function of the form of Eq. (1), where the observable is a Heisenberg Hamiltonian

$$O = \frac{1}{3(n-1)} \sum_{i=1}^{n-1} (X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}). \quad (13)$$

The initial state is the zero state, $\rho = |0\rangle\langle 0|^{\otimes n}$, and $\mathcal{E}_\alpha(\rho)$ is a variational *Ansatz* with $m = 3nL$ parameters composed of L -layers of single-qubit rotations followed by two-qubit ZZ gates of the form

$$U(\alpha) = \prod_{l=1}^L \left(\prod_{i=1}^n R_{ZZ}(\alpha_{i+2n,l}) \prod_{i=1}^n R_Z(\alpha_{i+n,l}) R_X(\alpha_{i,l}) \right). \quad (14)$$

We compute the surrogate of the gradient $\nabla \tilde{f}(\alpha)$ around the randomly chosen initial point α^* and use it to train the algorithm until $f(\alpha)$, the test loss, stops decreasing. Let α_0 denote the corresponding parameters at that step. Finally, we compute $\|\alpha^* - \alpha_0\|_x$, where $x \in \{\infty, 2, 1\}$, with respect to the number of parameters in the circuit m .

The results are presented in Fig. 3, where the diamonds are the numerical values of $\|\alpha^* - \alpha_0\|_x$ for $x = 1, 2, \infty$ after averaging for 20 runs with different initial points (with the maximum and minimum shaded), the solid lines are a fit of those points and the dashed lines represent the scalings we would expect from the bound. Namely, we recall that according to Theorem 1, $\|\alpha^* - \alpha_0\|_\infty \in \mathcal{O}(m^{-0.5})$, $\|\alpha^* - \alpha_0\|_2 \in \mathcal{O}(1)$, $\|\alpha^* - \alpha_0\|_1 \in \mathcal{O}(m^{0.5})$, compared to the $\|\alpha^* - \alpha_0\|_\infty \in \mathcal{O}(m^{-0.15})$, $\|\alpha^* - \alpha_0\|_2 \in \mathcal{O}(m^{0.2})$, $\|\alpha^* - \alpha_0\|_1 \in \mathcal{O}(m^{0.7})$, found in the experiment. We thus see that the experimental results have, roughly, between a $m^{0.2}$ and $m^{0.3}$ better scaling (i.e., the solid lines are angled above the

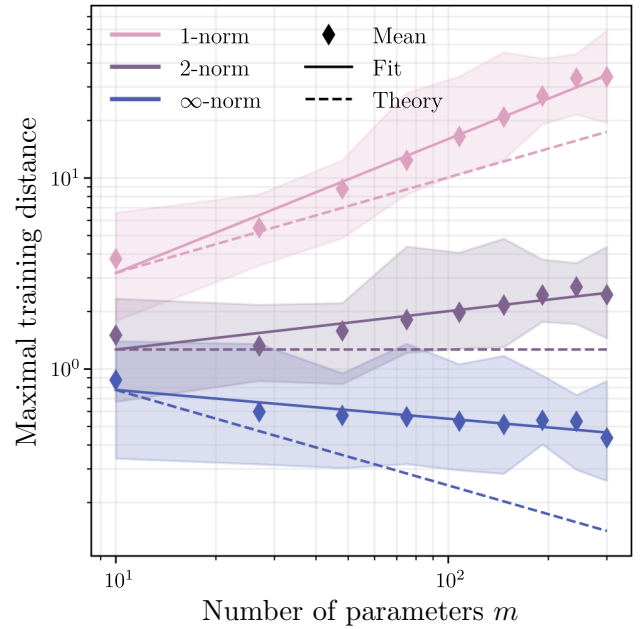


FIG. 3. Maximal training distance as function of the number of parameters. We consider a loss of the form of Eq. (1), where the observable is a Heisenberg Observable described in Eq. (13), the initial state is $\rho = |0\rangle\langle 0|^{\otimes n}$ and $\mathcal{E}_\alpha(\cdot)$ is a variational *Ansatz* with $m = 3nL$ parameters described in Eq. (14). We compute the surrogate of the gradient around a randomly chosen initial point α^* and use it to train the algorithm until the true loss $f(\alpha)$ stops decreasing and starts to increase, at α_0 . We compute the distance $\|\alpha^* - \alpha_0\|_x$, where $x \in \{\infty, 2, 1\}$, and plot it with respect to the number of parameters. We repeat the process 20 times and show the average in diamonds (with the maximum and minimum shaded), plot a fit as a solid line, and show the theoretical scalings according to Theorem 1 as a reference. The fitted lines correspond to the following functions $\|\alpha^* - \alpha_0\|_\infty = 1.10m^{-0.15}$, $\|\alpha^* - \alpha_0\|_2 = 0.80m^{0.20}$, $\|\alpha^* - \alpha_0\|_1 = 0.63m^{0.70}$.

dashed lines). This implies that the optimization process can use the surrogate for significantly longer distances than our bounds predict.

In Appendix D, we further provide a preliminary comparison between the quantum resources required for an iterative surrogate approach as compared to the standard hybrid quantum-classical approach. As there are many different hyper-parameters for this benchmarking problem, it is hard to fairly draw definitive conclusions. In particular, the standard gradient-descent approach is operationally simpler, as it requires tuning fewer hyper-parameters than the iterative surrogate strategy, which additionally introduces choices such as the frequency of the surrogate reconstruction. Moreover, gradients estimated directly on quantum hardware are unbiased, rendering the method more resilient to shot noise. Nevertheless, our numerics suggest that the bias introduced by the Taylor approximation can, in some regimes, be balanced by the increased statistical noise associated with repeated gradient estimation in the standard approach, although we do not claim this as a general result.

As such, we present both cases where the surrogate outperforms a standard variational quantum algorithm and vice versa. We therefore suggest that this Taylor surrogate should be viewed as a potential alternative to vanilla VQAs, rather than a replacement, with potential practical advantages given the problems associated with noise drift and/or queuing for cloud access.

IV. NEAR-CLIFFORD PATCHES

We will now specialize to particular families of problems for which the requirements of the quantum computer can be substantially reduced as compared to the Taylor surrogate approach. In particular, in this section, we consider a patch of a quantum landscape that is centered around a Clifford circuit. We utilize the Pauli propagation approach described in Sec. II B with a truncation method specialized for this near-Clifford region.

A. Error guarantees

Concretely, we consider the circuit composed of alternating layers of Clifford gates and uncorrelated Pauli rotations defined in Eq. (5). We then consider a hypercube of width $2r$ around the all-zero vector $\mathbf{0}$. In this region, the angles α_i are small and so we can leverage the fact that sine contributions to the coefficients c_P in Eq. (8) are much smaller than cosine contributions. We use this observation to prune Pauli paths with many sine coefficients, similar to the approach in Ref. [50]. Then, we show that this “small-angle” truncation strategy achieves an exponentially decreasing average error as a function of the sine expansion order κ (i.e., the number of sine terms kept).

1. Classical time complexity

We begin by presenting a theorem (derived in Appendix E 5), which studies the time complexity of the classical simulation algorithm.

Theorem 2 (Time complexity of small-angle Pauli propagation, Informal). Consider an expectation function $f(\alpha)$ of the form of Eq. (1) with an observable $O = \sum_{P \in \mathcal{P}_n} a_P P$, where at most $N_{\text{Paulis}} \in \mathcal{O}(\text{poly}(n))$ coefficients a_P are nonzero, with $\|\mathbf{a}\|_1 \in \mathcal{O}(1)$, and a parametrized circuit of the form of Eq. (5) with m independent parameters. If the expectation values $\text{Tr}[\rho P]$ of the initial state ρ with any Pauli operator P can be efficiently computed, a runtime of

$$t \in \mathcal{O}\left(N_{\text{Paulis}} \cdot m^{\log\left(\frac{1}{\epsilon^2 \delta}\right)}\right) \quad (15)$$

suffices to classically simulate $f(\alpha)$ up to an error ϵ with probability at least $1 - \delta$ for random $\alpha \sim \mathcal{D}_m(\mathbf{0}, r)$ with

$$r \in \mathcal{O}\left(\frac{1}{\sqrt{m}}\right). \quad (16)$$

Similarly, with a runtime of

$$t \in \mathcal{O}\left(N_{\text{Paulis}} \cdot m^{\log\left(\frac{1}{\epsilon}\right)}\right), \quad (17)$$

we can classically simulate $f(\alpha)$ up to an error ϵ for all α in the hypercube $\mathcal{V}_m(\mathbf{0}, r)$ with

$$r \in \mathcal{O}\left(\frac{1}{m}\right). \quad (18)$$

A more formal version of the theorem is stated in Supplemental Theorems E3 and E4.

Thus we see that Pauli propagation can be used to efficiently simulate Pauli expectation values for arbitrary circuits of the form Eq. (5) with high probability for all α in a hypercube of width $\mathcal{O}(1/\sqrt{m})$ and for all α in a smaller hypercube of width $\mathcal{O}(1/m)$. Notice that these results mirror the general result in Theorem 1, but for a family of circuits that is itself classically tractable. It follows that as long as we have a classical representation of the initial state ρ (analytically or, e.g., as a tensor network) from which we can efficiently compute expectation values of Paulis, then it is possible to classically simulate VQAs with high probability in the small angle regimes for which trainability guarantees have been derived [33,34,36,37]. Moreover, our numerics in Sec. IV B suggest that the above bounds are somewhat loose and improved resource scalings can be obtained in practice.

In Appendix E 4, we further provide a bound for the average error of an observable with a circuit where all angles are correlated (i.e., where all the rotation gates are

parametrized by the same α). Intriguingly, we find that the average-case simulation of such maximally correlated angles is asymptotically as hard as the worst case as a function of m . It is not clear whether the equal scaling is mostly due to our proof techniques, or whether simulating cases such as real-time dynamics of quantum systems is about as hard as the worst case.

An interesting corollary of our worst-case analysis is that we can provably efficiently simulate quantum dynamics of Hamiltonians with $\mathcal{O}(n)$ Pauli terms and $\mathcal{O}(1)$ -sized coefficients up to time $t = \mathcal{O}(1/n)$ on any geometry. This is because, for L Trotter layers with time step dt , we have $m = \Theta(nL)$ parameters and $dt = \mathcal{O}(1/m)$ is the largest guaranteed time step. We thus have $dt = \mathcal{O}(1/nL)$, and with $t = Ldt$ we find $t = \mathcal{O}(1/n)$ is guaranteed to be efficient.

Given our significantly more favorable average-case errors for uncorrelated angles it stands to reason that any independence of parameters interpolates between our two complexity bounds. In Sec. IV C we present a large-scale (127 qubits) numerical demonstration of Pauli propagation for a quantum simulation problem with a *time-dependent* Hamiltonian which supports this hypothesis.

2. Sample complexity

In the previous section, we assumed that we could efficiently estimate the overlap terms $\text{Tr}[\rho P]$ in Eq. (8). If these can be efficiently computed classically for any Pauli operator P , then the time complexity computed in the previous section fully quantifies the complexity of our small-angle patch simulation. If, instead, we do not have a classical representation of the state ρ , then we need a quantum device to measure the expectation value $\text{Tr}[\rho P]$. Here we quantify the number of measurements required, i.e., the sample complexity, in the case where ρ is non-classically simulable. As we will see, this sampling step is efficient in the sense that it requires only a polynomial number of Pauli measurements on quantum hardware.

There are several different tomographic methods that could be used to measure the necessary $\text{Tr}[\rho P]$. The simplest is to take the total shot budget N_s and divide them equally between each of the Pauli terms that need to be measured. In this case, analogously to Theorem 2, the number of terms that need to be measured scales polynomially with the number of parameters for a hypercube region with $r \in \mathcal{O}(1/\sqrt{m})$. More concretely, the following theorem holds.

Theorem 3 (Surrogation via Direct Pauli Measurements, Informal). Consider an expectation $f(\alpha)$ of the form of Eq. (1) with an observable $O = \sum_{P \in \mathcal{P}_n} a_P P$ containing at most $N_{\text{Paulis}} \in \mathcal{O}(\text{poly}(n))$ Pauli terms with $\|a\|_1 \in \mathcal{O}(1)$ and a parametrized circuit of the form of

Eq. (5) with m independent parameters. A shot budget of

$$N_s \in \mathcal{O}\left(N_{\text{Paulis}} m^{\log\left(\frac{1}{\epsilon^{2\delta}}\right)}\right) \quad (19)$$

suffices to classically surrogate $f(\alpha)$ with an error $\Delta f(\alpha)$ of at most ϵ with probability at least $1 - \delta$ for any random $\alpha \sim \mathcal{D}_m(\mathbf{0}, r)$.

A more formal version of the theorem together with the proof is provided in Appendix E 6 d in the form of Supplemental Theorem E9.

While the above strategy is simple, in general using the same number of shots to measure each of the Pauli terms is suboptimal. Indeed, if some of the Paulis have larger coefficients than others (on average over the patch of interest) it is advantageous to use more shots to measure the terms that contribute more to the final expectation value (see warm-up example in Appendix B 1 comparing different shot allocation strategies). In fact, it is precisely the exponentially decreasing contribution per sine-expansion order that we leverage for our efficient simulation guarantees. Hence, one should allocate the shots to each term in proportion to the average coefficient of that Pauli over the patch. In this case, the sample complexity becomes effectively independent of N_{Paulis} and instead scales with the 1-norm of the averaged coefficients. That is, we show that the number of shots scales as

$$N_s \in \mathcal{O}(\|c\|_{1,\text{avg}}^2), \quad (20)$$

where we call $\|c\|_{1,\text{avg}} := \sum_{P \in \mathcal{P}_n} \sqrt{\mathbb{E}_{\alpha \sim \mathcal{D}}[c_P(\alpha)^2]}$ an “effective” 1-norm. We describe this quantity in more detail in Appendixes B 2 and E 6 a. The surrogation guarantees with the effective 1-norm allocation strategy are further provided in Appendix E 6 b (average case) and Appendix E 6 c (worst case).

We note that a similar strategy can be employed if one is not interested in surrogating a whole patch but rather evaluating a single expectation value for a given α_0 without running $U(\alpha_0)$ on the quantum computer. This is often true in the context of dynamical simulation, where one seeks $f(\alpha)$ at only a handful of parameter values. In this case, by allocating the total shot budget to Pauli P in proportion $|a_P|$ one obtains a sample complexity that scales as $N_s \propto \|a\|_1^2$.

Another popular family of tomographic procedures are *classical shadows* [92–94]. Their variant for Pauli expectation values, which we here call Pauli classical shadows, enjoy tight accuracy guarantees but require more restricted families of circuits for optimal efficiency. Here we suppose that the Clifford layers consist exclusively of single-qubit Clifford gates and the Pauli rotation layers are $\mathcal{O}(1)$ -local. Now, each of the backpropagated Pauli operators have Pauli weight at most $\mathcal{O}(\kappa)$, where κ is the sine-truncation order. Thus, for $\kappa \in \mathcal{O}(\log(n))$, i.e., the regime

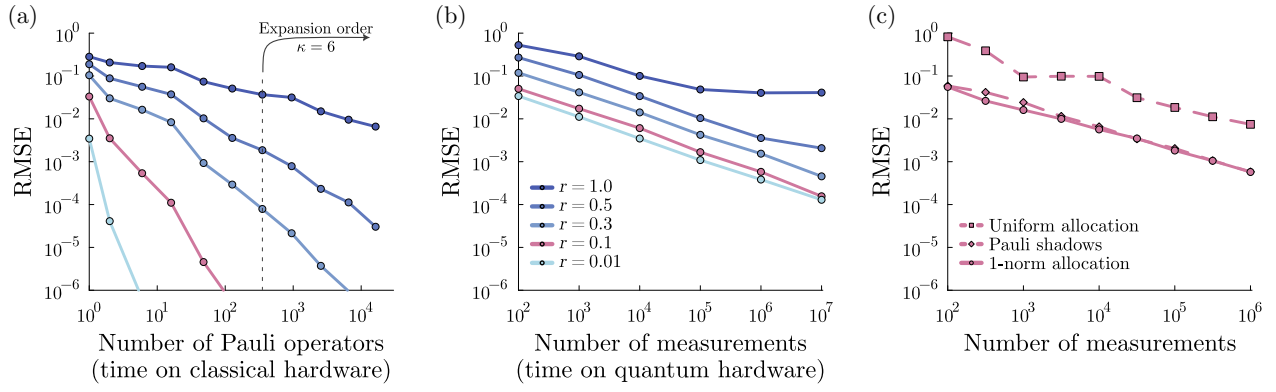


FIG. 4. Landscape and surrogation root mean square error as a function of classical and quantum resources. We consider a 16-qubit Hamiltonian variational *Ansatz* of a transverse field Ising model on a 4×4 grid and a 1-local Pauli Z expectation value near the middle of the grid. The initial state ρ is the all-zero state evolved exactly by four Trotter layers of the transverse field Ising Hamiltonian at $dt = 0.1$, and we surrogate the final four layers with free parameters. (a) Truncation root mean square error of our small-angle Pauli propagation surrogate for different r as a function of the number of backpropagated Pauli operators. This number corresponds to increasing sine-expansion order from $\kappa = 0$ to $\kappa = 10$. (b) Total root mean square error for expectation values using measurements of the initial state ρ for different r as a function of the number of quantum measurements at $\kappa = 6$. We use the effective 1-norm of the surrogate's coefficients described to allocate the measurements on ρ . (c) Comparison of different measurement allocation approaches at $\kappa = 6$ and $r = 0.1$. The uniform allocation over the operators expectedly underperforms and Pauli classical shadows are in this case on-par with our effective 1-norm allocation.

in which small-angle Pauli Propagation enjoys polynomial run times, we can employ classical shadows. This is captured by the following theorem.

Theorem 4 (Shadow Surrogation Guarantee, Informal). Consider an expectation function $f(\alpha)$ of the form of Eq. (1), an observable $O = \sum_{P \in \mathcal{P}_n} a_P P$ composed of at most $N_{\text{Paulis}} \in \mathcal{O}(\text{poly}(n))$ Pauli terms with $\|a\|_1 \in \mathcal{O}(1)$, and a parameterized quantum circuit of the form of Eq. (5) with the additional constraints that C_i are single-qubit Clifford operations and the Pauli rotations $U_i(\alpha_i)$ act on $\mathcal{O}(1)$ qubits. A shot budget of

$$N_s \in \mathcal{O}\left(\frac{\min(\log(n), \log(mN_{\text{Paulis}}))}{\epsilon}\right), \quad (21)$$

random Pauli measurements suffices to achieve a mean square error at most ϵ with high probability over a uniformly sampled hypercube $\mathcal{D}_m(\mathbf{0}, r)$ with $r \in \mathcal{O}(1/\sqrt{m})$.

A more formal version of this theorem is provided in Appendix E 6b as Supplemental Theorem E7. This tomographic procedure is well-studied and follows a “measure first, ask questions later” philosophy. Thus, one might wonder how this shot allocation strategy fares against the seemingly more informed effective 1-norm allocation described above. The general scalings in Appendix E 6b do not allow for a clear comparison between the approaches, but we compare them numerically in Fig. 4. While circuits with multi-qubit Clifford gates

would likely make Pauli classical shadows underperform, there is a significant upside in having collected a more general set of measurements that can be used to estimate other quantities with a similar error.

B. Numerical resource and error scalings

In this section, we present numerical experiments that indicate that both the time complexity and error of Pauli Propagation methods equipped with small-angle truncation outperforms the guarantees provided by our bounds.

First we test the error incurred by the truncated Pauli propagation together and the shot noise in an exactly verifiable system. We focus on the case of the HVA [95] for which reduced parameter range trainability guarantees were derived in Ref. [34]. This HVA family of circuits has the structure of the Trotterized time evolution under the Hamiltonian, but with mostly, or fully, free rotation angles. Here we choose a transverse field Ising (TFI) Hamiltonian

$$H_{\text{TFI}} = H_X + H_{ZZ}, \quad (22)$$

with

$$H_X = -\sum_i h_i X_i, \quad H_{ZZ} = -\sum_{\langle i,j \rangle} J_{ij} Z_i Z_j, \quad (23)$$

where X, Z are Pauli operators, and h_i, J_{ij} coefficients. The topology of the ZZ-interaction terms is that of nearest neighbors on a 4×4 grid in a 16-qubit system.

To demonstrate the freedom in our approach, we choose the initial state ρ as the exactly time-evolved all-zero

state for four Trotter layers with $h_j = J_{ij} = 1$ and $dt = 0.1$, and we surrogate the final four layers with the same HVA circuit *Ansatz* but with independent parameters. The observable is a Z observable in the middle of the grid.

In Fig. 4(a), we show the root mean square error (RMSE) of the truncated simulation as a function of the number of Pauli operators computed at increasing sine-truncation order κ , and for different angle ranges r . For example, the RMSE for $r = 0.1$ goes below 10^{-6} at expansion order $\kappa = 6$, which computes around 10^2 Pauli operators. As a reference, the corresponding classical surrogation procedure takes a few milliseconds on a standard laptop with expectation evaluations being significantly faster. If the initial state is classically simulable this time complexity effectively quantifies the total simulation cost. If not, then we additionally need to account for the effect of shot noise.

Moving onto the quantum resource requirements for nonclassically simulable initial states, in Fig. 4(b), we record the total error of the expectation value as a function of the number of measurements of ρ . We allocate the measurements according to the effective 1-norm allocation described before. The numerical results seem to indicate that the total error for $r \leq 0.5$ is dominated by shot noise even with 10^7 measurements, as those uncertainties are orders of magnitudes larger than the classical approximation error for small r . This argument can be exemplified by the extreme case of $r = 0$, where a single operator perfectly reconstructs the “landscape.” Yet, the total error scales as $\sim 1/\sqrt{N_s}$, where N_s is the number of measurements.

If Fig. 4(c), we compare the effective 1-norm allocation to uniform allocation over all operators and to Pauli classical shadows. While uniform allocation understandably underperforms, Pauli classical shadows and the effective 1-norm allocation are on-par for this example. However, the setup considered here is among the best-case scenarios for Pauli classical shadows, with more global observables or patches around nonidentity unitaries deteriorating its relative performance.

These results are complemented by the plots shown in Fig. 9 in Appendix E7. There we showcase a Heisenberg model HVA circuit structure instead of the TFI HVA in Fig. 4. The number of generated Pauli strings is larger, which, broadly speaking, increases the surrogation and measurement errors at the same classical or quantum resources, respectively. Due to the increased structure and number of commuting, overlapping terms in the Heisenberg Hamiltonian, the 1-norm estimation strategy underperforms without additional grouping of simultaneously measurable terms. How to best pair commuting observables and allocate quantum computing resources for the simulation of quantum landscape patches remains an open problem for future work.

C. Large-scale small-angle Pauli propagation

In this section, we numerically demonstrate small-angle Pauli propagation on a large-scale problem. When picking a system to showcase the algorithm, there is a fine balance to be struck between picking a complex problem that is beyond the realm of exact simulation methods and yet one where we can be confident in the results obtained. To address this tension we will draw motivation from Ref. [96], where the authors used analytically known scalings for benchmarking. While Ref. [96] takes this approach to benchmark a quantum simulation on a quantum device, we here apply the same logic to benchmark our approximate classical simulation at scale.

The system we consider is a time-dependent transverse field Ising (TFI) model with 127 spins and on a heavy-hex entangling topology [1]. The time-dependent TFI Hamiltonian is related to the time-independent TFI Hamiltonian in Eq. (22) and reads

$$H(t) = (1 - g(t))H_X + g(t)H_{ZZ}. \quad (24)$$

As before, we set the coefficients in H_X and H_{ZZ} [see Eq. (23)] to $h_i = J_{ij} = 1$. Importantly, the ramp function $g(t)$ is chosen such that it interpolates between the two Hamiltonians H_X and H_{ZZ} from time $t = 0$ to the final time $t = t_f$, i.e., $g(0) = 0, g(t_f) = 1$. Given the initial state $\rho = |+\rangle\langle+|^{\otimes n}$, the evolution begins in the ground state of the transverse Hamiltonian. This initial state and other stabilizer states like the all-zero state $|0\rangle\langle 0|$ allow for very fast and exact evaluation of the final overlaps $\text{Tr}[\rho P]$ for any Pauli operator P .

The phenomenon we leverage to benchmark our classical surrogate algorithm is the *Kibble-Zurek* (KZ) mechanism [97–99]. It predicts the density of defects in the evolved state relative to the ground state of the final Hamiltonian (here H_{ZZ}) as the system crosses a critical point with a finite annealing time t_f . Defects here are detected by nonzero values for quantities like $1 - \langle Z_i Z_j \rangle$ between any two sites i, j . In the limit of infinitely slowly ramping between H_X and H_{ZZ} , the state remains in the instantaneous ground state and there will be no defects. For a 1D TFI model, the density of defects n_{def} has been shown to scale as $n_{\text{def}} \propto t_f^{-1/2}$. Interestingly, the authors in Ref. [96] conjecture that experiments on the heavy-hex topology should follow a very similar scaling because the topology is closer to 1D than 2D.

Here we provide numerical evidence for this conjecture by simulating the time evolution of the time-dependent TFI model in Eq. (24) with 127 spins on a heavy-hex topology. We use a smaller time step of $dt = 0.3$, use the full heavy-hex lattice, and simulate to significantly longer times than possible with the quantum device in Ref. [96]. In this simulation, we use a combination of *weight* truncation [26] and our sine-expansion order truncation. Concretely, we truncate Pauli operators with Pauli weight of $W > 5$ (i.e., more

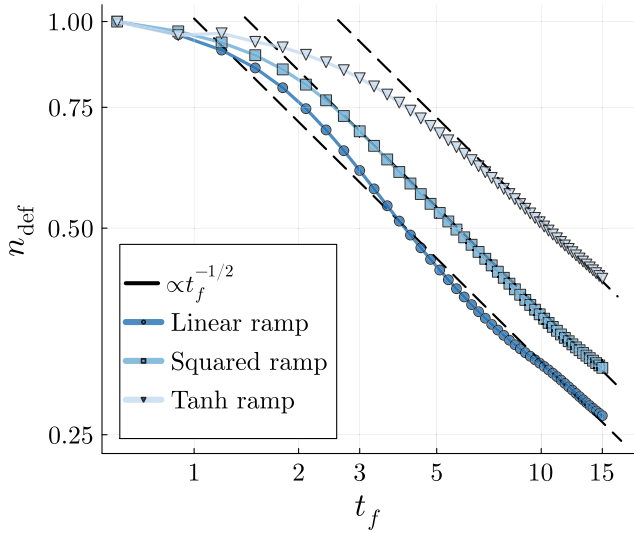


FIG. 5. Testing the Kibble-Zurek scaling of defects n_{def} in a time-dependent transverse field Ising model with 127 spins on a heavy-hex entangling topology. Using our Pauli propagation surrogate, we verify the conjectured scaling resembling $t_f^{-1/2}$ on the heavy-hex topology. The ramp function $g(t)$ in Eq. (24) is varied to confirm a similar scaling across various annealing protocols. The surrogation of the expectation landscape with 50 Trotter steps and $m = 13\,550$ gates was done in under 30 s, with every expectation evaluation taking under 1 s.

than five nonidentity Paulis) and with a number of sines $\kappa > 21$. Constructing the classical surrogate for the $Z_{62}Z_{63}$ Pauli operator in the middle of the lattice with 50 Trotter layers took under 30 s using one CPU thread on a laptop. We then record the density of defects on these lattice sites via $n_{\text{def}} = 1 - \langle Z_{62}Z_{63} \rangle$ on the surrogate landscape at points corresponding to different final evolution times and annealing ramps. This took under 1 s per evaluation.

Our results can be seen in Fig. 5. The linear ramp is defined via $g(t) = t/t_f$, the square ramp via $g(t) = (t/t_f)^2$, and the tanh ramp via $g(t) = \frac{1}{2} [\tanh(-3(1 - (t/t_f)) + 3(t/t_f)) + 1]$ which sweeps over the center part of a rescaled tanh function. Not only do we observe that all three annealing ramps exhibit the analytically known scaling of the defect density, but we can verify the quality of our results by observing that the norm of the Heisenberg-evolved observable is very high between 84% and 99% its original 2-norm after 50 circuit layers. Thus we see that our Pauli propagation simulation algorithm vastly exceeds its theoretical guarantees and, arguably, the capabilities of most current quantum hardware.

It is evident that the classical simulations could be pushed significantly further, for example, to attempt exploring when the tree-like scaling on the heavy-hex geometry transitions into a 2D scaling. This is coincidentally the regime where the classical simulation should become substantially harder. Our work merely highlights

that PP surrogates with their fast reevaluation capabilities can be used to explore entire families of quantum protocols, particularly using our sine-truncation in the small-angle regime, for later deployment on quantum devices.

V. DISCUSSION

This work described classically *surrogating* a subregion (dubbed a “patch”) of an expectation landscape produced by a parameterized quantum circuit. The surrogation procedure may require quantum resources but subsequent evaluations can be performed fully classically. This alleviates the load of quantum computers in a wide range of quantum algorithms, like VQAs, quantum machine learning, dynamical simulation, and even quantum sensing, by greatly reducing the necessary gate count and the number of measurements.

We presented three main scenarios demanding progressively more modest quantum resources. Section III considered arbitrarily located patches in very general circuits acting on potentially unknown (but quantumly preparable) initial states. Here, the surrogation prescribes executing the deepest (inside this work) yet relatively undemanding, nonadaptive circuits and measurements. Section IV considered patches centered around parameters for which the circuit is Clifford. This simplifies the task by enabling efficient classical simulation of the quantum circuit with our Pauli propagation algorithm. In this case, for states that cannot be classically simulated, but can be prepared on quantum hardware, our surrogates require only polynomially few Pauli measurements upon the initial state. And for classically tractable initial states ρ , whereby $\text{Tr}[\rho P]$ with a Pauli string P is efficiently classically evaluable, the surrogate can now be computed entirely classically.

We highlight that our scalings for efficient surrogatability and simulation align with prior work on the absence of barren plateau guarantees inside patches of width $\mathcal{O}(1/\sqrt{m})$ [33–39]. Thus our findings here complement the conjecture connecting the provable absence of barren plateaus and efficient classical surrogatability [8].

More positively, however, we have expanded the corpus of quantum-enhanced classical algorithms, consequently expanding the collection of problems that can be tackled with both classical and quantum resources. These hybrid approaches are emerging as a notable intermediary between fully classical and quantum settings, thus nuancing the threshold of practical quantum advantage [100, 101]. In particular, classical surrogation of a patch can augment a wide class of quantum algorithms, addressing shortcomings like that of insufficient coherence times and gate fidelities to reliably prepare *Ansatz* states, but which can faithfully prepare and measure initial states. As quantum hardware improves, computations could be linked up *in the middle* of the quantum circuit, as we have shown in Fig. 4, which in turn reduces the classical complexity.

It may also alleviate the need for experimentally difficult parametrized unitaries that can be more straightforwardly applied classically. The surrogated patch also has utility as a *symbolic* approximation to the parametrized landscape. In addition to being evaluated at particular points, it can be analytically studied and ergo integrated with, for example, symbolic and global minimization techniques [102], as has been shown to improve VQA convergence [103].

In general, we have shown that quantum preparation of a classical surrogate is often efficient and can be used to simplify an algorithm’s prescribed quantum operations. We caution however that we have *not* shown this process is necessarily faster than the original quantum algorithm running on compatible hardware. Constants and prefactors may well mean, for example, that surrogation of a variational cost function can involve more total measurements and operations than direct quantum evaluation. This is indeed supported by our numerics in Appendix D where we find that Taylor surrogation outperforms our bounds, and may well have practical advantages, but has a broadly compatible performance to vanilla variational quantum algorithms. Even in the case of near-Clifford patches where the quantum resources required to surrogate are minimal, there could also end up being polynomial speed ups from running directly on quantum hardware when the initial state is quantum. Still, by adding these hybrid methods to our algorithmic arsenal and analyzing their performance, we hope to have *patched* a hole in the community understanding of what is classically simulable, surrogatable, and worthwhile deploying to quantum computers.

ACKNOWLEDGMENTS

R.P. acknowledges the support of the SNF Quantum Flagship Replacement Scheme (Grant No. 215933). M.S.R. acknowledges funding from the 2024 Google PhD Fellowship. A.A. and Z.H. acknowledge support from the Sandoz Family Foundation-Monique de Meuron program for Academic Promotion. T.J. acknowledges support of the NCCR MARVEL, a National Centre of

Competence in Research, funded by the Swiss National Science Foundation (Grant No. 205602). M.C. acknowledges supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory (LANL) under Project No. 20230049DR, and by the Quantum Science Center (QSC), a National Quantum Information Science Research Center of the U.S. Department of Energy (DOE). This work was also supported by were supported by the U.S. DOE, Office of Science, Office of Advanced Scientific Computing Research, under Computational Partnerships program. S.T. acknowledges Exchange Faculty Travel Grant: TG168033 from Chulalongkorn University as well as funding from National Research Council of Thailand (NRCT) (grant number N42A680126). S.T. further acknowledges Thailand Science Research and Innovation Fund Chulalongkorn University (No. IND_FF_69_258_2300_062), and funding from The Office of the Permanent Secretary of the Ministry of Higher Education, Science, Research and Innovation. This research has received funding support from the NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation (Grant No. B39G690073).

CODE AVAILABILITY

The Pauli propagation simulations in this work were conducted with the [PauliPropagation.jl](#) package publically available on GitHub. Code for other numerical experiments are available from the authors upon reasonable request.

DATA AVAILABILITY

The data that support the findings of this article are openly available. The raw data for the numerical experiments are available from the authors upon reasonable request.

APPENDIX A: PRELIMINARIES

1. Notation

General notation table	
Symbol	Definition
n	Number of qubits
α	A vector of variational real parameters and its i th component
α_i	i th component of α
m	Number of variational parameters i.e., the size of α
$f(\alpha)$	The expectation value $\text{Tr}[O\mathcal{E}_\alpha(\rho)]$ with an observable O , an initial state ρ and a parametrized quantum channel $\mathcal{E}_\alpha(\cdot)$
$\bar{f}(\alpha)$	Approximate function that can be efficiently implemented.
$\Delta f(\alpha)$	Difference between an approximate function and the original function $ \bar{f}(\alpha) - f(\alpha) $ evaluated at α
$\mathcal{V}_m(\alpha^*, r)$	Hypercube of width $2r$ around some fixed point α^* in the parameter space, $\mathcal{V}_m(\alpha^*, r) := \{\alpha\} \text{ such that } \alpha_i \in [\alpha_i^* - r, \alpha_i^* + r], \forall i$

$\mathcal{D}_m(\alpha^*, r)$	Uniform distribution of α over $\mathcal{V}_m(\alpha^*, r)$ also referred to as $\text{Unif}[\mathcal{V}_m(\alpha^*, r)]$
$\tilde{f}(\alpha)$	Surrogate of $f(\alpha)$.
$\hat{f}(\alpha)$	Empirical estimator of $f(\alpha)$.
$\widetilde{\hat{f}}(\alpha)$	Empirical estimator of the surrogate.
$f^{(P)}(\alpha)$	The subscript (P) refers to $f^{(P)}(\alpha) = \text{Tr}[\rho U^\dagger(\alpha) P U(\alpha)]$.
κ	Order of expansion of a surrogate $\tilde{f}(\alpha)$ of $f(\alpha)$.
N_s	Number of measurements.
$\hat{O}$	Empirical estimate of an observable with a state $\text{Tr}[\rho O]$.
$\hat{O}(\alpha)$	Empirical estimate of a backpropagated observable with a state $\text{Tr}[\rho O(\alpha)]$.
$\mathcal{S}$	Set of “experimentally friendly” observables. A typical choice is a set of Pauli i.e., $\mathcal{S} \subseteq \mathcal{P}_n$.
Λ	Upper bound on the infinity norm of all the elements in $\mathcal{S}$, i.e., $\Lambda \geq \ P\ _\infty \forall P \in \mathcal{S}$.
$\beta(\cdot)$	$\beta : \mathcal{S} \rightarrow]0, 1]$ such that $\sum_{P \in \mathcal{S}} \beta(P) = 1$ and $\beta(P) > 0 \forall P \in \mathcal{S}$.

General surrogate: Notation table

γ	Constant such that the k th partial derivatives are bounded as $ \partial_{i_1, i_2, \dots, i_k} f(\alpha) \leq \gamma^k \ O\ _\infty$
$\mathcal{E}_\alpha$	Parametrized quantum channel.
$\partial_{i_1, i_2, \dots, i_k}$	Partial derivatives with respect to parameters $\{\alpha_{i_1}, \alpha_{i_2}, \dots, \alpha_{i_k}\}$, i.e., $\partial_{i_1, i_2, \dots, i_k} := \frac{\partial^k}{\prod_{j=1}^k \partial \alpha_{i_j}}$.
$\partial^{\mathbf{k}}$	Partial derivatives of order k_l with respect to each parameter α_l where $\mathbf{k} = (k_1, k_2, \dots, k_m)$, i.e., $\partial^{\mathbf{k}} := \prod_{l=1}^m \frac{\partial^{k_l}}{\partial \alpha_{i_l}^{k_l}}.$
b_j	The set $\{b_j\}$ is defined as $\partial_{i_1, i_2, \dots, i_k} f(\alpha) = \sum_{j=1}^{N_d} b_j f(\alpha_j)$.
N_d	Is the number of coefficients b_j .
b_0	$b_0 := \max_j b_j $.
δ_i	$\delta_i = \alpha_i - \alpha_i^*$
$\Delta_{i_1, \dots, i_k}(\mathbf{x}, \mathbf{x}^*)$	$\Delta_{i_1, i_2, \dots, i_k}(\mathbf{x}, \mathbf{x}^*) = (x_{i_1} - x_{i_1}^*)(x_{i_2} - x_{i_2}^*) \cdots (x_{i_k} - x_{i_k}^*)$

Pauli propagation surrogate: Notation table

Symbol	Definition
$\mathcal{P}_n$	n -qubit Pauli basis, i.e., $\{I, X, Y, Z\}^{\otimes n}$.
$ P $	Weight of a Pauli $P \in \mathcal{P}_n$, i.e., number of qubits P acts nontrivially on.
N_{Paulis}	For $O = \sum_{P \in \mathcal{P}_n} a_P P$, N_{Paulis} corresponds to the number of different $a_P \neq 0$.
a_P	Coefficients in which an observable is decomposed in the Pauli basis.
$c_P(\alpha)$	Coefficients of propagated observable $O(\alpha) = U(\alpha)^\dagger O U(\alpha)$ in the Pauli basis.
$O(\alpha)$	Backpropagated observable.
$\tilde{O}(\alpha)$	Truncated version of the backpropagated observable.
O_ω	Backpropagated observable along a certain path ω .
ω	Pauli path.
$\Phi_\omega(\alpha)$	Product of the coefficients obtained when propagating a Pauli P through a Pauli path.
$\phi_{\omega_i}(\alpha_i)$	Individual coefficients that are picked up after backpropagating a Pauli through one Pauli rotation. Check Eq. (E19) for the possible values.
d_ω	Overlap of the backpropagated Pauli along the path ω with the initial state $\text{Tr}[\rho P_\omega]$.
$n_{\text{pth}}^{(\kappa)}$	Number of paths found by the Pauli Propagation algorithm with truncation order κ .
T	Time that it takes to do each Pauli Propagation Step.

2. Concentration of measure

We recall some useful tail bounds which we will employ throughout the paper. For further details, see also Ref. [104] and references therein.

Lemma A1 (Markov's inequality). Let X be a positive random variable (i.e., $X \geq 0$) and $\epsilon > 0$. Then, we have

$$\Pr[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}. \quad (\text{A1})$$

Lemma A2 (Hoeffding's inequality [105]). Given $\epsilon, \delta \in (0, 1)$, $c > 0$, let $N_s = \log(2/\delta)c^2/(2\epsilon^2)$. Suppose $X_1, X_2, \dots, X_{N_s}$ are independent random variables taking values in $[a, b]$ such that $|a - b| \leq c$. Let X denote their mean. Then for any $\epsilon > 0$,

$$\Pr[|X - \mathbb{E}[X]| \geq \epsilon] \leq \delta. \quad (\text{A2})$$

3. Central limit theorem and moments of a normal distribution

We present and explain the Central Limit theorem and discuss the moments of a random variable that follows a normal distribution. For more information we refer the reader to Ref. [106]

Supplemental Theorem A1 (Lindeberg-Lévy: Central Limit theorem). Suppose that $\{\alpha_i\}_{i=1}^m$ is an independent sequence of m random variables having the same distribution with mean $\mathbb{E}[\alpha_i] = \mu$ and finite positive variance $\text{Var}[\alpha_i] = \sigma_1^2$. If $S_m = \sum_{i=1}^m \alpha_i$, then

$$X_m = \frac{S_m - m\mu}{\sigma_1\sqrt{m}} \xrightarrow{m \rightarrow \infty} \mathcal{N}(0, 1), \quad (\text{A3})$$

i.e., the random variable X_m tends to a normal distribution with mean $\mu_{\mathcal{N}} = 0$, and variance $\sigma_{\mathcal{N}} = 1$.

Using Supplemental Theorem A1 for a sequence of random variables $\{\alpha_i\}_{i=1}^m$ with $\mathbb{E}[\alpha_i] = 0$ we get that the probability distribution of the random variable X_m tends to behave as a normal distribution with variance 1 and mean 0, or in other words,

$$X_m = \frac{S_m}{\sigma_m} \xrightarrow{m \rightarrow \infty} \mathcal{N}(0, 1), \quad (\text{A4})$$

where we have used that $\sigma_1\sqrt{m} = \sigma_m$ as all the α_i are independent and have the same mean.

The k th moment of such distribution is defined as the expected value of the variable to the k th power, i.e., $\mathbb{E}[X_m^k]$. This can be computed as follows [107]:

$$\mathbb{E}[X_m^k] = \int_{-\infty}^{\infty} X_m^k \frac{1}{\sqrt{2\pi}} e^{-\frac{x_m^2}{2}} = \begin{cases} 0 & \text{if } k \text{ odd,} \\ (k-1)!! & \text{if } k \text{ even,} \end{cases} \quad (\text{A5})$$

where $a!!$ is the double factorial defined for $a \in \mathbb{N}$ as $a!!(a-1)!! = a!$, or in other words $a!! = a(a-2)(a-4)(a-6) \cdots 1$.

4. Binomial inequalities

In subsequent sections in the appendix, we will make frequent use of the following binomial inequalities:

$$\left(\frac{m}{k}\right)^k \leq \binom{m}{k} \leq \frac{m^k}{k!} < \left(\frac{em}{k}\right)^k, \quad (\text{A6})$$

where we assume $0 \leq k \leq m$ and e is the Euler number. In fact, the rightmost inequality is loose enough so that for

$k \leq m$ we have

$$\sum_{j=0}^k \binom{m}{j} \leq \left(\frac{em}{k}\right)^k. \quad (\text{A7})$$

We also prove the following more general lemmas.

Lemma A3. For any $x \in [0, 1]$ such that $k \leq mx$, we have

$$\sum_{j=0}^k \binom{m}{j} x^j \leq \left(\frac{emx}{k}\right)^k. \quad (\text{A8})$$

Proof.

$$\begin{aligned} \sum_{j=0}^k \binom{m}{j} x^j &\leq \sum_{j=0}^k \frac{(mx)^j}{j!} = \sum_{j=0}^k \left(\frac{mx}{k}\right)^j \frac{k^j}{j!} \\ &\leq \left(\frac{mx}{k}\right)^k \sum_{j=0}^k \frac{k^j}{j!} \leq \left(\frac{emx}{k}\right)^k, \end{aligned} \quad (\text{A9})$$

where we first use Eq. (A6). The second inequality is obtained from the condition $k \leq xm$ i.e., $(xm/k)^j$ increases with j so reaches its maximum for $j = k$. The last inequality is obtained by adding terms in the sum (up to $j = \infty$) and recognizing the exponential series. ■

Lemma A4. For any x such that $x \leq k/m$ we have

$$\sum_{j=k}^m \binom{m}{j} x^j \leq \left(\frac{emx}{k}\right)^k. \quad (\text{A10})$$

Proof. The proof of Eq. (A10) is similar to Eq. (A8) and reads as follows:

$$\begin{aligned} \sum_{j=k}^m \binom{m}{j} x^j &\leq \sum_{j=k}^m x^j \frac{m^j}{j!} = \sum_{j=k}^m \left(\frac{xm}{k}\right)^j \frac{k^j}{j!} \\ &\leq \left(\frac{xm}{k}\right)^k \sum_{j=k}^m \frac{k^j}{j!} \leq \left(\frac{emx}{k}\right)^k, \end{aligned} \quad (\text{A11})$$

where we first use Eq. (A6). The second inequality is obtained from the condition $xm \leq k$ i.e., $(xm/k)^j$ decreases with j so reaches a maximum for $j = k$. The last inequality is obtained by adding terms in the sum (from $j = 0$ to $j = \infty$) and recognizing the exponential series. ■

5. Lagrange multipliers

In this appendix we present the method of Lagrange multipliers to find the minima of a function subjected to some constraints. We refer the reader to Ref. [108]. The optimization algorithm works as follows.

a. Lagrange multipliers method: Optimization algorithm

- (1) Let $f(\mathbf{x})$ to be the function to optimize.
- (2) Let $g(\mathbf{x}) = 0$ to be the constrain function.
- (3) Define the Lagrangian: $\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$.
- (4) Compute the derivatives: $(\partial/\partial x_i)\mathcal{L}(\mathbf{x}, \lambda)$ and $(\partial/\partial \lambda)\mathcal{L}(\mathbf{x}, \lambda)$.
- (5) Solve the following system of equations

$$\begin{cases} \frac{\partial}{\partial x_i}\mathcal{L}(\mathbf{x}, \lambda) = 0, \\ \frac{\partial}{\partial \lambda}\mathcal{L}(\mathbf{x}, \lambda) = 0. \end{cases}$$

- (6) Discard $\lambda^{(sol)}$.
- (7) The points $\mathbf{x}^{(sol)}$ are the solution to the problem.

6. Taylor remainder theorem

We now introduce the Taylor Remainder Theorem, which will be used below to bound the error arising from making a k -order expansion in the expectation function. For further details on the Taylor reminder theorem we refer the reader to Ref. [109].

Supplemental Theorem A2 (Taylor Reminder Theorem). Consider a multivariate differentiable function $f(\mathbf{x})$ such that $f : \mathbb{R}^m \rightarrow \mathbb{R}$ and some positive integer κ . The function $f(\mathbf{x})$ can be expanded around some fixed point $\mathbf{x}^* \in \mathbb{R}^m$ as

$$f(\mathbf{x}) = \sum_{k=0}^{\kappa} \sum_{i_1, i_2, \dots, i_k=1}^m \frac{1}{k!} \partial_{i_1, i_2, \dots, i_k} f(\mathbf{x}^*) \Delta_{i_1, i_2, \dots, i_k}(\mathbf{x}, \mathbf{x}^*) + \epsilon_{\kappa, \mathbf{x}^*}(\mathbf{x}), \quad (\text{A12})$$

where we are denoting

$$\partial_{i_1, i_2, \dots, i_k} f(\mathbf{x}^*) := \left(\frac{\partial^k f(\mathbf{x})}{\partial x_{i_1} \partial x_{i_2} \dots \partial x_{i_k}} \right) \Big|_{\mathbf{x}=\mathbf{x}^*}, \quad (\text{A13})$$

$$\Delta_{i_1, i_2, \dots, i_k}(\mathbf{x}, \mathbf{x}^*) = (x_{i_1} - x_{i_1}^*)(x_{i_2} - x_{i_2}^*) \dots (x_{i_k} - x_{i_k}^*). \quad (\text{A14})$$

Then, the remainder of this function $\epsilon_{\kappa, \mathbf{x}^*}(\mathbf{x})$ can be expressed as follows

$$\epsilon_{\kappa, \mathbf{x}^*}(\mathbf{x}) = \sum_{i_1, i_2, \dots, i_{\kappa+1}=1}^m \frac{1}{(\kappa+1)!} \partial_{i_1, i_2, \dots, i_{\kappa+1}} f(\tilde{\mathbf{x}}) \Delta_{i_1, i_2, \dots, i_{\kappa+1}}(\mathbf{x}, \mathbf{x}^*), \quad (\text{A15})$$

with $\tilde{\mathbf{x}} = \tau \mathbf{x} + (1 - \tau)\mathbf{x}^*$ for some $\tau \in [0, 1]$.

APPENDIX B: SHOT NOISE ANALYSIS FOR ESTIMATING OBSERVABLES WITH RESTRICTED MEASUREMENTS

In this appendix we analyze the effect of shot noise for different strategies of computing families of observables. We will use these results in the later appendixes to compute the effect of shot noise for our different surrogation strategies.

Given an observable O and a quantum state ρ , we consider the task of estimating the expectation values $\text{Tr}[\rho O]$ with quantum experiments. In many practical settings, a direct measurement of the observable O on the state ρ might be experimentally unfeasible. Thus, we define a set $\mathcal{S}$ of “experimentally friendly” observables such that we can decompose O as

$$O = \sum_{P \in \mathcal{S}} a_P P, \quad (\text{B1})$$

where $|\mathcal{S}|$ is of a manageable size and each P is easily measurable. A typical choice of $\mathcal{S}$ is a subset of the Pauli basis, i.e., $\mathcal{P}_n := \{I, X, Y, Z\}^{\otimes n}$. While in Secs. C 7 and E we will precise consider such basis, we here make no assumption, and leave $\mathcal{S}$ as a general set of observables. By measuring each $P \in \mathcal{S}$ individually, we can estimate an approximate value of $\text{Tr}[O\rho]$. We start by providing a general measurement protocol for estimating $\text{Tr}[\rho O]$ up to a desired error ϵ .

Lemma B1 (Parametric estimator with restricted measurements). Let $O = \sum_{P \in \mathcal{S}} a_P P$, where a_P are nonzero real coefficients and $\mathcal{S}$ is a set of observables, for which each element can be measured efficiently. Moreover, assume that $\|P\|_{\infty} \leq \Lambda$ for all $P \in \mathcal{S}$. Let $\beta : \mathcal{S} \rightarrow]0, 1]$ be a probability distribution whose support is $\mathcal{S}$ [i.e., $\beta(\cdot)$ satisfies $\sum_{P \in \mathcal{S}} \beta(P) = 1$ and $\beta(P) > 0 \forall P \in \mathcal{S}$].

- (1) There exists a randomized measurement protocol which uses Λ^2/ϵ^2 measurements on different copies of an unknown state ρ , to output an estimate $\hat{o}$ satisfying

$$\sqrt{\mathbb{E}[(\hat{o} - \text{Tr}[O\rho])^2]} < \epsilon \sqrt{\sum_{P \in \mathcal{S}} \frac{a_P^2}{\beta(P)}}. \quad (\text{B2})$$

- (2) There exists a probabilistic measurement protocol which uses $2\Lambda^2 \log(2/\delta)/\epsilon^2$ measurements on different copies of an unknown state ρ , to output an estimate $\hat{o}$ satisfying

$$|\hat{o} - \text{Tr}[O\rho]| < \epsilon \cdot \max_{P \in \mathcal{S}} \frac{|a_P|}{\beta(P)}, \quad (\text{B3})$$

with probability at least $1 - \delta$.

Proof. Given $N_s \geq 1$, for $i = 1, 2, \dots, N_s$, we perform the following steps:

- (1) Sample an operator $P_i \in \mathcal{S}$ with probability $\beta(P_i)$.
- (2) Measure P_i on (an unused copy of) the state ρ .
- (3) Denote the measurement outcome as $\hat{x}_i$.
- (4) Define the random variable

$$\hat{X}_i = \frac{a_{P_i}}{\beta(P_i)} \hat{x}_i. \quad (\text{B4})$$

We notice that $\hat{X}_i$ is an unbiased estimator for $\text{Tr}[\rho O]$,

$$\mathbb{E}\hat{X}_i = \sum_{P \in \mathcal{S}} \beta(P) \frac{a_P}{\beta(P)} \text{Tr}[P\rho] = \text{Tr}[O\rho]. \quad (\text{B5})$$

We also note that

$$|\hat{X}_i| \leq \Lambda \max_{P \in \mathcal{S}} \frac{|a_P|}{\beta(P)}, \quad (\text{B6})$$

and therefore $\hat{X}_i$ takes value in an interval of size at most $2\Lambda \max_{P \in \mathcal{S}} \{|a_P|/\beta(P)\}$. Let $\hat{o}$ be the mean of $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_{N_s}$. By Lemma A2 (Hoeffding's inequality), we have that

$$\Pr \left[|\hat{o} - \text{Tr}[O\rho]| < \epsilon \cdot \Lambda \max_{P \in \mathcal{S}} \frac{|a_P|}{\beta(P)} \right] \geq 1 - \delta, \quad (\text{B7})$$

provided that $N_s \geq 2 \log(2/\delta)/\epsilon^2$ which proves the second statement of the lemma.

Moreover, we have

$$\mathbb{E}[(\hat{o} - \text{Tr}[O\rho])^2] = \text{Var}[\hat{o}] \quad (\text{B8})$$

$$= \frac{1}{N_s^2} \sum_{i=1}^{N_s} \text{Var}[\hat{X}_i] \quad (\text{B9})$$

$$\leq \frac{1}{N_s^2} \sum_{i=1}^{N_s} \mathbb{E}[\hat{X}_i^2] \quad (\text{B10})$$

$$= \sum_{P \in \mathcal{S}} \frac{a_P^2}{\beta(P)} \frac{1}{N_s^2} \sum_{i=1}^{N_s} \mathbb{E}[\hat{x}_i^2] \quad (\text{B11})$$

$$\leq \sum_{P \in \mathcal{S}} \frac{a_P^2 \Lambda^2}{\beta(P) N_s}, \quad (\text{B12})$$

where the last inequality is due to $|\hat{x}_i| \leq \Lambda$. This proves the first statement of the lemma by setting $N_s = \Lambda^2/\epsilon^2$. ■

For any given values of the coefficients a_P , one can analytically compute the distribution $\beta(P)$ minimizing the function

$$\max_{P \in \mathcal{S}} \frac{|a_P|}{\beta(P)} \quad (\text{B13})$$

and run the corresponding measurement protocol, as we demonstrate in the following appendix. Notice that we

assumed $\beta(P) > 0$ for $a_P \neq 0$. This condition will naturally be met with the following settings for β .

1. Warm-up: Estimating a single observable

The protocol presented in Lemma B1 is parameterized by a distribution β , which defines the probability of measuring each observable $P \in \mathcal{S}$ in a single round of the protocol. After $N_s = \Omega(\log(1/\delta)/\epsilon^2)$ rounds, each observable $P \in \mathcal{S}$ is measured $\beta(P) \times N_s$ times, in expectation. Intuitively, more measurement shots should be allocated to the observables P that have larger coefficients a_P (in absolute value). The following corollary formalizes this intuition.

Corollary B1. Let $O = \sum_{P \in \mathcal{S}} a_P P$ be an observable. There exists a randomized measurement protocol which uses $2 \|\mathbf{a}\|_1^2 \log(2/\delta) \Lambda^2/\epsilon^2$ measurements on different copies of an unknown state ρ , which outputs an estimate $\hat{o}$ satisfying

$$|\hat{o} - \text{Tr}[O\rho]| < \epsilon. \quad (\text{B14})$$

Proof. The corollary follows from Lemma B1 by setting $\beta(\cdot)$ as

$$\beta(P) = \frac{|a_P|}{\sum_{P' \in \mathcal{S}} |a_{P'}|} := \frac{|a_P|}{\|\mathbf{a}\|_1}. \quad (\text{B15})$$

■

This shot allocation strategy is not novel and has been explored in a number of contexts including the variational quantum algorithms [110,111]. Crucially, this choice of β minimizes the loss function $\mathcal{L}(\beta) = \max_P |a_P|/\beta(P)$, which can be verified using Lagrange multipliers (see Appendix B3). One might wonder whether a simpler protocol, such as allocating an equal probability to all operators in $\mathcal{S}$, could still achieve comparable performance. However, a uniform allocation of measurement shots becomes highly suboptimal when the observable O is concentrated around a few elements of $\mathcal{S}$. In such cases, a uniform strategy would result in an exponentially larger sample complexity, as we demonstrate in the following discussion.

a. Comparison with uniform allocation

One of the simplest choices for $\beta(\cdot)$ is the uniform distribution over the set $\mathcal{S}$, that is,

$$\beta(P) = \begin{cases} \frac{1}{|\mathcal{S}|} & \text{if } P \in \mathcal{S}, \\ 0 & \text{otherwise,} \end{cases} \quad (\text{B16})$$

which corresponds to measuring each observable (in $\mathcal{S}$) $N_s/|\mathcal{S}|$ times in expectation. By Lemma B1, we have

that $N_s = 2 \log(2/\delta)/\epsilon^2$ measurement shots are sufficient to ensure

$$|\hat{o}_{\text{uniform}} - \text{Tr}[O\rho]| \leq \epsilon |\mathcal{S}| \max_P |c_P| < \epsilon \Lambda |\mathcal{S}|, \quad (\text{B17})$$

with probability at least $1 - \delta$. Setting $\epsilon = \tilde{\epsilon}/\Lambda |\mathcal{S}|$, we obtain that $N_s = 2 |\mathcal{S}|^2 \Lambda^2 \log(2/\delta)/\tilde{\epsilon}^2$ many measurements are required to achieve an error $\tilde{\epsilon}$ with probability at least $1 - \delta$. To demonstrate the limitations of this measurement protocol, consider the following extreme example of an observable:

$$O_{\text{peaked}} = Z^{\otimes n} + \tau \sum_{P \in \{I, X, Y, Z\}^{\otimes n} \setminus \{Z^{\otimes n}\}} \frac{P}{4^n - 1},$$

for some constant $\tau = \Theta(1)$. In the above expression, the observable is dominated by a single Pauli term $Z^{\otimes n}$, while all other Pauli strings contribute only marginally with coefficients scaled by an exponentially small factor $\tau/(4^n - 1)$.

In this case $|\mathcal{S}| = 4^n$ and $\Lambda = 1$ (i.e., $\mathcal{S}$ is the set of n qubits Pauli), and thus the samples complexity for the uniform allocation scales as $\mathcal{O}(16^n \log(1/\delta)/\tilde{\epsilon}^2)$.

On the other hand, we have that

$$\|\mathbf{a}\|_1 = 1 + \tau \in \Theta(1), \quad (\text{B18})$$

and therefore the sample complexity of the protocol described in Corollary B1 is $\mathcal{O}(\log(1/\delta)/\tilde{\epsilon}^2)$ for achieving error $\tilde{\epsilon}$ and success probability $1 - \delta$. Crucially, the uniform allocation requires exponentially more measurements.

2. Estimating parametrized observables

Throughout this paper, we often encounter *parametrized* observables $O(\alpha) = \sum_{P \in \mathcal{S}} c_P(\alpha) P$, where α is sampled from a given distribution $\mathcal{D}$. Here, we aim at estimating a function $\hat{o}(\alpha)$ minimizing the following *mean square error*:

$$\mathbb{E}_{\alpha \sim \mathcal{D}} (\text{Tr}[O(\alpha)\rho] - \hat{o}(\alpha))^2. \quad (\text{B19})$$

In such case, designing an optimal measurement protocol is more challenging, since the vector $\mathbf{c}(\alpha)$ is now a random variable. Consequently, the measurement distribution $\beta(\cdot)$ should depend on the distribution $\mathcal{D}$.

Broadly speaking, we show that, while the sample complexity for a single observable (i.e., a fixed value vector α_0) is dominated by the 1-norm $\|\mathbf{c}(\alpha_0)\|_1$, the sample complexity for the multiple observables cases can be described via a suitable “effective 1-norm,” which we define in the following.

Definition B1 (Effective 1-norms). Let α be a parameter vector, and let $O(\alpha) = \sum_{P \in \mathcal{S}} c_P(\alpha) P$ be a parametrized observable.

- (a) Let us assume that α is sampled from a distribution $\mathcal{D}$. Then the associated average-case effective 1-norm is defined as

$$\|\mathbf{c}\|_{1,\text{avg}} := \sum_{P \in \mathcal{S}} \sqrt{\mathbb{E}_{\alpha \sim \mathcal{D}} [c_P(\alpha)^2]}. \quad (\text{B20})$$

- (b) Let us assume that α lies in a set $\mathcal{A}$. Then the associated *worst-case effective 1-norm* is defined as

$$\|\mathbf{c}\|_{1,\text{worst}} := \sum_{P \in \mathcal{S}} \max_{\alpha \in \mathcal{A}} |c_P(\alpha)|. \quad (\text{B21})$$

Lemma B2 (Mean square error). Let us assume that α is sampled from a distribution $\mathcal{D}$. There exists a randomized measurement protocol which uses $N_s = \Lambda^2 \|\mathbf{c}\|_{1,\text{avg}}^2 / \epsilon^2$ measurements on different copies of a state ρ , and outputs an estimate $\hat{o}(\alpha)$ for each α satisfying

$$\sqrt{\mathbb{E}_{\alpha \sim \mathcal{D}, \mathcal{M}} (\text{Tr}[O(\alpha)\rho] - \hat{o}(\alpha))^2} \leq \epsilon, \quad (\text{B22})$$

where $\mathbb{E}_{\mathcal{M}}$ denotes an average over the measurement outcomes.

Proof. Let $\beta(\cdot)$ be defined as

$$\beta(P) = \frac{\sqrt{\mathbb{E}_{\alpha \sim \mathcal{D}} [c_P(\alpha)^2]}}{\|\mathbf{c}\|_{1,\text{avg}}}. \quad (\text{B23})$$

The measurement protocol here is similar to Lemma B1, but in the fourth step, we use the outcome of the third step, $\hat{x}_i$ to define the random variable

$$\hat{X}_i(\alpha) = \frac{c_{P_i}(\alpha)}{\beta(P_i)} \hat{x}_i, \quad (\text{B24})$$

for each α . Recall that $\hat{X}_i(\alpha)$ is an unbiased estimator of $\text{Tr}[\rho O(\alpha)]$. Let

$$\hat{o}(\alpha) = \frac{1}{N_s} \sum_{i=1}^{N_s} \hat{X}_i(\alpha). \quad (\text{B25})$$

We have also seen in Eq. (B12) that

$$\mathbb{E}_{\mathcal{M}} [(\hat{o}(\alpha) - \text{Tr}[\rho O(\alpha)])^2] \leq \sum_{P \in \mathcal{S}} \frac{c_P(\alpha)^2 \Lambda^2}{\beta(P) N_s}, \quad (\text{B26})$$

which is true for any α . Therefore, averaging over $\alpha \sim \mathcal{D}$ and inserting the definition $\beta(P)$ leads to the desired

result, i.e.,

$$\mathbb{E}_{\alpha \sim \mathcal{D}, \mathcal{M}}[(\hat{\alpha}(\alpha) - \text{Tr}[\rho O(\alpha)])^2] \leq \sum_{P \in \mathcal{S}} \frac{\mathbb{E}_{\alpha \sim \mathcal{D}}[c_P(\alpha)^2] \Lambda^2}{\beta(P) N_s} \quad (\text{B27})$$

$$= \frac{\|c\|_{1,\text{avg}}^2 \Lambda^2}{N_s}. \quad (\text{B28})$$

Finally, we complete the proof by setting $N_s = \Lambda^2 \|c\|_{1,\text{avg}}^2 / \epsilon^2$. ■

Lemma B3 (Maximum error). Let us assume that α lies in the set $\mathcal{A}$. There exists a randomized measurement protocol which uses $N_s = 2 \log(2/\delta) \Lambda^2 \|c\|_{1,\text{worst}}^2 / \epsilon^2$ measurements on different copies of a state ρ , and output an estimate $\hat{\alpha}(\alpha)$ satisfying

$$|\text{Tr}[O(\alpha)\rho] - \hat{\alpha}(\alpha)| < \epsilon, \quad (\text{B29})$$

with probability $1 - \delta$ (over the internal randomness of the measurement protocol).

Proof. The measurement protocol is similar to Lemma B1 with the sampling distribution $\beta(\cdot)$ defined as

$$\beta(P) = \frac{\max_{\alpha \in \mathcal{A}} |c_P(\alpha)|}{\|c\|_{1,\text{worst}}}. \quad (\text{B30})$$

Then, for any fixed $\alpha \in \mathcal{A}$ we can use Hoeffding's inequality (i.e., Lemma A2) as in Lemma B1 to get

$$|\text{Tr}[O(\alpha)\rho] - \hat{\alpha}(\alpha)| < \tilde{\epsilon} \cdot \Lambda \max_{P \in \mathcal{S}} \frac{|c_P(\alpha)|}{\beta(P)}, \quad (\text{B31})$$

with probability at least $1 - \delta$ using $N_s = 2 \log(2/\delta) / \tilde{\epsilon}^2$. From previous definition of $\beta(P)$, we have $\max_P |c_P(\alpha)| / \beta(P) \leq \|c\|_{1,\text{worst}}$. Thus, substituting $\tilde{\epsilon} = \epsilon / \|c\|_{1,\text{worst}} \Lambda$ completes the proof. ■

Notice that the sampling distribution, $\beta(\cdot)$ used in Lemma B3 leads to the same bound on the MSE as in Lemma B2, but with the average-case effective 1-norm being replaced by the worst-case one.

3. Optimal sampling distributions

In order to find the distribution $\beta(P)$ that minimize the MSE in Eq. (B26), we used Lagrange multipliers method

(see Appendix A 5 a) to minimize the function

$$\sum_{P \in \mathcal{S}} \frac{\mathbb{E}_{\alpha} [c_P^2(\alpha)]}{\beta(P)}, \quad (\text{B32})$$

under the constraint $\sum_{P \in \mathcal{S}} \beta(P) = 1$. The appropriate Lagrangian is given by

$$\mathcal{L} = \sum_{P \in \mathcal{S}} \frac{\mathbb{E}_{\alpha} [c_P^2(\alpha)]}{\beta(P)} + \lambda \left(\sum_{P \in \mathcal{S}} \beta(P) - 1 \right), \quad (\text{B33})$$

where λ is the Lagrangian multiplier. Next we compute the derivatives with respect to $\beta(P)$ and equate them to zero. The previous leads to

$$\frac{\partial \mathcal{L}}{\partial \beta(P)} = -\frac{\mathbb{E}_{\alpha} [c_P^2(\alpha)]}{\beta(P)^2} + \lambda = 0. \quad (\text{B34})$$

Then, the solution to the system of equations $\{\partial_{\beta(P)} \mathcal{L} = 0\}_{P \in \mathcal{S}}$ with the normalization constraint is

$$\beta(P) = \frac{\sqrt{\mathbb{E}_{\alpha} [c_P^2(\alpha)]}}{\sum_{P \in \mathcal{S}} \sqrt{\mathbb{E}_{\alpha} [c_P^2(\alpha)]}} = \frac{\sqrt{\mathbb{E}_{\alpha} [c_P^2(\alpha)]}}{\|c\|_{1,\text{avg}}}. \quad (\text{B35})$$

The derivation of the optimal distribution $\beta(P)$ in Lemma B3 is similar to Lemma B2, but here we need to minimize Eq. (B31) for any choice of $\alpha \in \mathcal{A}$ which is equivalent to minimizing the function

$$\max_P \max_{\alpha} \frac{|c_P(\alpha)|}{\beta(P)}. \quad (\text{B36})$$

Let us define the vector $\mathbf{v}$ with entries $v_P = \max_{\alpha} |c_P(\alpha)| / \beta(P)$. So, we want to minimize $\|\mathbf{v}\|_{\infty}$ under the constraint $\sum_{P \in \mathcal{S}} \beta(P) = 1$. Let us consider the k -norm of $\mathbf{v}$ and then we will take the limit $k \rightarrow \infty$. The Lagrangian is given by

$$\mathcal{L} = \left(\sum_{P \in \mathcal{S}} v_P^k \right)^{\frac{1}{k}} + \lambda \left(\sum_{P \in \mathcal{S}} \beta(P) - 1 \right), \quad (\text{B37})$$

which is minimized if, for every $\beta(P)$, we have

$$-\frac{(\max_{\alpha} |c_P(\alpha)|)^k}{\beta(P)^{k+1}} \left(\sum_{P \in \mathcal{S}} v_P^k \right)^{\frac{1-k}{k}} + \lambda = 0. \quad (\text{B38})$$

The solution is given by

$$\beta(P) = \frac{(\max_{\alpha} |c_P(\alpha)|)^{\frac{k}{k+1}}}{\sum_{P' \in \mathcal{S}} (\max_{\alpha} |c_{P'}(\alpha)|)^{\frac{k}{k+1}}}. \quad (\text{B39})$$

Finally, by taking the limit $k \rightarrow \infty$, we end up with

$$\beta(P) = \frac{\max_{\alpha} |c_P(\alpha)|}{\sum_{P' \in \mathcal{S}} \max_{\alpha} |c_{P'}(\alpha)|} = \frac{\max_{\alpha} |c_P(\alpha)|}{\|c\|_{1,\text{worst}}}. \quad (\text{B40})$$

APPENDIX C: GENERAL SURROGATABILITY GUARANTEE

1. Summary of the key results

In this appendix we will start by summarizing the main results of which our general surrogatability guarantee is comprised. Their proofs are then provided in subsequent sections of this appendix.

a. Taylor surrogate construction

We recall that we are interested in expectation functions of the general form

$$f(\alpha) = \text{Tr}[\mathcal{E}_\alpha(\rho)O], \quad (\text{C1})$$

where $\mathcal{E}_\alpha(\cdot)$ is an arbitrary parameterized quantum channel with m variational parameters α acting on a n – qubit quantum state ρ and O is some Hermitian operator. We are interested in constructing the surrogate $\tilde{f}_\kappa(\alpha)$ in the hypercube region $\mathcal{V}_m(\alpha^*, r)$, i.e.,

$$\mathcal{V}_m(\alpha^*, r) := \{\alpha\} \quad \text{such that } \alpha_i \in [\alpha_i^* - r, \alpha_i^* + r] \quad \forall i. \quad (\text{C2})$$

This expectation function can be expressed using a multivariate Taylor's expansion as

$$f(\alpha) = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i_1, i_2, \dots, i_k=1}^m \partial_{i_1, \dots, i_k} f(\alpha^*) \Delta_{i_1, \dots, i_k}(\alpha, \alpha^*), \quad (\text{C3})$$

where we introduce two compressed notations for ease of notation

$$\partial_{i_1, \dots, i_k} f(\alpha^*) = \left(\frac{\partial^k f(\alpha)}{\partial \alpha_{i_1} \partial \alpha_{i_2} \dots \partial \alpha_{i_k}} \right) \Big|_{\alpha=\alpha^*}, \quad (\text{C4})$$

$$\Delta_{i_1, \dots, i_k}(\alpha, \alpha^*) = (\alpha_{i_1} - \alpha_{i_1}^*)(\alpha_{i_2} - \alpha_{i_2}^*) \dots (\alpha_{i_k} - \alpha_{i_k}^*). \quad (\text{C5})$$

The κ th-order surrogate of the expectation function is made by simply truncating the infinite sum in Eq. (C3) to order κ

$$\tilde{f}_\kappa(\alpha) = \sum_{k=0}^{\kappa} \frac{1}{k!} \sum_{i_1, i_2, \dots, i_k=1}^m \partial_{i_1, \dots, i_k} f(\alpha^*) \Delta_{i_1, \dots, i_k}(\alpha, \alpha^*). \quad (\text{C6})$$

In order to construct the surrogate, we are required to obtain all partial derivatives of the expectation function up to order κ . These partial derivatives can be obtained by evaluating the expectation function $f(\alpha)$ at different points on the landscape.

The following three key assumptions are made for the surrogatability guarantee in our proofs.

Assumption C1. The partial derivatives of $f(\alpha)$ within the region are bounded as

$$|\partial_{i_1, i_2, \dots, i_k} f(\alpha)| \leq \gamma^k \|O\|_\infty, \quad \forall \alpha \in \mathcal{V}_m(\alpha^*, r), \quad (\text{C7})$$

with some constant $\gamma \in \mathcal{O}(1)$.

Assumption C2. Computing each partial derivative (up to the order κ) requires at most N_d evaluations of the expectation function which scales as

$$N_d \in \mathcal{O}(m^\kappa). \quad (\text{C8})$$

Assumption C3. The partial derivative can be expressed as a linear combination of expectation functions evaluated at different parameters

$$\partial_{i_1, i_2, \dots, i_k} f(\alpha) = \sum_{j=1}^{N_d} b_j f(\alpha_j), \quad (\text{C9})$$

where each b_j is an associated coefficient to $f(\alpha_j)$ such that $b_0 = \max_j |b_j|$ with some positive $b_0 \in \mathcal{O}(\text{poly}(m))$, and can be efficiently computed classically. Furthermore, a set of $\{\alpha_j\}_{j=1}^{N_d}$ depends on α and can be efficiently determined classically. Note that for each combination of the derivatives $\partial_{i_1, \dots, i_k}$ the dependence on $\{b_j\}_{j=1}^{N_d}$ and $\{\alpha_j\}_{j=1}^{N_d}$ can be different. However, we need the condition $b_0 = \max_j |b_j|$ to hold for all the derivatives.

We note that two well-known approaches to compute gradients that satisfy Assumptions C2 and C3 are parameter shift-rule and numerical finite difference.

In what follows, we present a series of formal statements with the ultimate aim of providing the surrogatability guarantee over the region. That is, Supplement Theorem C1 is presented as a formal version of Theorem 1.

b. Worst-case error

The worst-case error of the κ th-order surrogate in the region $\mathcal{V}_m(\alpha^*, r)$ around the parameters α^* is defined as

$$\epsilon_{\kappa, \alpha^*}^{\text{WC}} := \max_{\alpha \in \mathcal{V}_m(\alpha^*, r)} |\tilde{f}_\kappa(\alpha) - f(\alpha)|. \quad (\text{C10})$$

The following proposition (proved below) tells how the worst-case error scales with respect to the surrogate order.

Proposition C1 (Worst-case error of the surrogate). Consider an expectation function $f(\alpha)$ as defined in Eq. (C1) with $\|O\|_\infty \in \mathcal{O}(1)$. Furthermore, consider the

hypercube region $\mathcal{V}_m(\alpha^*, r)$ around some initial fixed point α^* with the perturbation r as defined in Eq. (C2) such that

$$r \in \mathcal{O}\left(\frac{1}{m}\right). \quad (\text{C11})$$

Assume that all the derivatives (up to order $k = \kappa + 1$) are bounded as

$$|\partial_{i_1, i_2, \dots, i_k} f(\alpha)| \leq \gamma^k \|O\|_\infty, \quad \forall \alpha \in \mathcal{V}_m(\alpha^*, r), \quad (\text{C12})$$

with $\gamma \in \mathcal{O}(1)$ (i.e., Assumption C1).

Then, we have that $\tilde{f}_\kappa(\alpha)$ according to Eq. (C6), with the κ order of the truncation, is a classical surrogate of $f(\alpha)$ over this regime with a worst-case error bounded by

$$\epsilon_{\kappa, \alpha^*}^{\text{WC}} \leq \frac{\|O\|_\infty}{(\kappa + 1)!} (\gamma r m)^{\kappa+1}. \quad (\text{C13})$$

In particular, an arbitrarily small but constant error $\epsilon_{\kappa, \alpha^*}^{\text{WC}} \in \mathcal{O}(1)$ can be achieved with $\kappa \in \mathcal{O}(1)$.

c. Mean square error

The mean square error (MSE) of the κ th-order surrogate associated with an uniform distribution $\mathcal{D}_m(\alpha^*, r)$ over the hypercube region $\mathcal{V}_m(\alpha^*, r)$ is defined as

$$(\epsilon_{\kappa, \alpha^*}^{\text{MSE}})^2 = \mathbb{E}_{\alpha \sim \mathcal{D}_m(\alpha^*, r)} \left[(\tilde{f}_\kappa(\alpha) - f(\alpha))^2 \right]. \quad (\text{C14})$$

The next proposition (proved below) analytically shows the MSE scaling with respect to the surrogate order.

Proposition C2 (Mean Square Error of the surrogate). Consider an expectation function $f(\alpha)$ as defined in Eq. (C1) with $\|O\|_\infty \in \mathcal{O}(1)$. Furthermore, consider the hypercube region $\mathcal{V}_m(\alpha^*, r)$ around some initial fixed point α^* with the perturbation r as defined in Eq. (C2) such that

$$r \in \mathcal{O}(1/\sqrt{m}). \quad (\text{C15})$$

Assume that all the derivatives (of order $k > \kappa$) are bounded as (i.e., Assumption C1)

$$|\partial_{i_1, i_2, \dots, i_k} f(\alpha)| \leq \gamma^k \|O\|_\infty, \quad \forall \alpha \in \mathcal{V}_m(\alpha^*, r), \quad (\text{C16})$$

with $\gamma \in \mathcal{O}(1)$.

Then, we have that $\tilde{f}_\kappa(\alpha)$ according to Eq. (C6), with the κ order of the truncation, is a classical surrogate of $f(\alpha)$ over this regime with the MSE associated with the uniform distribution $\mathcal{D}_m(\alpha^*, r)$ bounded by

$$\epsilon_{\kappa, \alpha^*}^{\text{MSE}} \leq \left(\frac{2\gamma^2 m r^2}{3} \right)^{\frac{\kappa+1}{2}} \frac{\|O\|_\infty}{\sqrt{(\kappa + 1)!}} e^{\frac{\gamma^2 m r^2}{3}}. \quad (\text{C17})$$

In particular, for an arbitrarily small but constant error $\epsilon_{\kappa, \alpha^*}^{\text{MSE}} \in \mathcal{O}(1)$ can be achieved for $\kappa \in \mathcal{O}(1)$.

d. Scaling of the number of expectation function evaluations with truncation order

The classical surrogate can be constructed from the partial derivatives at α^* according to Eq. (C6). These partial derivatives $\partial_{i_1, \dots, i_k} f(\alpha^*)$ can be computed by evaluating the expectation function $f(\alpha)$ at different parameter values. Let N denote the total number of expectation function evaluations to determine the partial derivatives required to construct the κ th-order surrogate. The following proposition bounds N in terms of the surrogate order.

Proposition C3 (Scaling of the number of terms). The classical surrogate can be constructed by evaluating the expectation functions at N parameter settings, where

$$N \leq N_d \left[\frac{e(m + \kappa - 1)}{\kappa} \right]^\kappa, \quad (\text{C18})$$

with N_d an upper bound on the number of expectation function evaluations required to compute each partial derivative (up to $\kappa + 1$ order).

The way we evaluate the expectation function to compute the partial derivatives has not been specified yet. In some certain cases, the expectation function evaluations might be computable classically. However, in general we will need access to quantum computers to estimate expectation function values and in turn the partial derivatives. Since only statistical estimates of the expectation function can be obtained from quantum computers through measurement shots, this leads to an additional error for the surrogate due to the statistical fluctuation.

e. General surrogation guarantee

The following Supplemental Theorem presents the final surrogatability guarantee taking into account the presence of shot noise. Here we assume that the expectation function can be estimated on a quantum computer by measuring in the eigenbasis of the observable O . That is, given an eigendecomposition $O = \sum_j \lambda_j |\lambda_j\rangle\langle\lambda_j|$ where λ_j is an eigenvalue with an associated eigenstate $|\lambda_j\rangle$, an unbiased estimator of the expectation function after N_s measurement shots can be expressed as

$$\hat{f}(\alpha) = \frac{1}{N_s} \sum_{i=1}^{N_s} \hat{o}_i, \quad (\text{C19})$$

where $\hat{o}_i$ is an outcome of i th measurement which takes a value λ_j with a probability $p_j = \text{Tr}[\mathcal{E}_\alpha(\rho) |\lambda_j\rangle\langle\lambda_j|]$. For observables of complex many-body systems it may not be straightforward to measure in the eigenbasis of O . In this case, it is instead natural to expand O as a sum of “easier” operators to measure (e.g., a sum of Paulis). While we do

not explicitly present this case here, the generalization is straightforward.

Supplemental Theorem C1 (General Surrogatability Guarantee, Formal). Consider an expectation function $f(\alpha)$ as defined in Eq. (C1) with $\|O\|_\infty \in \mathcal{O}(1)$. Assume that the derivatives are bounded as

$$|\partial_{i_1, i_2, \dots, i_k} f(\alpha)| \leq \|O\|_\infty \gamma^k, \quad \forall \alpha \in \mathcal{V}_m(\alpha^*, r), \quad (\text{C20})$$

with $\gamma \in \mathcal{O}(1)$ (i.e., Assumption C1). We additionally assume that each partial derivative (up to the order κ) can be written as a linear combination of at most N_d expectation function evaluations such that

$$N_d \in \mathcal{O}(m^\kappa) \quad (\text{C21})$$

(i.e., Assumption C2) and where the biggest coefficient in absolute value scales at most as $b_0 \in \mathcal{O}(\text{poly}(m))$ (i.e., Assumption C3). Moreover, we assume that each of these expectation function evaluations can be estimated by measuring in the eigenbasis on a quantum computer as in Eq. (C19).

(a) Worst-case error: when the perturbation r scales as

$$r \in \mathcal{O}\left(\frac{1}{m}\right), \quad (\text{C22})$$

with a number of measurement shots N_s and probability at least $1 - \delta$ (for any $\delta \in \Theta(1)$) we can construct $\widehat{f}_\kappa(\alpha)$ such that

$$\begin{aligned} & \left| \widehat{f}_\kappa(\alpha) - f(\alpha) \right| \\ & \leq \|O\|_\infty \left(b_0 N_d e^{rm} \sqrt{\frac{2 \log(2/\delta)}{N_s}} + \frac{(\gamma r m)^{\kappa+1}}{(\kappa+1)!} \right). \end{aligned} \quad (\text{C23})$$

Thus the worst-case error can be reduced to an arbitrarily small constant, $\left| \widehat{f}_\kappa(\alpha) - f(\alpha) \right| \in \Theta(1)$ by increasing the order κ with $\kappa \in \mathcal{O}(1)$ and the number of shots with $N_s \in \mathcal{O}(\text{poly}(m))$.

(b) Average-case error: when the perturbation r scales as

$$r \in \mathcal{O}\left(\frac{1}{\sqrt{m}}\right), \quad (\text{C24})$$

with N_s measurement shots, we can construct $\widehat{f}_\kappa(\alpha)$ such that,

$$\begin{aligned} & \mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\widehat{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] \\ & \leq 2 \left(\frac{b_0 N_d (\kappa+1) (rm)^\kappa \|O\|_\infty}{\sqrt{N_s} \kappa!} \right)^2 \\ & \quad + 2 \left(\frac{2\gamma^2 m r^2}{3} \right)^{\kappa+1} \frac{\|O\|_\infty^2}{(\kappa+1)!} e^{\frac{2\gamma^2 m r^2}{3}}. \end{aligned} \quad (\text{C25})$$

Thus the mean square error can be arbitrarily reduced to any constant in $\Theta(1)$ by increasing the order κ with $\kappa \in \mathcal{O}(1)$ and the number of shots N_s with $N_s \in \mathcal{O}(\text{poly}(m))$. Then by applying Markov's inequality Lemma A1 we find the informal version of the theorem.

Supplemental Theorem C1 is the formal version of Theorem 1. Indeed, we can set both the truncation and the empirical mean square errors to be $(\epsilon/2)^2$. We have seen that with $r \in \mathcal{O}(1/\sqrt{m})$, the mean square truncation error can be made arbitrarily small by increasing $\kappa \in \mathcal{O}(1)$. Then, equating both errors from the expression in Eq. (C25) leads to

$$N_s \in \mathcal{O}(m^\kappa b_0^2 N_d^2) = \mathcal{O}(\text{poly}(m)), \quad (\text{C26})$$

where we used Assumptions C2 and C3 for the scaling of both b_0 and N_d . Therefore, using polynomial resources, we can get an arbitrarily small mean square error. A similar analysis can be made the worst-case error guarantee with $r \in \mathcal{O}(1/m)$ to get

$$N_s \in \mathcal{O}(N_d^2 b_0^2 \log(1/\delta)) = \mathcal{O}(\text{poly}(m)). \quad (\text{C27})$$

We note that in the proof of Supplemental Theorem C1, the surrogate is expressed as the linear combination of the expectation function evaluated at different parameters. In order to construct a statistical estimate of the surrogate, we individually estimate each term in the sum and also the total measurement shout budget is evenly distributed to all the terms. This strategy is suboptimal in general since it could be the case that there are terms in the sum that commute with each other and hence they can be simultaneously estimated and/or the total shot budget could be distributed in a more informative manner. Nevertheless, we now show that even with this suboptimal strategy one can achieve a surrogate that achieves an arbitrarily small constant error with high probability with a polynomial number of measurement shots.

Note: In principle, the analysis can be generalized to allow for independent perturbation radii for each parameter. The intuition is simple: in the derivation of the Taylor

remainder in Eq. (C63), the dominant contribution arises from the variance of the sum of the parameter variations, suggesting that the expression mr^2 could be replaced by $\sum_{i=1}^m r_i^2$ where $\delta_i \in [-r_i, r_i]$. Nevertheless, establishing this generalization with full rigor would require considerably more technical work and would add little conceptual value to the main results. For this reason, and to keep the presentation focused, we restrict the analysis to hypercubes, while noting that the extension follows the same intuition.

f. Guarantee that our assumptions are met by common parameterized quantum circuits

The following proposition establishes that a broad family of parameterized quantum circuits (indeed, essentially all currently considered by the community currently with independent parameters between circuit layers) satisfy the assumptions of our bounds.

Proposition C4 (Bounded derivatives for expectation value with parametrized unitary channel). Let us assume a expectation function of the form of Eq. (1) i.e., $f(\alpha) = \text{Tr}[\rho U^\dagger(\alpha) O U(\alpha)]$, where $U(\alpha) = \prod_{l=1}^m V_l U_l(\alpha_l)$ such that V_l are arbitrary nonparametrized unitaries and each $U_l(\alpha_l) = \exp(-iH_l \alpha_l)$ is a parametrized gate with generator H_l that is hermitian. Then, the k th-order derivatives of the form $\partial^k f(\alpha) / \prod_{l=1}^m \partial \alpha_l^{k_l} := \partial^k f(\alpha)$ where $\sum_{l=1}^m k_l = k$ and $k_l \in \mathbb{N}$ for each l are bounded by

$$|\partial^k f(\alpha)| \leq \|O\|_\infty \prod_{l=1}^m (2 \|H_l\|_\infty)^{k_l} \quad (\text{C28})$$

$$\leq \|O\|_\infty \left(2 \max_l \|H_l\|_\infty \right)^k \quad (\text{C29})$$

Corollary C1 (Surrogation Guarantee for Unitary Channel). Consider a generic expectation function $f(\alpha)$ as defined in Eq. (1) with m independent parameters such that the spectral norm of each generators as well as the initial observable one are $\mathcal{O}(1)$, and assume that (up to) order- κ partial derivatives can be estimated efficiently with quantum computers. It is possible to efficiently classically surrogate $f(\alpha)$ over any uniformly sampled hypercube $\mathcal{D}_m(\alpha^*, r)$ around an arbitrary point α^* , with

$$r \in \mathcal{O}\left(\frac{1}{\sqrt{m}}\right). \quad (\text{C30})$$

In our general surrogate guarantee, Assumptions C2 and C3 can be argued from a practical perspective. Let us clarify the scope of this theorem by first considering the following simple example. Assume a circuit composed by uncorrelated Pauli rotations and any nonparametrized gates (not restricted to near-Clifford here). In this case, the

parameter-shift rule, which satisfies both Assumptions C2 and C3, can be readily applied. Indeed, any order of derivatives with respect to a single parameters involves two loss evaluations with coefficients magnitude $1/2$. From this, we can directly state that $b_0 \in \mathcal{O}(1)$. For k th-order partial derivatives, the largest number of loss evaluations correspond to derivatives with respect to k distinct parameters which leads to up to 2^k loss evaluations in total (with coefficients magnitude $1/2^k$) [112]. This means that truncating at order κ involves at most $N_d \in \mathcal{O}(2^\kappa)$ loss evaluation for each partial derivatives considered in the surrogate. Therefore, both assumptions are satisfied. For this case, we provide a numerical study in Appendix D where we use Taylor surrogate iteratively to approximate a ground state.

The standard parameter-shift rule applies to gates whose generators have a two distinct eigenvalues, including single-qubit rotations and multi-qubit Pauli rotations. More generally, we can use the generalized parameter-shift rule [91] to compute derivatives of generators with more than two distinct eigenvalues. One example are Givens rotations whose generators are sums of commuting Pauli operators. These require four circuit evaluations.

g. How the polynomial guarantees break down for larger patch sizes

For completeness, we note that if we want the resource requirements to scale at most polynomially with the system size and number of variational parameters, the only region for which we [113] could prove generic guarantees for is the aforementioned $r \in \mathcal{O}(1/\sqrt{m})$. Indeed, this ensures that the terms we require to compute and the MSE ϵ remain constant.

We can see this as follows. We start by recalling that the number of terms we need to compute, given by Proposition C3, scales as

$$N \in \mathcal{O}(m^{w\kappa}), \quad (\text{C31})$$

where κ is the truncation order and $w \in [1, 2]$ is some constant, and the MSE given in Proposition C2 is proportional to

$$\epsilon^2 \leq \frac{(cmr^2)^{\kappa+1}}{(\kappa+1)!} e^{cmr^2}, \quad (\text{C32})$$

where $c = 2\gamma^2/3 \in \mathcal{O}(1)$ is a constant. Next we will proceed to show what happens when we want to extend beyond the $r \in \mathcal{O}(1/\sqrt{m})$.

Let us first simplify the MSE expression using Stirling approximation for the factorial

$$k! \geq \sqrt{2\pi k} (k/e)^k \geq (k/e)^k. \quad (\text{C33})$$

Therefore, the MSE ϵ can be bounded as follows:

$$\epsilon^2 \leq \left(\frac{cemr^2}{\kappa + 1} \right)^{\kappa+1} e^{cmr^2}. \quad (C34)$$

From the previous equation, we see that in order to make the error bound arbitrarily small, we would necessarily need $cemr^2 < \kappa + 1$. Let us fix $\kappa + 1 = Kcemr^2 \in \mathcal{O}(mr^2)$ where $K > 1$ is a constant. Then we can rewrite the upper bound in the MSE from the previous equation as

$$\epsilon^2 \leq \left(\frac{1}{K} \right)^{Kcemr^2} e^{cmr^2} = \left(\frac{e}{K^K} \right)^{cmr^2}. \quad (C35)$$

Since we are interested in extending r to larger region (beyond $\mathcal{O}(1/\sqrt{m})$), we can assume $r \in \mathcal{O}(1/m^\beta)$ with $0 \leq \beta < 1/2$. In this case, we would have $\kappa \in \mathcal{O}(mr^2) = \mathcal{O}(m^{1-2\beta})$ where $1 - 2\beta > 0$. Substituting back the relation between κ and m into the number of resources required in Eq. (C31), the count scales as

$$N \in \mathcal{O}\left(m^{wm^{1-2\beta}}\right), \quad (C36)$$

which scales super-exponentially for $\beta = 0$, where $r \in \mathcal{O}(1)$.

More positively, we can reduce to *subexponential* (albeit still super-polynomial) if $0 < \beta < 1/2$. Note that if r decreases logarithmically with m , the resources also scales *super-polynomially*.

a. Pauli Propagation. The previous observations can be equivalently shown for Pauli propagation truncation case (see Appendix E). In this case, the maximal resource requirement (see Eq. (E56)) is proportional to

$$N \leq \left(\frac{em}{\kappa} \right)^\kappa, \quad (C37)$$

whereas the error will be (see Supplemental Theorem E1) of the following form

$$\epsilon^2 \leq \left(\frac{emr^2}{\kappa} \right)^\kappa. \quad (C38)$$

To make the error arbitrarily small, we would need at least $\kappa \in \Omega(mr^2)$. So assuming $r \in \mathcal{O}(1/m^\beta)$ with $0 \leq \beta < 1/2$, we have $\kappa \in \mathcal{O}(m^{1-2\beta})$. Since the resource scales exponentially with $\kappa \in \mathcal{O}(m^{1-2\beta})$, it will be exponential with m for the case $\beta = 0$, i.e., constant r and subexponential, but super-polynomial for $0 < \beta < 1/2$. Similarly, if κ decreases logarithmically with m , the resource scaling will be super-polynomial (but subexponential in general). Obviously, the case $r \sim 1$ makes no sense here since sine can be larger than cosine. Also, setting $\kappa = m$, means that we are not truncating anymore.

Within the Pauli-propagation framework, the required number of terms is inherently problem-dependent. While the worst-case complexity is super-polynomial, Pauli propagation can still be practical in specific settings. Equation (C37) should thus be interpreted as an upper bound arising from a maximal-splitting scenario.

2. Absolute error: Proof of Proposition C1

Proof. The worst-case error of the surrogate $\tilde{f}_\kappa(\alpha)$ of the order κ is written as

$$\epsilon_{\kappa, \alpha^*}^{\text{WC}} := \max_{\alpha \in \mathcal{V}_m(\alpha^*, r)} |\tilde{f}_\kappa(\alpha) - f(\alpha)|, \quad (C39)$$

where f_α is the expectation function, and $\mathcal{V}_m(\alpha^*, r)$ is the hypercube in the parameter space such that

$$\alpha_i \in [\alpha_i^* - r, \alpha_i^* + r]. \quad (C40)$$

By invoking the Taylor Remainder Theorem (i.e., Supplemental Theorem A2), we see that this is equivalent to bounding the remainder of the Taylor expansion. In particular, by denoting the parameters corresponding to the worst-case error

$$\alpha' = \arg \max_{\alpha \in \mathcal{V}_m(\alpha^*, r)} |\tilde{f}_\kappa(\alpha) - f(\alpha)|, \quad (C41)$$

we can use Supplemental Theorem A2 to explicitly write the worst-case error as

$$\begin{aligned} \epsilon_{\kappa, \alpha^*}^{\text{WC}} &= |\tilde{f}_\kappa(\alpha') - f(\alpha')| \\ &= \left| \frac{1}{(\kappa + 1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m \Delta_{i_1, \dots, i_{\kappa+1}}(\alpha', \alpha^*) \partial_{i_1, \dots, i_{\kappa+1}} f(\vartheta) \right|, \end{aligned} \quad (C42)$$

where $\vartheta = \tau \alpha' + (1 - \tau) \alpha^*$ for some $\tau \in [0, 1]$.

By applying the triangle inequality we can further simplify this expression to obtain

$$\begin{aligned} \epsilon_{\kappa, \alpha^*}^{\text{WC}} &= \left| \frac{1}{(\kappa + 1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m \Delta_{i_1, \dots, i_{\kappa+1}}(\alpha', \alpha^*) \partial_{i_1, \dots, i_{\kappa+1}} f(\vartheta) \right| \\ &\leq \frac{1}{(\kappa + 1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m |\Delta_{i_1, \dots, i_{\kappa+1}}(\alpha', \alpha^*) \partial_{i_1, \dots, i_{\kappa+1}} f(\vartheta)| \end{aligned} \quad (C44)$$

$$\leq \frac{1}{(\kappa + 1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m |\Delta_{i_1, \dots, i_{\kappa+1}}(\alpha', \alpha^*)| |\partial_{i_1, \dots, i_{\kappa+1}} f(\vartheta)| \quad (C45)$$

$$= \frac{1}{(\kappa + 1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m |\Delta_{i_1, \dots, i_{\kappa+1}}(\alpha', \alpha^*)| |\partial_{i_1, \dots, i_{\kappa+1}} f(\vartheta)| \quad (C46)$$

where in the last equality we use that the absolute value of the product is the product of absolute values, i.e., $|xy| = |x||y|$.

Next, let us focus on the terms $\partial_{i_1, \dots, i_{\kappa+1}} f(\boldsymbol{\theta})$. Since the assumption in Eq (C12) of Proposition C1 holds, the derivatives in the previous sum are bounded as

$$|\partial_{i_1, i_2, \dots, i_k} f(\boldsymbol{\alpha})| \leq \|O\|_{\infty} \gamma^k, \quad \forall \boldsymbol{\alpha} \in \mathcal{V}_m(\boldsymbol{\alpha}^*, r). \quad (\text{C47})$$

Combining the previous result with Eq. (C46), we can simplify the bound as

$$\epsilon_{\kappa, \boldsymbol{\alpha}^*}^{\text{WC}} \leq \frac{\|O\|_{\infty} \gamma^{\kappa+1}}{(\kappa+1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m |\Delta_{i_1, \dots, i_{\kappa+1}}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)|. \quad (\text{C48})$$

Next we can recall the definition of $\Delta_{i_1, \dots, i_{\kappa+1}}(\boldsymbol{\alpha}', \boldsymbol{\alpha}^*)$ in Eq. (C5)

$$\begin{aligned} \Delta_{i_1, \dots, i_{\kappa+1}}(\boldsymbol{\alpha}', \boldsymbol{\alpha}^*) \\ = (\alpha'_{i_1} - \alpha^*_{i_1})(\alpha'_{i_2} - \alpha^*_{i_2}) \cdots (\alpha'_{i_{\kappa+1}} - \alpha^*_{i_{\kappa+1}}) \end{aligned} \quad (\text{C49})$$

$$= \delta'_{i_1} \delta'_{i_2} \cdots \delta'_{i_{\kappa+1}}, \quad (\text{C50})$$

where we denote $\delta'_k = \alpha'_k - \alpha^*_k$. Therefore, we have

$$\epsilon_{\kappa, \boldsymbol{\alpha}^*}^{\text{WC}} \leq \frac{\|O\|_{\infty} \gamma^{\kappa+1}}{(\kappa+1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m |\delta'_{i_1} \cdots \delta'_{i_{\kappa+1}}| \quad (\text{C51})$$

$$\leq \frac{\|O\|_{\infty} \gamma^{\kappa+1}}{(\kappa+1)!} \sum_{i_1, \dots, i_{\kappa+1}=1}^m \gamma^{\kappa+1} \quad (\text{C52})$$

$$= \frac{\|O\|_{\infty}}{(\kappa+1)!} (\gamma r m)^{\kappa+1}, \quad (\text{C53})$$

where in the second inequality we used that $\max_{\alpha_i} |\alpha_i - \alpha^*_i| = r$ for all the α_i and to reach the last line we use $\sum_{i=1}^m 1 = m$ for all indices. Notice that $r \in \mathcal{O}(1/m)$, $\gamma \in \mathcal{O}(1)$ and $\|O\|_{\infty} \in \mathcal{O}(1)$ implies $\gamma r m \in \mathcal{O}(1)$ and therefore, the error is $\mathcal{O}(1)$ and decreases with κ , which completes the proof of the proposition. ■

3. Mean-square error: Proof of Proposition C2

Proof. The MSE is of the form

$$(\epsilon_{\kappa, \boldsymbol{\alpha}^*}^{\text{MSE}})^2 = \mathbb{E}_{\boldsymbol{\alpha} \sim \mathcal{D}_m(\boldsymbol{\alpha}^*, r_m)} \left[(\tilde{f}_{\kappa}(\boldsymbol{\alpha}) - f(\boldsymbol{\alpha}))^2 \right] \quad (\text{C54})$$

$$= \mathbb{E}_{\boldsymbol{\alpha} \sim \mathcal{D}_m(\boldsymbol{\alpha}^*, r_m)} \left[\left(\sum_{k=\kappa+1}^{\infty} \frac{1}{k!} \sum_{i_1, \dots, i_k=1}^m \Delta_{i_1, \dots, i_k}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) \partial_{i_1, \dots, i_k} f(\boldsymbol{\alpha}^*) \right)^2 \right], \quad (\text{C55})$$

where $\mathcal{D}_m(\boldsymbol{\alpha}^*, r_m)$ is a uniform distribution of $\boldsymbol{\alpha}$ in the hypercube region, i.e., $\mathcal{V}_m(\boldsymbol{\alpha}^*, r_m) = \{\boldsymbol{\alpha}\}$ such that $\alpha_i \in [\alpha^*_i - r_m, \alpha^*_i + r_m]$ for all i . Note that we explicitly add the subscript m in r_m indicating that it is associated with $\boldsymbol{\alpha}$ with m parameters. This will come in handy later in our proof. To reach the second equality, we use the definition of the Taylor expansion on the expectation function in Eq. (C3) and the definition of the surrogate in Eq. (C6) to see that $\tilde{f}_{\kappa}(\boldsymbol{\alpha}) - f(\boldsymbol{\alpha})$ is an infinite sum of terms.

To further proceed, we introduce a perturbation vector $\boldsymbol{\delta} := \boldsymbol{\alpha} - \boldsymbol{\alpha}^*$, whose i th components is $\delta_i = \alpha_i - \alpha^*_i$ for any index i . Then, we also denote the average of $\boldsymbol{\delta}$ over the perturbed region $\mathbb{E}_{\boldsymbol{\delta}}[\cdot] := \mathbb{E}_{\boldsymbol{\delta} \sim \mathcal{D}_m(\mathbf{0}, r_m)}[\cdot]$ where $\mathcal{D}_m(\mathbf{0}, r_m)$ is a uniform distribution over the region $\mathcal{V}_m(\mathbf{0}, r_m) = \{\boldsymbol{\delta}\}$ with $\delta_i \in [-r_m, r_m]$. More specifically, $\mathcal{D}_m(\mathbf{0}, r_m)$ is obtained by sampled independently each component δ_i over $[-r_m, r_m]$. Then, the MSE can be written as

$$(\epsilon_{\kappa, \boldsymbol{\alpha}^*}^{\text{MSE}})^2 = \mathbb{E}_{\boldsymbol{\alpha} \sim \mathcal{D}_m(\boldsymbol{\alpha}^*, r)} \left(\sum_{k=\kappa+1}^{\infty} \frac{1}{k!} \sum_{i_1, \dots, i_k=1}^m \Delta_{i_1, \dots, i_k}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) \partial_{i_1, \dots, i_k} f(\boldsymbol{\alpha}^*) \right)^2 \quad (\text{C56})$$

$$= \sum_{k, l=\kappa+1}^{\infty} \frac{1}{l! k!} \sum_{i_1, \dots, i_k=1}^m \sum_{j_1, \dots, j_l=1}^m \mathbb{E}_{\boldsymbol{\alpha} \sim \mathcal{D}_m(\boldsymbol{\alpha}^*, r)} [\Delta_{i_1, \dots, i_k}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) \Delta_{j_1, \dots, j_l}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)] \partial_{i_1, \dots, i_k} f(\boldsymbol{\alpha}^*) \partial_{j_1, \dots, j_l} f(\boldsymbol{\alpha}^*) \quad (\text{C57})$$

$$\leq \sum_{k, l=\kappa+1}^{\infty} \frac{1}{l! k!} \sum_{i_1, \dots, i_k=1}^m \sum_{j_1, \dots, j_l=1}^m \mathbb{E}_{\boldsymbol{\delta}} [\delta_{i_1} \cdots \delta_{i_k} \delta_{j_1} \cdots \delta_{j_l}] \|O\|_{\infty}^2 \gamma^{k+l} \quad (\text{C58})$$

$$= \|O\|_\infty^2 \sum_{k,l=\kappa+1}^{\infty} \frac{\gamma^{k+l}}{l!k!} \mathbb{E}_\delta \left[\sum_{i_1=1}^m \delta_{i_1} \cdots \sum_{i_k=1}^m \delta_{i_k} \sum_{j_1=1}^m \delta_{j_1} \cdots \sum_{j_l=1}^m \delta_{j_l} \right] \quad (\text{C59})$$

$$= \|O\|_\infty^2 \sum_{k,l=\kappa+1}^{\infty} \frac{\gamma^{k+l}}{l!k!} \mathbb{E}_\delta \left[\left(\sum_{i=1}^m \delta_i \right)^{k+l} \right], \quad (\text{C60})$$

where the inequality is by the assumption of the bounded derivatives as specified in Eq. (C16),

$$|\partial_{i_1, i_2, \dots, i_k} f(\alpha)| \leq \|O\|_\infty \gamma^k, \quad \forall \alpha \in \mathcal{V}_m(\alpha^*, r_m), \quad (\text{C61})$$

together with the fact that the averages of the perturbations are always non-negative due to the symmetry in the perturbations. In particular, since $\delta_k \in [-r_m, r_m]$, we have that the uniform average results in $\mathbb{E}_{\delta_k}[\delta_k^{2a+1}] = 0$ and $\mathbb{E}_{\delta_k}[\delta_k^{2a}] \geq 0$ for any integer $a \in \mathbb{N}$. Hence, this justified Eq. (C58). To reach the last equality, we simply massage the expression by pulling out the average and rewriting the sums.

From here onward we focus on bounding the average term $\mathbb{E}_\delta \left[\left(\sum_{i=1}^m \delta_i \right)^{k+l} \right]$. Particularly, one of our key ingredients is the central limit theorem. To see this, we begin with denoting

$$S_m = \sum_{i=1}^m \delta_i, \quad (\text{C62})$$

together with its variance

$$\sigma_m^2 := \text{Var}_\delta[S_m] = \frac{mr_m^2}{3}, \quad (\text{C63})$$

where we use the fact that the variables $\{\delta_i\}$ are independent and $\text{Var}_{\delta_i}[\delta_i] = r_m^2/3$ for a uniform distribution. Furthermore, we can define a random variable

$$X_m = \frac{S_m}{\sigma_m}, \quad (\text{C64})$$

together with its variance which is equal to 1, i.e.,

$$\text{Var}_\delta[X_m] = 1. \quad (\text{C65})$$

The central limit theorem can now be applied to X_m (see Supplemental Theorem A1), which essentially states that when m approaches infinity, the probability distribution of the random variable X_m approaches a normal distribution with the zero mean and variance equal to one, i.e.,

$\mathcal{N}(0, 1)$. As stated in Appendix A3, the moments of a random variable under $\mathcal{N}(0, 1)$ are

$$\mathbb{E}_\delta[X_{m \rightarrow \infty}^a] = \begin{cases} 0 & \text{if } a \text{ is odd,} \\ (a-1)!! & \text{if } a \text{ is even,} \end{cases} \quad (\text{C66})$$

for any integer $a \in \mathbb{N}$ and $x!!$ being the double factorial function, defined for $x \in \mathbb{N}$ as $x!! = x(x-2)(x-4)(x-6) \cdots 1$.

Next we will upper bound $\mathbb{E}[X_m^a]$ with $\mathbb{E}[X_{m \rightarrow \infty}^a]$ for any m . That is, the expected value of X_m^a tends to a maximum with increasing m . To see this, first notice that X_{m+1} can be written in terms of X_m as

$$X_{m+1} = \frac{S_m + \delta_{m+1}}{\sigma_{m+1}} = \frac{\sigma_m}{\sigma_{m+1}} X_m + \frac{\delta_{m+1}}{\sigma_{m+1}}. \quad (\text{C67})$$

We can bound the a th moment of X_{m+1} with the a th moment of X_m as

$$\mathbb{E}_\delta[X_{m+1}^a] = \mathbb{E}_\delta \left[\left(\frac{\sigma_m}{\sigma_{m+1}} X_m + \frac{\delta_{m+1}}{\sigma_{m+1}} \right)^a \right] \quad (\text{C68})$$

$$= \sum_{i=0}^a \binom{a}{i} \mathbb{E}_\delta[X_m^i] \left(\frac{\sigma_m}{\sigma_{m+1}} \right)^i \frac{\mathbb{E}_{\delta_{m+1}}[\delta_{m+1}^{a-i}]}{\sigma_{m+1}^{a-i}} \quad (\text{C69})$$

$$\geq \left(\frac{\sigma_m}{\sigma_{m+1}} \right)^a \mathbb{E}_\delta[X_m^a], \quad (\text{C70})$$

where in the second line we use the binomial expansion $(x+y)^a = \sum_{i=0}^a \binom{a}{i} x^i y^{a-i}$ and the linearity of the expectation, while in the inequality we dropped all the terms (which are all non-negative) except for $i = a$.

To proceed, we consider the scenario where $\sigma_m/\sigma_{m+1} \geq 1$ which implies that

$$\mathbb{E}_\delta[X_{m+1}^a] \geq \mathbb{E}_\delta[X_m^a]. \quad (\text{C71})$$

To determine when $\sigma_m/\sigma_{m+1} \geq 1$, we further consider r_m to be of the form $r_m = cm^x$ with some constant c . Then, we can see that $\sigma_m/\sigma_{m+1} \geq 1$ is valid only if $x \leq -1/2$.

More precisely, by using Eq. (C63) we have

$$\frac{\sigma_m}{\sigma_{m+1}} = \left(\frac{m}{m+1} \right)^{2x+1} \geq 1 \quad \text{iff } x \leq -\frac{1}{2}. \quad (\text{C72})$$

Hence, from here onward, we consider that

$$r_m \in \mathcal{O} \left(\frac{1}{\sqrt{m}} \right), \quad (\text{C73})$$

justifying Eq. (C71) and leading to

$$\begin{aligned} \mathbb{E}_\delta[X_m^a] &\leq \mathbb{E}_\delta[X_{m+1}^a] \leq \dots \leq \mathbb{E}_\delta[X_{m \rightarrow \infty}^a] \\ &= (a-1)!! \quad \text{if } a \text{ is even,} \end{aligned} \quad (\text{C74})$$

where in the last equality we used Eq. (C66). We remind that only even a terms do not vanish, since for odd a we have $\mathbb{E}_\delta[X_m^a] = 0$.

We are now ready to revisit Eq. (C60) by applying the previously obtained bound. In particular, we have

$$(\epsilon_{\kappa, \alpha^*}^{\text{MSE}})^2 \leq \|O\|_\infty^2 \sum_{k,l=\kappa+1}^{\infty} \frac{\gamma^{k+l}}{k!l!} \mathbb{E}_\delta \left[\left(\sum_{i=1}^m \delta_i \right)^{k+l} \right] \quad (\text{C75})$$

$$= \|O\|_\infty^2 \sum_{k,l=\kappa+1}^{\infty} \frac{\gamma^{k+l}}{k!l!} \sigma_m^{k+l} \mathbb{E}_\delta[X_m^{k+l}] \quad (\text{C76})$$

$$\leq \|O\|_\infty^2 \sum_{k,l=\kappa+1}^{\infty} \frac{(\gamma \sigma_m)^{k+l}}{k!l!} \mathbb{E}_\delta[X_{m \rightarrow \infty}^{k+l}] \quad (\text{C77})$$

$$= \|O\|_\infty^2 \sum_{\substack{k,l=\kappa+1 \\ k+l \text{ even}}}^{\infty} \frac{(\gamma \sigma_m)^{k+l}}{k!l!} (k+l-1)!!, \quad (\text{C78})$$

where in the first equality we used that $S_m = X_m \sigma_m$. In the second inequality we used the bound derived in Eq. (C74).

In the next few steps, we will do a series of changes of indices, which leads to a more manageable expression. First, we denote $k = \kappa + 1 + i$, $l = \kappa + 1 + j$ and explicitly add a condition to force that $i + j$ is even. Particularly, we have

$$\begin{aligned} &\sum_{\substack{k,l=\kappa+1 \\ k+l \text{ even}}}^{\infty} \frac{(\gamma \sigma_m)^{k+l}}{k!l!} (k+l-1)!! \\ &= \sum_{i,j=0}^{\infty} \frac{(2\kappa+1+i+j)!!}{(\kappa+1+i)!(\kappa+1+j)!} (\gamma \sigma_m)^{2\kappa+2+i+j} \\ &\quad \times \left(\frac{1+(-1)^{i+j}}{2} \right), \end{aligned} \quad (\text{C79})$$

where the last fraction ensures $i + j$ is even.

We can now make another change in the variables that will help with the computation. More precisely, for any h_{ij} we will use

$$\sum_{i,j=0}^{\infty} h_{i,j} = \sum_{y'=0}^{\infty} \sum_{\substack{i,j=0 \\ i+j=y'}}^{y'} h_{i,j} = \sum_{y'=0}^{\infty} \sum_{x=0}^{y'} h_{x,y'-x}, \quad (\text{C80})$$

where in the first equality we are adding all the combinations of $i + j$ that add up to y' , and then in the second equality we substitute $i = x$ and $j = y' - x$ (and thus we keep $y' = i + j$). This last change in indices will allow us to write the sum more simply. Indeed if we apply it Eq. (C79) we find

$$\begin{aligned} &\sum_{i,j=0}^{\infty} \frac{(2\kappa+1+i+j)!!}{(\kappa+1+i)!(\kappa+1+j)!} (\gamma \sigma_m)^{2\kappa+2+i+j} \frac{1+(-1)^{i+j}}{2} \\ &= \sum_{y'=0}^{\infty} \sum_{x=0}^{y'} \frac{(2\kappa+1+y')!!}{(\kappa+1+x)!(\kappa+1+y'-x)!} (\gamma \sigma_m)^{2\kappa+2+y'} \\ &\quad \times \frac{1+(-1)^{y'}}{2}. \end{aligned} \quad (\text{C81})$$

It is clear from the previous expression that y' has to be even, thus we only keep the terms of y' that are divisible by 2. Enforcing this condition is equivalent to (i) replacing $y' = 2y$ in the terms and in the limit of the second sum, and (ii) keeping the limit of the first sum from $y = 0$ to ∞ . Implementing (i) and (ii) leads to

$$\sum_{y'=0}^{\infty} \sum_{x=0}^{y'} \frac{(2\kappa+1+y')!!}{(\kappa+1+x)!(\kappa+1+y'-x)!} (\gamma \sigma_m)^{2\kappa+2+y'} \frac{1+(-1)^{y'}}{2} = \sum_{y=0}^{\infty} \sum_{x=0}^{2y} \frac{(2\kappa+1+2y)!!}{(\kappa+1+x)!(\kappa+1+2y-x)!} (\gamma \sigma_m)^{2\kappa+2+2y}. \quad (\text{C82})$$

We have finished the process of changing the variables and will proceed to bound the factorial terms. We start by recalling the identity $(2k+1)!! = (2k+2)!/2^{k+1}(k+1)!$ with $k = \kappa + y$, which leads to

$$\sum_{y=0}^{\infty} \sum_{x=0}^{2y} \frac{(2\kappa + 1 + 2y)!!}{(\kappa + 1 + x)!(\kappa + 1 + 2y - x)!} (\gamma \sigma_m)^{2\kappa+2+2y} = \sum_{y=0}^{\infty} \frac{(\gamma \sigma_m)^{2\kappa+2+2y}}{2^{\kappa+y+1}(\kappa + y + 1)!} \sum_{x=0}^{2y} \frac{(2\kappa + 2 + 2y)!}{(\kappa + 1 + x)!(\kappa + 1 + 2y - x)!}. \quad (C83)$$

Let us focus now on upper bounding the second sum involving x . We write the factorial terms in a binomial form as

$$\sum_{x=0}^{2y} \frac{(2\kappa + 2 + 2y)!}{(\kappa + 1 + x)!(\kappa + 1 + 2y - x)!} = \sum_{x=0}^{2y} \binom{2\kappa + 2 + 2y}{\kappa + 1 + x} \leq \sum_{x=-\kappa-1}^{2y+\kappa+1} \binom{2\kappa + 2 + 2y}{\kappa + 1 + x} = 2^{2\kappa+2+2y}, \quad (C84)$$

where in the inequality we use that all the terms in the summation are positive, meaning that we can increase the limits of the sum. Then, in the last equality we use that the sum of all these terms equal the exponential (as per the Binomial Theorem/Binomial expansion). Replacing this bound in Eq. (C83) leads to

$$\sum_{y=0}^{\infty} \frac{(\gamma \sigma_m)^{2\kappa+2+2y}}{2^{\kappa+y+1}(\kappa + y + 1)!} \sum_{x=0}^{2y} \frac{(2\kappa + 2 + 2y)!}{(\kappa + 1 + x)!(\kappa + 1 + 2y - x)!} \leq \sum_{y=0}^{\infty} \frac{(2\gamma^2 \sigma_m^2)^{\kappa+1+y}}{(\kappa + y + 1)!} \quad (C85)$$

$$\leq \sum_{y=0}^{\infty} \frac{(2\gamma^2 \sigma_m^2)^{\kappa+1+y}}{(\kappa + 1)!y!} \quad (C86)$$

$$= \frac{(2\gamma^2 \sigma_m^2)^{\kappa+1}}{(\kappa + 1)!} e^{2\gamma^2 \sigma_m^2}, \quad (C87)$$

where in the second inequality we use that $(\kappa + 1 + y)! \geq (\kappa + 1)!y!$, and we use the fact that $\sum_{k=0}^{\infty} x^k/k! = e^x$.

Finally, we revisit Eq. (C78) by writing in terms of the variables and bounding with Eq. (C87), which leads to

$$(\epsilon_{\kappa, \alpha^*}^{\text{MSE}})^2 \leq \|O\|_{\infty}^2 \frac{(2\gamma^2 \sigma_m^2)^{\kappa+1}}{(\kappa + 1)!} e^{2\gamma^2 \sigma_m^2} \quad (C88)$$

$$= \frac{\|O\|_{\infty}^2}{(\kappa + 1)!} \left(\frac{2m\gamma^2 r_m^2}{3} \right)^{\kappa+1} e^{\frac{2m\gamma^2 r_m^2}{3}}, \quad (C89)$$

where to reach the last equality we use Eq. (C63). We remind again that we explicitly use the condition that $r_m \in \mathcal{O}(1/\sqrt{m})$ in Eq. (C72) for the bound to be valid. Moreover, assuming $\gamma \in \mathcal{O}(1)$ implies $m\gamma^2 r_m^2 \in \mathcal{O}(1)$. Therefore, we can see that the MSE above can be made arbitrarily small by increasing κ . Especially, if $\|O\|_{\infty} \in \mathcal{O}(1)$ we have $\epsilon_{\kappa, \alpha^*}^{\text{MSE}} \in \mathcal{O}(1)$. ■

4. Scaling of the number of expectation function terms for given surrogation order: Proof of Proposition C3

Proof. We start this proof by recalling the expression of the surrogate in Eq. (C6)

$$\tilde{f}_{\kappa}(\alpha) = \sum_{k=0}^{\kappa} \frac{1}{k!} \sum_{i_1, i_2, \dots, i_k} \partial_{i_1, \dots, i_k} f(\alpha^*) \Delta_{i_1, \dots, i_k}(\alpha, \alpha^*). \quad (C90)$$

Now, let us rewrite this expression as a sum over unique partial derivatives. Indeed, we can group terms using the symmetry of the partial derivatives (e.g., $\partial_{i,j} f = \partial_{j,i} f$) to give

$$\tilde{f}_{\kappa}(\alpha) = \sum_{k=0}^{\kappa} \frac{1}{k!} \sum_{i_1, i_2, \dots, i_k=1}^m \partial_{i_1, \dots, i_k} f(\alpha^*) \Delta_{i_1, \dots, i_k}(\alpha, \alpha^*) \quad (C91)$$

$$= \sum_{k=0}^{\kappa} \frac{1}{k!} \sum_{i_1, i_2, \dots, i_k=1}^m \prod_{i_l=i_1}^{i_k} (\delta_{i_l} \partial_{i_l}) f(\alpha^*), \quad (C92)$$

where we noted $\Delta_{i_1, \dots, i_k}(\alpha, \alpha^*) = \prod_{i_l=i_1}^{i_k} \delta_{i_l}$, with $\delta_{i_l} = \alpha_{i_l} - \alpha_{i_l}^*$. We can recognize that this expression contains all the terms of a multinomial expansion. Indeed, the sum of products $\sum_{i_1, i_2, \dots, i_k=1}^m \prod_{i_l=i_1}^{i_k} (\delta_{i_l} \partial_{i_l})$ contains all the possible combinations of products $(\delta_{i_l} \partial_{i_l})$ up to a maximum power of k . Therefore, we can just rewrite this expression in a more compact form making use of the multinomial expression. That is, as

$$\tilde{f}_{\kappa}(\alpha) = \sum_{k=0}^{\kappa} \frac{\left(\sum_{i=1}^m \delta_i \frac{\partial}{\partial \alpha_i} \right)^k}{k!} f(\alpha^*) \quad (C93)$$

$$= \sum_{k=0}^{\kappa} \frac{1}{k!} \sum_{k_1+\dots+k_m=k} \binom{k}{k_1, \dots, k_m} \prod_{l=1}^m (\delta_l^{k_l} \partial_l^{k_l}) f(\alpha^*) \quad (\text{C94})$$

$$= \sum_{k=0}^{\kappa} \sum_{k_1+\dots+k_m=k} \prod_{l=1}^m \left(\frac{\delta_l^{k_l}}{k_l!} \partial_l^{k_l} \right) f(\alpha^*), \quad (\text{C95})$$

where the second equality is obtained by expanding the multinomial. In this expression we have introduced $k_l \in \mathbb{N}$ which is the order of the derivative with respect to parameter α_l for each $l \in \{1, 2, \dots, m\}$. The last equality is obtained by simplifying the expression of the multinomial coefficient and the factorial term

$$\binom{k}{k_1, \dots, k_m} \frac{1}{k!} = \prod_{l=1}^m \frac{1}{k_l!}. \quad (\text{C96})$$

The number of different terms N that need to be evaluated in order to construct the surrogate/simulate the expectation function landscape corresponds to the number of different derivatives. Indeed, for example, in the case of surrogation, each derivative would be a different term that would need to be estimated using a quantum computer. In other words, we assume that each partial derivative can be obtained by at worst measuring N_d different parameter settings of the expectation function on a quantum computer. For the k th order in the surrogate, there exists $\binom{m+k-1}{k}$ unique partial derivatives (which is the number of possible $\mathbf{k} = (k_1, k_2, \dots, k_m) \in \mathbb{N}^m$ such that $\sum_{l=1}^m k_l = k$), leading to at most $\binom{m+k-1}{k} N_d$ expectation values to be measured in a quantum computer. Then we can readily compute an upper bound in the total number of terms N we will have to measure by summing all the orders in the surrogate,

$$N = \sum_{k=0}^{\kappa} \binom{m+k-1}{k} N_d \leq N_d \left[\frac{e(m+\kappa-1)}{\kappa} \right]^{\kappa}, \quad (\text{C97})$$

where the inequality is obtained using $\binom{n}{k} \leq n^k/k!$ as follows:

$$\sum_{k=0}^{\kappa} \binom{m+k-1}{k} \leq \sum_{k=0}^{\kappa} \frac{(m+k-1)^k}{k!} \quad (\text{C98})$$

$$= \sum_{k=0}^{\kappa} \frac{\kappa^k}{k!} \left(\frac{m+k-1}{\kappa} \right)^k \quad (\text{C99})$$

$$\leq \left(\frac{m+\kappa-1}{\kappa} \right)^{\kappa} \sum_{k=0}^{\kappa} \frac{\kappa^k}{k!} \quad (\text{C100})$$

$$\leq \left(\frac{e(m+\kappa-1)}{\kappa} \right)^{\kappa}, \quad (\text{C101})$$

where the second inequality is obtained by noticing that the term $((m+k-1)/\kappa)^k$ reaches its largest value when $k = \kappa$ (as it increases with k for all $\kappa \leq m-1$), and the last inequality is obtained by adding terms in the sum up to $k \rightarrow \infty$ and recognizing the exponential series, i.e., $\sum_{k=0}^{\infty} \kappa^k/k! = e^{\kappa}$. ■

5. General surrogation guarantee: Proof of Supplemental Theorem C1

Proof. Assumptions C2 and C3 implies that the κ -order surrogate $\tilde{f}_{\kappa}(\alpha)$ can be expressed as a linear combination of expectation functions evaluated at different parameters. In particular, from Assumption C3, we can express each partial derivatives in Eq. (C95) as

$$\prod_{l=1}^m \frac{\partial \tilde{f}_{\kappa}(\alpha^*)}{\partial \alpha_l^{k_l}} = \sum_{j=1}^{N_d} b_j^{(\mathbf{k})} f(\alpha_j^{(\mathbf{k})}), \quad (\text{C102})$$

where $\mathbf{k} = (k_1, k_2, \dots, k_m) \in \mathbb{N}^m$. Therefore, Eq. (C95) can be rewritten as follows:

$$\tilde{f}_{\kappa}(\alpha) = \sum_{k=0}^{\kappa} \sum_{k_1+\dots+k_m=k} \prod_{l=1}^m \frac{\delta_l^{k_l}}{k_l!} \sum_{j=1}^{N_d} b_j^{(\mathbf{k})} f(\alpha_j^{(\mathbf{k})}) \quad (\text{C103})$$

$$= \sum_{k=0}^{\kappa} \sum_{|\mathbf{k}|=k} \sum_{j=1}^{N_d} \left(b_j^{(\mathbf{k})} \prod_{l=1}^m \frac{\delta_l^{k_l}}{k_l!} \right) f(\alpha_j^{(\mathbf{k})}) \quad (\text{C104})$$

$$= \sum_{|\mathbf{k}| \leq \kappa} \sum_{j=1}^{N_d} \left(b_j^{(\mathbf{k})} \prod_{l=1}^m \frac{\delta_l^{k_l}}{k_l!} \right) f(\alpha_j^{(\mathbf{k})}) \quad (\text{C105})$$

$$= \sum_{l=1}^N c_l(\alpha) f(\alpha_l), \quad (\text{C106})$$

where in the last equality we use an index l for each possible choice of $(\mathbf{k}, j)$ such that $\mathbf{k} \in \mathbb{N}^m$ with $|\mathbf{k}| = k_1 + \dots + k_m \leq \kappa$ and $j \in \{1, \dots, N_d\}$ [notice that there are N possible choice of $(\mathbf{k}, j)$]. So, we defined $f(\alpha_l) = f(\alpha_j^{(\mathbf{k})})$ and $c_l(\alpha) = b_j^{(\mathbf{k})} \prod_{l=1}^m \delta_l^{k_l}/k_l!$. The total number of terms in the sum, N is given in Eq. (C97) as well as its upper bound (see Proposition C3).

(i). *The worst-case error with the perturbation $r \in \mathcal{O}(1/m)$.*

The total error can be upper bounded by the sum of both truncation error and shot noise error using the triangle inequality

$$\left| \tilde{f}_{\kappa}(\alpha) - f(\alpha) \right| \leq \left| \tilde{f}_{\kappa}(\alpha) - \tilde{f}_{\kappa}(\alpha) \right| + \left| \tilde{f}_{\kappa}(\alpha) - f(\alpha) \right|. \quad (\text{C107})$$

The term $|\tilde{f}_{\kappa}(\alpha) - f(\alpha)|$ can be bounded using Proposition C1. We now focus on the term $|\tilde{f}_{\kappa}(\alpha) - \tilde{f}_{\kappa}(\alpha)|$. Lemma B3

can be applied to the surrogate of the expectation function expressed in the form of Eq. (C106) by noticing that for each l we can define operator $O(\alpha_l)$ such that $f(\alpha_l) = \text{Tr}[\rho O(\alpha_l)]$ and $\|O(\alpha_l)\|_\infty \leq \|O\|_\infty$ with associate coefficients $c_l(\alpha)$. Let $\mathbf{c}_\alpha$ be the vector of length N with the l th component $c_l(\alpha)$. Therefore, the measurement strategy presented in Lemma B3 can be applied to output an estimator $\hat{f}_\kappa(\alpha)$ of $\tilde{f}_\kappa(\alpha)$ such that N_s measurements on different copies of ρ are sufficient to ensure

$$\left| \hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right| < \|\mathbf{c}_\alpha\|_{1,\text{worst}} \|O\|_\infty \sqrt{\frac{2 \log(2/\delta)}{N_s}}, \quad (\text{C108})$$

with probability at least $1 - \delta$. Now, using the assumption that the coefficients $b_j^{(k)}$ are bounded by b_0 , i.e., $|b_j^{(k)}| \leq b_0$, we have

$$\|\mathbf{c}_\alpha\|_1 = \sum_{l=1}^N |c_l(\alpha)| \quad (\text{C109})$$

$$= \sum_{|k| \leq \kappa} \sum_{j=1}^{N_d} |b_j^{(k)}| \prod_{l=1}^m \frac{|\delta_l|^{k_l}}{k_l!} \quad (\text{C110})$$

$$\leq b_0 N_d \sum_{|k| \leq \kappa} \prod_{l=1}^m \frac{|\delta_l|^{k_l}}{k_l!} \quad (\text{C111})$$

$$\leq b_0 N_d \sum_{|k| \leq \kappa} \prod_{l=1}^m \frac{r^{k_l}}{k_l!} \quad (\text{C112})$$

$$= b_0 N_d \sum_{k=0}^{\kappa} \frac{(rm)^k}{k!}, \quad (\text{C113})$$

where the second inequality follows from $|\delta_l| \leq r$ and the last equality is obtained by recognizing the multinomial sum $\sum_{|k|=k} \binom{k}{k_1, \dots, k_m} \prod_{l=1}^m r^{k_l} = (mr)^k$ where $\binom{k}{k_1, \dots, k_m} = k! / \prod_{l=1}^m k_l!$. Notice that this bound is valid for any $\alpha \in \mathcal{V}_m(\alpha^*, r)$ due to the second inequality, so this is also an upper bound of $\|\mathbf{c}_\alpha\|_{1,\text{worst}}$. Additionally, we can upper bound Eq. (C113) by

$$b_0 N_d \sum_{k=0}^{\kappa} \frac{(rm)^k}{k!} \leq b_0 N_d e^{rm}, \quad (\text{C114})$$

which is obtained by extending the limit of the sum to ∞ and recognizing the exponential series. Assuming that $rm \in \mathcal{O}(1)$ ensures that $e^{rm} \sim \mathcal{O}(1)$. By inserting previous bound on the effective 1-norm in Eq. (C108) to bound the shot noise error, and using Proposition C1 to bound the

truncation error, the total error is bounded as follows:

$$\left| \hat{f}_\kappa(\alpha) - f(\alpha) \right| \leq b_0 N_d e^{rm} \|O\|_\infty \sqrt{\frac{2 \log(2/\delta)}{N_s}} + \frac{\|O\|_\infty}{(\kappa + 1)!} (\gamma rm)^{\kappa+1}, \quad (\text{C115})$$

which happens with probability at least $1 - \delta$. Assuming that $b_0, N_d \in \mathcal{O}(\text{poly}(m))$, then we can see that using polynomial number of shots is enough to decrease the empirical error while the truncation error can be reduced by increasing κ .

(ii). *The average-case error with the perturbation $r \in \mathcal{O}(1/\sqrt{m})$.*

First, let us bound the total mean square error by both contribution of the empirical error and the truncation error using that $(x + y)^2 \leq 2(x^2 + y^2)$ for any reals x and y . This leads to

$$\mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\hat{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] \leq 2 \mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] + 2 \mathbb{E}_\alpha \left[\left(\tilde{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right]. \quad (\text{C116})$$

Notice that the average here is over both the random parameters α and the output of the measurement denoted by $\mathcal{M}$ (except for the truncation error which depends only on α).

Now, the estimate $\hat{f}_\kappa(\alpha)$ obtained by following the measurement protocol described in Lemma B2 leads to the following upper bound on the empirical mean square error

$$\mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] \leq \frac{\|O\|_\infty^2 \|\mathbf{c}\|_{1,\text{avg}}^2}{N_s}, \quad (\text{C117})$$

where $\|\mathbf{c}\|_{1,\text{avg}}$ is the averaged effective 1-norm (see Definition B1) of the vector $\mathbf{c}_\alpha$.

Finally, combining this result together with the MSE upper bound of the surrogate from Proposition C2 leads to

$$\begin{aligned} \mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\hat{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] &\leq 2 \frac{\|O\|_\infty^2 \|\mathbf{c}\|_{1,\text{avg}}^2}{N_s} + 2 \left(\frac{2\gamma^2 m r^2}{3} \right)^{\kappa+1} \frac{\|O\|_\infty^2}{(\kappa + 1)!} e^{\frac{2\gamma^2 m r^2}{3}}. \end{aligned} \quad (\text{C118})$$

Notice that Eq. (C113) is also a valid upper bound for $\|\mathbf{c}\|_{1,\text{avg}}$ as it corresponds to a worst-case 1-norm obtained by upper bounding each $|c_l(\alpha)|$. Therefore, the sum in

Eq. (C113) is bounded as follows:

$$\sum_{k=0}^{\kappa} \frac{(rm)^k}{k!} \leq (\kappa + 1) \frac{(rm)^{\kappa}}{\kappa!}, \quad (\text{C119})$$

under the condition that $rm > \kappa$ from Eq. (A8) (since $r \in \mathcal{O}(1/\sqrt{m})$, we have that $rm \in \mathcal{O}(\sqrt{m})$). From the assumptions, the bound on the effective norm scales polynomially with the number of parameters. Finally, we have

$$\begin{aligned} \mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\widehat{f}_k(\alpha) - f(\alpha) \right)^2 \right] \\ \leq 2 \left(\frac{b_0 N_d (\kappa + 1) (rm)^{\kappa} \|O\|_{\infty}}{\sqrt{N_s} \kappa!} \right)^2 \\ + 2 \left(\frac{2\gamma^2 m r^2}{3} \right)^{\kappa+1} \frac{\|O\|_{\infty}^2}{(\kappa + 1)!} e^{\frac{2\gamma^2 m r^2}{3}}. \end{aligned} \quad (\text{C120})$$

Notice that both measurement strategies from Lemmas B2 and B3 can be used to get previous bound as we upper bounded the average-case 1-norm by the worst-case one. ■

6. Parametrized quantum circuits that satisfy assumptions: Proof of Proposition C4 and Corollary C1

Corollary C1 is a direct consequence of Supplemental Theorem C1 and Proposition C4. Let us first derive Proposition C4.

Proof. The expectation function considered here is of the form of Eq. (C1) with unitary channel, i.e., $f(\alpha) = \text{Tr}[\rho U^{\dagger}(\alpha) O U(\alpha)]$, where $U(\alpha) = \prod_{l=1}^m V_l U_l(\alpha_l)$ such that V_l are arbitrary nonparametrized unitaries and each $U_l(\alpha_l) = \exp(-iH_l \alpha_l)$ is a parametrized gate with Hermitian generator H_l . Now, let us bound the k th-order derivatives of the form $\partial^k f(\alpha) / \prod_{l=1}^m \partial \alpha_l^{k_l} := \partial^k f(\alpha)$ where $\sum_{l=1}^m k_l = k$ and $k_l \in \mathbb{N}$ for each l . First, let us use Hölder inequality as follows:

$$|\partial^k f(\alpha)| = |\text{Tr}[\rho \partial^k U^{\dagger}(\alpha) O U(\alpha)]| \quad (\text{C121})$$

$$\leq \|\rho\|_1 \|\partial^k U^{\dagger}(\alpha) O U(\alpha)\|_{\infty} \quad (\text{C122})$$

$$= \|\partial^k (U^{\dagger}(\alpha) O U(\alpha))\|_{\infty}, \quad (\text{C123})$$

where the last equality is given by the normalization of quantum states. Now, the goal is to show that

$$\|\partial^k (U^{\dagger}(\alpha) O U(\alpha))\|_{\infty} \leq \|O\|_{\infty} \prod_{l=1}^m (2 \|H_l\|_{\infty})^{k_l} \quad (\text{C124})$$

by induction on l . First notice that the operator we want to bound can be rewritten as follows:

$$\partial^k ((U^{\dagger}(\alpha) O U(\alpha))) = \frac{\partial^{k_1}}{\partial \alpha_1^{k_1}} \left(U_1^{\dagger}(\alpha_1) V_1^{\dagger} \left(\dots \frac{\partial^{k_m}}{\partial \alpha_m^{k_m}} (U_m^{\dagger}(\alpha_m) V_m^{\dagger} O V_m U_m(\alpha_m)) \dots \right) V_1 U_1(\alpha_1) \right). \quad (\text{C125})$$

Then, each derivative consists in applying a commutator with the generator or more formally for any operator A_{l+1} (which will be specified later), we have

$$\begin{aligned} \frac{\partial^{k_l}}{\partial \alpha_l^{k_l}} \left(U_l^{\dagger}(\alpha_l) A_{l+1} U_l(\alpha_l) \right) \\ = \frac{\partial^{k_l-1}}{\partial \alpha_l^{k_l-1}} \left(U_l^{\dagger}(\alpha_l) i[H_l, A_{l+1}] U_l(\alpha_l) \right) \end{aligned} \quad (\text{C126})$$

$$= \frac{\partial^{k_l-2}}{\partial \alpha_l^{k_l-2}} \left(U_l^{\dagger}(\alpha_l) i[H_l, i[H_l, A_{l+1}]] U_l(\alpha_l) \right) \quad (\text{C127})$$

$$= \dots \quad (\text{C128})$$

$$= i^{k_l} U_l^{\dagger}(\alpha_l) \underbrace{[H_l, \dots, [H_l, A_{l+1}]]}_{k_l \text{ times}} U_l(\alpha_l). \quad (\text{C129})$$

From this result and the unitarily invariance properties of the norm, we have

$$\begin{aligned} \left\| \frac{\partial^{k_l}}{\partial \alpha_l^{k_l}} \left(U_l^{\dagger}(\alpha_l) V_l^{\dagger} A_{l+1} V_l U_l(\alpha_l) \right) \right\|_{\infty} \\ = \left\| [H_l, \dots, [H_l, V_l^{\dagger} A_{l+1} V_l] \dots] \right\|_{\infty} \end{aligned} \quad (\text{C130})$$

$$\leq (2 \|H_l\|_{\infty})^{k_l} \|V_l^{\dagger} A_{l+1} V_l\|_{\infty} \quad (\text{C131})$$

$$= (2 \|H_l\|_{\infty})^{k_l} \|A_{l+1}\|_{\infty}, \quad (\text{C132})$$

where the last equality follows from unitarily invariance of the norm, and the inequality is obtained by induction using

the property that for any operators A and B , we have

$$\|[A, B]\|_\infty \leq \|AB\|_\infty + \|BA\|_\infty \quad (\text{C133})$$

$$\leq 2 \|A\|_\infty \|B\|_\infty, \quad (\text{C134})$$

where we first use triangle inequality, and then submultiplicativity of the spectral norm. Now, using Eq. (C132) with $A_l = \partial^{k_l} / \partial \alpha_l^{k_l} \left(U_l^\dagger(\alpha_l) V_l^\dagger \left(\dots \left(\partial^{k_m} / \partial \alpha_m^{k_m} \right) \left(U_m^\dagger(\alpha_m) V_m^\dagger O V_m U_m(\alpha_m) \right) \dots \right) V_l U_l(\alpha_l) \right)$ gives $\|A_l\|_\infty \leq (2 \|H_l\|_\infty)^{k_l} \|A_{l+1}\|_\infty$. Finally, using it recursively starting from Eq. (C125) leads to the desired result

$$\|\partial^k ((U^\dagger(\alpha) O U(\alpha))\|_\infty = \|A_1\|_\infty \quad (\text{C135})$$

$$\leq (2 \|H_1\|_\infty)^{k_1} \|A_2\|_\infty \quad (\text{C136})$$

$$\leq \dots \quad (\text{C137})$$

$$\leq \|O\|_\infty \prod_{l=1}^m (2 \|H_l\|_\infty)^{k_l}, \quad (\text{C138})$$

where we used the convention that $A_{m+1} = O$ in the last step. Now, we can simply upper bound the norm of each generator by the largest one, i.e., by $\max_l \|H_l\|_\infty$, which finally leads to

$$|\partial^k f(\alpha)| \leq \|O\|_\infty \prod_{l=1}^m (2 \|H_l\|_\infty)^{k_l} \quad (\text{C139})$$

$$\leq \|O\|_\infty \left(2 \max_l \|H_l\|_\infty \right)^k. \quad (\text{C140})$$

■

Finally, we can prove Corollary C1 as follows.

Proof. Now, from the assumption that each generator has a spectral norm of order $\mathcal{O}(1)$, then there exist a constant $\gamma \in \mathcal{O}(1)$ such that $2 \|H_l\|_\infty \leq \gamma$ for each l . Therefore, we recover Assumption C1 using Proposition C4, which allows us to apply Supplemental Theorem C1. Therefore, the bounded assumption for the observable i.e., $\|O\|_\infty \in \mathcal{O}(1)$ together with Supplemental Theorem C1 completes the proof of the corollary. ■

7. Application to circuits with Pauli generators

In this appendix we apply our surrogation technique to the special case when the circuit generators are Pauli strings. Furthermore, we will also explore the connection between the Taylor expansion based surrogate at the heart of Theorem 1 and the Pauli propagation surrogation we

will discuss in Appendix E. Concretely, we will focus on a circuit with the following general structure:

$$U(\alpha) = \prod_{j=1}^m V_j U_j(\alpha_j), \quad (\text{C141})$$

where V_j are a set of fixed unitary matrices (but not necessarily Clifford) and $U_j(\alpha_j) = e^{-i(\alpha_j/2)P_j}$ are parametrized rotations with P_j Pauli strings. We will further suppose that the parameters α_j are uncorrelated.

Under the previous constraints, we first can compute the Fourier expansion of the expectation function. Indeed, as it will be explicitly shown below (Appendix C7a), the expectation function can be naturally expressed in terms of sine/cosine functions as

$$f(\alpha) = \sum_{\omega \in \{0,1,2\}^m} \Upsilon_\omega(\delta) \text{Tr}[O_\omega(\alpha^*) \rho_0], \quad (\text{C142})$$

where $\delta = \alpha - \alpha^*$ with $\alpha \in \mathcal{V}_m(\alpha^*, r)$ a hypercube of the form of Eq. (3) sampled uniformly as in Eq. (4). Here $\omega = (\omega_1, \omega_2, \dots, \omega_m)$ is an index vector with each ω_j taking a value of 0, 1, or 2 and the Fourier coefficients are given by

$$\Upsilon_\omega(\delta) = \prod_{i=1}^m v_{\omega_j}(\delta_j), \quad \text{where} \quad (\text{C143})$$

$$v_{\omega_j}(\delta_j) = \begin{cases} \cos^2\left(\frac{\delta_j}{2}\right) & \text{if } \omega_j = 0, \\ 2 \sin\left(\frac{\delta_j}{2}\right) \cos\left(\frac{\delta_j}{2}\right) & \text{if } \omega_j = 1, \\ \sin^2\left(\frac{\delta_j}{2}\right) & \text{if } \omega_j = 2. \end{cases} \quad (\text{C144})$$

The operators $O_\omega(\alpha^*)$ are defined recursively and can be computed efficiently. The recursive method to obtain them is detailed in Eq. (C173) in Appendix C7a.

For an individual term in the sum associated with ω , denote $n_1(\omega)$ and $n_2(\omega)$ as the number of times $\omega_j = 1$ and $\omega_j = 2$ appear in a given ω . Then we can define the order of each term as

$$K(\omega) = n_1(\omega) + 2n_2(\omega). \quad (\text{C145})$$

Notice that this count the number of sine contributions to the corresponding coefficient. We can then consider an order κ surrogate of the expectation function that is built by keeping only terms with $K(\omega) \leq \kappa$. That is, we have

$$\tilde{f}_\kappa(\alpha) := \sum_{\omega; K(\omega) \leq \kappa} \Upsilon_\omega(\delta) \text{Tr}[O_\omega(\alpha^*) \rho_0]. \quad (\text{C146})$$

This surrogate is roughly equivalent to the Taylor surrogate [i.e., Eq. (C6)]. Indeed, for a given order, the error of the two techniques. Trivially, we can compare the surrogate via Taylor and the Taylor expansion of the sin / cos

surrogate. We see that these are equivalent, as the terms dropped in the sin/cos surrogate are also dropped in the Taylor one (this comes from the fact that $\cos \alpha \sim 1, \sin \alpha \sim \alpha$). Therefore, the only difference between the two surrogate methods lies in the difference between the term $\Upsilon(\delta)$ and its Taylor expansion at order κ . This difference is of order $\kappa + 1$.

Now that we have established the connection between the Taylor and the sin/cos surrogation techniques, we devote the rest of this appendix to explaining how the latter connects with the Pauli Propagation technique. To do so, let us start by assuming that the measurement operator is given by a single Pauli, i.e., $O = P$ (we will later generalize to the case of $O = \sum_i a_i P_i$). Furthermore, let us now assume that the V_i gates are Clifford gates.

Pauli Propagation (which we will detail further in Appendix E) is understood in the Heisenberg picture. Here, we study the backward evolved measurement operator

$$U^\dagger(\alpha)PU(\alpha) = \prod_{j=1, l=1}^m V_j^\dagger U_j^\dagger(\alpha_j)PU_l(\alpha_l)V_l. \quad (\text{C147})$$

Let us start by using Eq. (C157) to compute the effect of applying the layer closest to the Pauli observable. For a general observable this would look like

$$e^{i\frac{\alpha_m}{2}P_m}Pe^{-i\frac{\alpha_m}{2}P_m} = v_0(\alpha_m)P + v_1(\alpha_m)\frac{i}{2}[P_m, P] + v_2(\alpha_m)P_mPP_m. \quad (\text{C148})$$

However, Pauli strings either commute or anti-commute. Therefore, we can use this to our advantage. Indeed, if $P_mP = PP_m$ we can trivially simplify the previous expression to

$$e^{i\frac{\alpha_m}{2}P_m}Pe^{-i\frac{\alpha_m}{2}P_m} = [v_0(\alpha_m) + v_2(\alpha_m)]P = P. \quad (\text{C149})$$

Furthermore, in the case that $P_mP = -PP_m$, we can also simplify the terms in Eq. (C148) using $(i/2)[P_m, P] = iP_mP$ and $P_mPP_m = -P$ to give

$$e^{i\frac{\alpha_m}{2}P_m}Pe^{-i\frac{\alpha_m}{2}P_m} = [v_0(\alpha_m) - v_2(\alpha_m)]P + v_1(\alpha_m)iP_mP, \quad (\text{C150})$$

which by simple trigonometric identities (and by recalling Eq. (C144)) can be simplified to $v_0(\alpha_m) - v_2(\alpha_m) = \cos(\alpha_m)$, and $v_1(\alpha_m) = \sin(\alpha_m)$. This means that Eq. (C148) can be rewritten as

$$e^{i\frac{\alpha_m}{2}P_m}Pe^{-i\frac{\alpha_m}{2}P_m} = \begin{cases} P & \text{iff}[P_m, P] = 0, \\ \cos(\alpha_m)P + \sin(\alpha_m)P' & \text{iff}\{P_m, P\}, \end{cases} \quad (\text{C151})$$

where $P' := iP_mP$ is a Pauli matrix (up to a sign). We can now define the Pauli Path coefficients defined in Eq. (E19)

that we rewrite here as

$$\phi_{\omega_i}(\alpha_i) = \begin{cases} 1, & \text{if } \omega_i = 0, \\ \cos(\alpha_i), & \text{if } \omega_i = 1, \\ \sin(\alpha_i), & \text{if } \omega_i = -1. \end{cases} \quad (\text{C152})$$

Thus we obtain

$$e^{i\frac{\alpha_m}{2}P_m}Pe^{-i\frac{\alpha_m}{2}P_m} = \begin{cases} \phi_0(\alpha_m)P = P & \text{iff}[P_m, P] = 0, \\ \phi_1(\alpha_m)P + \phi_2(\alpha_m)P' & \text{iff}\{P_m, P\}. \end{cases} \quad (\text{C153})$$

Now it is worth noting that when we apply any Clifford gate into any Pauli matrix, it returns a Pauli matrix. Therefore, we can repeat this procedure for all the circuit in Eq. (C147) to obtain

$$U^\dagger(\alpha)PU(\alpha) = \sum_{\omega} \Phi_{\omega}(\alpha)P_{\omega}, \quad (\text{C154})$$

where P_{ω} are the propagated Pauli matrices. In other words, the final Pauli matrices obtained after applying the commutation rules for all of the gates in the circuit.

Finally, it is worth stressing that if $O = \sum_i a_i P_i$ instead of a single Pauli, due to the linearity of the expectation function the results still apply. Indeed, if we start from such observable, the expectation function can be written as

$$f(\alpha) = \text{Tr}[\rho U^\dagger(\alpha)OU(\alpha)] = \sum_i a_i \text{Tr}[\rho U^\dagger(\alpha)P_i U(\alpha)], \quad (\text{C155})$$

and we can apply the propagation technique for each P_i . This previous leads to

$$f(\alpha) = \sum_i \sum_{\omega} a_i \Phi_{\omega}(\alpha) \text{Tr}[\rho P_{\omega}] = \sum_{\omega} c_{\omega}(\alpha) \text{Tr}[\rho P_{\omega}], \quad (\text{C156})$$

where we have introduced the notation $c_{\omega}(\alpha) = \sum_i a_i \Phi_{\omega}(\alpha)$.

As the error associated with this Pauli propagation surrogation strategy are of the same order as the Taylor surrogate [i.e., Eq. (C6)], Theorem 1 can be used to indirectly bound the efficiency of Pauli Propagation. However, substantially tighter bounds can be obtained by analyzing Pauli propagation directly as we will proceed to do in the next full appendix (Appendix E).

a. Fourier expansion

Let us now study the Fourier expansion in Eq. (C142). We will explicitly show that the expectation function

can naturally be expressed in terms of sine and cosine functions. In particular, since $P_j^2 = \mathbb{1}$, we find

$$e^{-i\frac{\alpha_j}{2}P_j} = \cos\left(\frac{\alpha_j}{2}\right)\mathbb{1} - i\sin\left(\frac{\alpha_j}{2}\right)P_j. \quad (\text{C157})$$

In addition, the rotation gates can be re-expressed as perturbations α around any point $\alpha \in \mathcal{V}_m(\alpha^*, r)$ for all j . This

can be trivially seen by just writing $\alpha = \alpha^* + \delta$,

$$U(\alpha) = \prod_{j=1}^m U_j(\delta_j)U_j(\alpha_j^*)V_j \quad (\text{C158})$$

$$= \prod_{j=1}^m U_j(\delta_j)\tilde{V}_j, \quad (\text{C159})$$

where the first equality holds due to $e^{-i(\alpha_j/2)P_j} = e^{-i(\alpha_j^*/2)P_j}e^{-i(\delta_j/2)P_j}$ and in the second equality we denote $\tilde{V}_j := V_j e^{-i(\alpha_j^*/2)P_j}$. Also, we introduce the notation

$$\rho_l = \prod_{j=1}^l U_j(\alpha_j)\tilde{V}_j \rho_0 \left(\prod_{j=1}^l U_j(\alpha_j)\tilde{V}_j \right)^\dagger \quad (\text{C160})$$

Now, the expectation function as defined in Eq. (1) can be expanded as

$$f(\alpha) = \text{Tr}[\rho_{m-1}\tilde{V}_m^\dagger U_m^\dagger(\delta_m)O U_m(\delta_m)\tilde{V}_m] \quad (\text{C161})$$

$$= \text{Tr}\left[\rho_{m-1}\tilde{V}_m^\dagger \left(\cos\left(\frac{\delta_m}{2}\right)\mathbb{1} + i\sin\left(\frac{\delta_m}{2}\right)P_m\right)O \left(\cos\left(\frac{\delta_m}{2}\right)\mathbb{1} - i\sin\left(\frac{\delta_m}{2}\right)P_m\right)\tilde{V}_m\right] \quad (\text{C162})$$

$$= \cos^2\left(\frac{\delta_m}{2}\right)\text{Tr}[\rho_{m-1}(\tilde{V}_m^\dagger O \tilde{V}_m)] + 2\sin\left(\frac{\delta_m}{2}\right)\cos\left(\frac{\delta_m}{2}\right)\text{Tr}\left[\rho_{m-1}\left(\tilde{V}_m^\dagger \frac{i}{2}[P_m, O]\tilde{V}_m\right)\right] \quad (\text{C163})$$

$$+ \sin^2\left(\frac{\delta_m}{2}\right)\text{Tr}[\rho_{m-1}(\tilde{V}_m^\dagger P_m O P_m \tilde{V}_m)] \quad (\text{C164})$$

$$= v_0(\delta_m)\text{Tr}[\rho_{m-1}O_0] + v_1(\delta_m)\text{Tr}[\rho_{m-1}O_1] + v_2(\delta_m)\text{Tr}[\rho_{m-1}O_2], \quad (\text{C165})$$

where we defined

$$O_0 = \tilde{V}_m^\dagger O \tilde{V}_m, \quad (\text{C166})$$

$$O_1 = \tilde{V}_m^\dagger \frac{i}{2}[P_m, O]\tilde{V}_m, \quad (\text{C167})$$

$$O_3 = \tilde{V}_m^\dagger P_m O P_m \tilde{V}_m \quad (\text{C168})$$

and

$$v_0(\delta_m) = \cos^2\left(\frac{\delta_m}{2}\right), \quad (\text{C169})$$

$$v_1(\delta_m) = 2\sin\left(\frac{\delta_m}{2}\right)\cos\left(\frac{\delta_m}{2}\right), \quad (\text{C170})$$

$$v_2(\delta_m) = \sin^2\left(\frac{\delta_m}{2}\right). \quad (\text{C171})$$

By iteratively repeating the previous steps, we obtain the form of the expectation function by Fourier expansion as

$$f(\alpha) = \sum_{\omega \in \{0,1,2\}^m} \Upsilon_\omega(\delta)\text{Tr}[O_\omega(\alpha^*)\rho_0], \quad (\text{C172})$$

where $\omega = (\omega_1, \omega_2, \dots, \omega_m)$ is an index vector with each ω_j taking a value of 0, 1, or 2. We also introduce a notation $\omega_{(i,j)} = (\omega_i, \omega_{i+1}, \dots, \omega_j)$ with $i \leq j$ and $\omega = \omega_{(1,m)}$. Now, the operators $O_{\omega_{(1,m)}} := O_\omega(\alpha^*)$ are defined recursively by

$$O_{\omega_{(j,m)}} = \begin{cases} \tilde{V}_j^\dagger O_{\omega_{(j+1,m)}} \tilde{V}_j & \text{if } \omega_j = 0, \\ \frac{i}{2} \tilde{V}_j^\dagger [P_j, O_{\omega_{(j+1,m)}}] \tilde{V}_j & \text{if } \omega_j = 1, \\ \tilde{V}_j^\dagger P_j O_{\omega_{(j+1,m)}} P_j \tilde{V}_j & \text{if } \omega_j = 2, \end{cases} \quad (\text{C173})$$

with $\tilde{V}_j = e^{-i(\alpha_j^*/2)P_j} V_j$, as defined above, and the Fourier coefficients are given by $\Upsilon_\omega(\delta) = \prod_{j=1}^m \nu_{\omega_j}(\delta_j)$, where

$$\nu_{\omega_j}(\delta_j) = \begin{cases} \cos^2\left(\frac{\delta_j}{2}\right) & \text{if } \omega_j = 0, \\ 2 \sin\left(\frac{\delta_j}{2}\right) \cos\left(\frac{\delta_j}{2}\right) & \text{if } \omega_j = 1, \\ \sin^2\left(\frac{\delta_j}{2}\right) & \text{if } \omega_j = 2. \end{cases} \quad (\text{C174})$$

APPENDIX D: ITERATIVE TRAINING STRATEGY WITH PATCH SURROGATE

In this appendix, we study an iterative training procedure for parameterized quantum circuits based on patch surrogate. Here, instead of employing a standard hybrid classical-quantum optimization strategy where a loss gradient is estimated at each training iteration, we construct a surrogate of loss gradients around a patch and then perform several classical gradient steps within this patch. After few training iterations, we need to reset the surrogate around a patch. We investigate whether this iterative strategy can be advantageous. The appendix is structured as follows:

- (a) In Appendix D 1, we provide a detailed description of the iterative surrogate approach.
- (b) In Appendix D 2, we numerically compare the surrogate with the standard gradient descent.
- (c) In Appendix D 3, further numerical details are provided including a detailed description of our shot allocation strategy.
- (d) Lastly, in Appendix D 4, the analytical results relating shot noise and shot allocation are derived.

1. Description of a surrogate approach

We consider a quantum surrogate-based strategy where the loss gradients are surrogated and then trained purely classically. More specifically, we consider a Taylor surrogate of the loss gradients $\nabla f(\alpha)$ around a point α^* and train this surrogate classically using gradient descent (GD) for a small number of steps. To ensure the faithfulness of the training, the process of resurrogating the loss gradients is

then repeated using the newly obtained point as the next expansion center.

For simplicity, we consider a second-order Taylor expansion of the gradient of the form

$$\nabla \tilde{f}(\alpha) = \nabla f(\alpha^*) + H(\alpha^*)(\alpha - \alpha^*), \quad (\text{D1})$$

where $H(\alpha^*)$ is the Hessian of the loss at the expansion point. The individual components read

$$\partial_l \tilde{f}(\alpha) = \partial_l f(\alpha^*) + \sum_{k=1}^m \partial_{l,k} f(\alpha^*) \delta_k, \quad \delta_k = \alpha_k - \alpha_k^*. \quad (\text{D2})$$

We note that it is straightforward to extend this construction to higher-order surrogate models. Algorithm 1 summarizes the overall protocol. Moreover, whereas other parts of our work primarily focus on surrogating the loss itself, in practical gradient-based training it is often more appropriate to surrogate the loss gradient directly. A related strategy, termed Quantum Analytic Descent, was introduced in Ref. [31]; it is based on a second-order truncation of the sine coefficients in the Fourier expansion (see Appendix C 7 a).

Since standard GD is unbiased as opposed to the surrogate strategy which suffers from Taylor error, we are interested in comparing the performance on both methods by fixing an equivalent total shot budget. This is obtained by choosing

$$N_{\text{grad}} \approx \frac{N_{\text{surrogate}}}{N_{\text{iters}}}, \quad (\text{D3})$$

where N_{grad} is the number of measurement shots for one iteration of standard GD, N_{iters} is the total number of training iterations for standard GD, and $N_{\text{surrogate}}$ is the total number of measurement shots used in the surrogate training strategy. In this case, the shot noise with standard GD will increase by a factor N_{iters} , the number of iterations between surrogate updates. We aim to study how, in the task of estimating the ground state of a given Hamiltonian,

ALGORITHM 1. Iterative Training via Taylor Surrogate.

Input: Initial parameters α_1 , number of surrogates N_{surr} , steps per surrogate N_{iters}

for $i = 1$ **to** N_{surr} **do**

if $i = 1$ **then**

 Set $\alpha^* = \alpha_1$

else

 Set $\alpha^* = \alpha_i$

end

 Construct Taylor surrogate $\nabla \tilde{f}(\alpha)$ around α^* by estimating $\nabla f(\alpha^*)$ and $H(\alpha^*)$

 Perform N_{iters} steps of classical gradient descent with $\nabla \tilde{f}(\alpha)$ starting from α^*

 Let α_{i+1} be the final iterate

end

the additional shot noise incurred by standard gradient-descent balances against the Taylor truncation error introduced by the surrogate, and how this trade-off impacts overall optimization performance.

2. Summary of the results: Comparison study with standard gradient descent

In this appendix, we empirically investigate the performance of the surrogate approach and compare its resource consumption with that of the standard VQA approach.

a. Setup

We consider the task of estimating the ground state of a Hamiltonian. The target Hamiltonian is a rescaled Heisenberg chain given by

$$O = \frac{1}{3(n-1)} \sum_{i=1}^{n-1} (X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}). \quad (\text{D4})$$

The variational *Ansatz* consists of L layers of single-qubit rotations followed by two-qubit ZZ gates of the form

$$U(\alpha) = \prod_{l=1}^L \left(\prod_{i=1}^n R_{ZZ}(\alpha_{i+2n,l}) \prod_{i=1}^n R_Z(\alpha_{i+n,l}) R_X(\alpha_{i,l}) \right), \quad (\text{D5})$$

with $|\psi(\alpha)\rangle = U(\alpha) |0\rangle^{\otimes n}$. The total number of parameters is $m = 3nL$.

We explore different strategies for allocating measurement shots for the surrogate; details are provided in Appendix D3 a. One approach is straightforward and easily generalizable to higher-order expansions, while the second is slightly more optimized for practical implementation. The key difference is that the optimized strategy evaluates each loss only once for repeated Hessian diagonal elements, whereas the naive approach recomputes the same loss separately for each diagonal element. In Appendix D3 a, we also compare these methods with standard GD under an equivalent total shot budget. Additionally, we employ a two-phase training schedule for the surrogate, where the number of iterations between surrogate updates, N_{iters} , is adjusted once during the simulation (see Appendix D3 b).

b. Results discussion

a. Numerical details in simulations. Let *Surrogate 1* and *Surrogate 2* denote the naive and the optimized surrogate strategies, respectively, and let *standard GD* denote standard gradient descent with an equivalent quantum resource budget. As detailed in Appendix D3 a, our shot allocation is fully determined by fixing the number of shots per

loss evaluation, F_2 - that is, the total number of shots to compute the gradient is given by $2mF_2$. In our experiments, we set $F_2 = 600$. During the first training phase, we use $N_{\text{iters}}^{(1)} = 10$ and $N_{\text{sur}}^{(1)} = 20$. In the second phase, we set $N_{\text{iters}}^{(2)} = 8$ and $N_{\text{sur}}^{(2)} = 200$, corresponding to a total of 1800 GD steps. At each GD step, the test loss is computed as the difference between the exact expectation value of the observable and the ground state energy obtained numerically using *scipy.sparse.linalg.eigsh*. All methods are run independently 10 times with different random initial parameters. Importantly, the three methods share the same initial parameters in each run. The learning rate is fixed at $\eta = 0.1$.

b. Results and discussion. In Fig. 6, we show the test loss as a function of the number of shots [at depths (a) $L = 3$ and (c) $L = 6$] and the number of iterations [(b) $L = 3$ and (d) $L = 6$] for a 10-qubit Heisenberg Hamiltonian. We selected examples where the best final test loss is achieved either by standard GD ($L = 6$) or by Surrogate 2 ($L = 3$); however, the differences are small, particularly in terms of the interquartile range (IQR). We also compute the number of shots required to reduce the test loss below 10^{-1} for $L = 2, 3, 4, 5, 6$. The results are shown in Fig. 7 for $n = 8$ (a) and $n = 10$ (b) qubits. Again, there is no clear distinction in performance between the strategies.

The empirical results observed from Figs. 6 and 7 show that GD and the surrogate-based approaches exhibit comparable performance in terms of quantum resource usage, although both methods achieve good accuracy for the tasks considered. Intuitively, while standard GD suffers from larger shot noise, its stochastically unbiased nature compensates for this drawback. Nevertheless, this does not rule out the possibility that more sophisticated surrogate approaches could yield a distinct performance advantage. Indeed, early evidence for a Fourier-based surrogate method has been reported in Ref. [31]. A plausible explanation is that such approaches preserve the periodicity of the loss function and retain partial contributions from higher-order perturbative terms, although this remains a hypothesis.

From a practical perspective, the key distinction is that GD consumes quantum resources at every step, whereas the surrogate expends a larger quantum budget up front and then reuses the resulting information across multiple iterations. This can be advantageous in two separate ways. First, it can better tolerate noise drift on real devices: by amortizing quantum evaluations into fewer, more substantial data-collection phases, it reduces the need to repeatedly query the hardware under slowly changing calibration conditions. As a result, successive optimization updates are less exposed to time-dependent variations in gate errors and measurement bias that can otherwise introduce

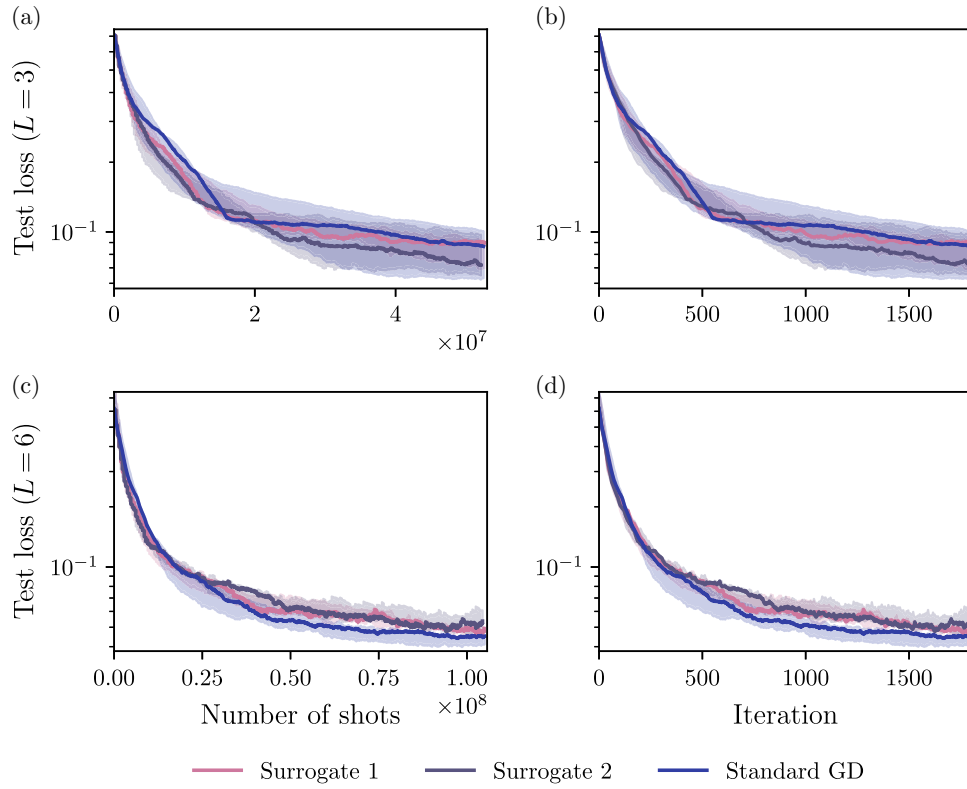


FIG. 6. Test loss during training for the 10-qubit Heisenberg Hamiltonian as a function of quantum shots (left) and iterations (right) for $L = 3$ (top) and $L = 6$ (bottom). Surrogate 1 and 2 correspond to the naive and optimized shot allocation, respectively, while standard gradient descent uses an equivalent shot budget. Solid lines show the median test loss over 10 runs, and the shaded area indicates the interquartile range.

inconsistent gradient estimates. Second, on cloud-based quantum platforms, it can mitigate the impact of variable queuing delays and inflated wall-clock runtimes that arise when jobs are repeatedly submitted.

c. Further numerical experiment: Training inside a single patch

We now focus on a single surrogate and investigate how many GD steps can be performed from random initial

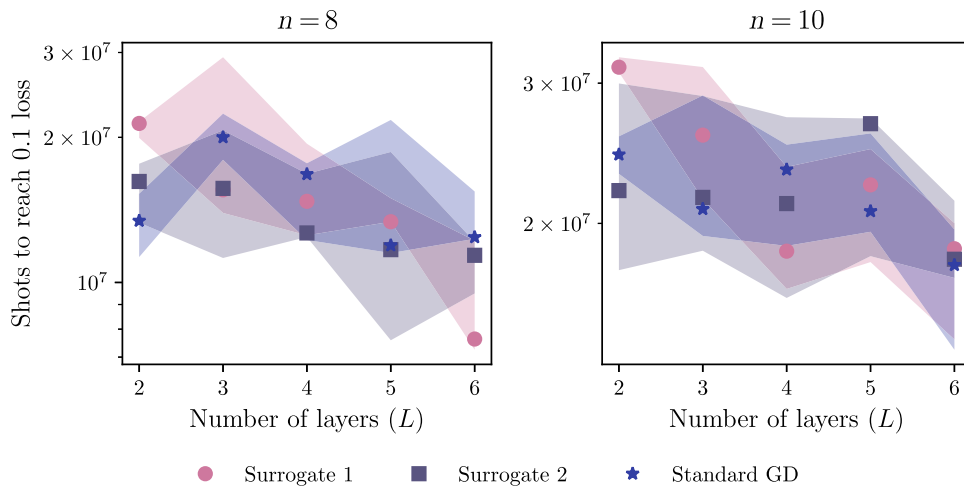


FIG. 7. Number of shots required to reach $f(\alpha) \leq 0.1$ as a function of L for (a) $n = 8$ and (b) $n = 10$ qubits in the Heisenberg Hamiltonian. Surrogate 1 and 2 correspond to the naive and optimized shot allocation, respectively. Standard gradient descent uses an equivalent shot budget. Shaded areas indicate the interquartile range.

parameters before the test loss begins to increase. The goal is to identify practically relevant values for N_{iters} .

a. Simulation details. We fix $F_2 = 600$ and $\eta = 0.1$ as in previous simulations and consider the naive surrogate method. The gradient is surrogated around random initial parameters α^* , and GD is performed classically until the test loss starts to increase. We record the maximal value of N_{iters} for each run. We set $n = L$ and perform $n_{\text{run}} = 20$ independent runs for each $n = L \in \{2, 3, \dots, 10\}$.

b. Results and discussion. The results are shown in Fig. 8, which plots the optimal number of iterations N_{iters} as a function of the number of parameters m . The shaded area corresponds to the minimum-maximum range over all $n_{\text{run}} = 20$ random initializations, the diamond markers indicate the mean, and the solid line shows a power-law fit (linear fit in log-log scale). The minimal value observed is $N_{\text{iters}} = 9$, suggesting that the choice of $N_{\text{iters}} = 8$ –10 in previous simulations was reasonable. These values were set heuristically based on preliminary tests. In practice, the test loss would not be accessible and so how to obtain the optimal N_{iters} remains an open problem. The corresponding distances to the center of the patch are shown in Fig. 3. One practical approach for selecting the number of iterations could be to restrict parameters within the patch of radius $r \in \mathcal{O}(1/\sqrt{m})$, where theoretical guarantees exist. However, Fig. 3 indicates that the surrogate can be trained beyond this patch, meaning such a restriction would be suboptimal.

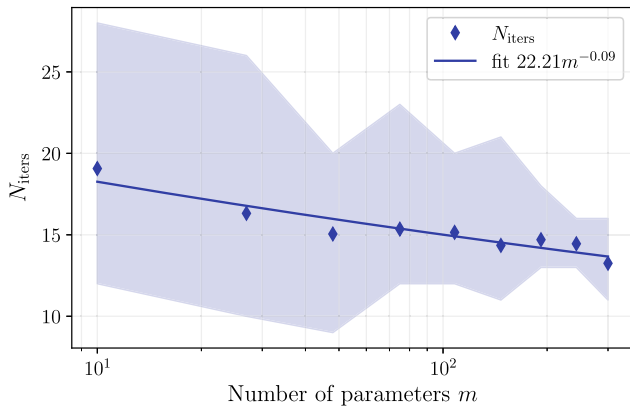


FIG. 8. Maximal number of iterations that can be performed using the surrogate before the test loss increases (while the surrogate loss continues to decrease). Shaded area shows the min-max range over $n_{\text{run}} = 20$ different initial random parameters, diamond markers indicate the mean, and the solid line is a power-law fitting curve (linear fit in log-log scale).

3. Further numerical details for Taylor surrogate

a. Shot allocation in simulation

To account for finite sampling noise, we assign a number of measurement shots $N(\delta\alpha)$ to each loss evaluation $f(\alpha^* + \delta\alpha)$. Instead of explicitly sampling loss evaluations according to the optimal randomized measurement distribution of Lemma B3, we adopt a deterministic shot-allocation scheme. Specifically, we approximate the optimal randomized measurement process by fixing the shot allocation to its average over the optimal sampling distribution, which preserves the same asymptotic scaling where the local patch size scales as $r \sim \mathcal{O}(1/m)$ while being marginally simpler to implement numerically. We fix a reference number of shots F_2 for gradient evaluations [i.e., for shifts of the form $\pm(\pi/2)e_l$] and define the remaining allocations proportionally,

$$F_3 \equiv \frac{F_2}{m}, \quad (\text{Hessian-related shifts}), \quad (\text{D6})$$

$$F_1 \equiv \frac{F_2}{\sqrt{m}}, \quad (\text{central-point evaluation}). \quad (\text{D7})$$

This reflects the intuition that larger offsets or higher-order terms can be estimated more coarsely.

a. (1) Naive strategy (without periodicity). In this straightforward approach, all loss evaluations required by the parameter-shift rules are treated independently. In particular, the Hessian is estimated using four separate evaluations for each element,

$$f\left(\alpha^* \pm \frac{\pi}{2}(e_l \pm e_k)\right), \quad k, l = 1, \dots, m. \quad (\text{D8})$$

For $k = l$, this includes multiple evaluations of $f(\alpha^*)$. Denoting $N_{l,i}(0)$ as the number of shots used for the two independent central-point evaluations corresponding to the l th diagonal term ($i = 1, 2$), we set $N_{l,i}(0) = F_3$ for simplicity. The total number of shots required to construct one surrogate is then,

$$N_{\text{surrogate}}^{\text{naive}} = 2mF_2 + 4 \cdot \frac{m(m+1)}{2}F_3 \quad (\text{D9})$$

$$= 2mF_2 + 2m(m+1)F_3. \quad (\text{D10})$$

Substituting $F_3 = F_2/m$ gives

$$N_{\text{surrogate}}^{\text{naive}} = (4m + 2)F_2. \quad (\text{D11})$$

b. (2) *Optimized strategy (with periodicity).* By exploiting the periodicity property,

$$f(\alpha + 2\pi e_l) = f(\alpha), \quad (\text{D12})$$

the diagonal Hessian elements can be computed more efficiently,

$$\partial_l^2 f(\alpha^*) = \frac{f(\alpha^* + \pi e_l) - f(\alpha^*)}{2}. \quad (\text{D13})$$

Each diagonal element requires only one additional loss evaluation, while the central point $f(\alpha^*)$ is reused across all terms. The total number of shots per surrogate is then,

$$N_{\text{surrogate}}^{\text{opt}} = F_1 + 2mF_2 + F_3 \left[m + 4 \cdot \frac{m(m-1)}{2} \right] \quad (\text{D14})$$

$$= F_1 + 2mF_2 + F_3 m(2m-1). \quad (\text{D15})$$

Substituting $F_1 = F_2/\sqrt{m}$ and $F_3 = F_2/m$ gives

$$N_{\text{surrogate}}^{\text{opt}} = \left(4m - 1 + \frac{1}{\sqrt{m}} \right) F_2. \quad (\text{D16})$$

c. *Summary.* Both strategies scale similarly as $\mathcal{O}(mF_2)$, but the optimized strategy is slightly more shot-efficient due to reuse of the central evaluation and compression of diagonal Hessian evaluations. Nevertheless, the naive approach can slightly reduce shot noise in practice, resulting in comparable performance in terms of quantum resources.

d. *Standard GD.* Standard GD requires estimating the gradient at each iteration, which involves $2m$ loss evaluations via the parameter-shift rule. Let F_2^{GD} denote the number of shots per evaluation. To ensure a fair comparison with the surrogate methods under an equivalent total shot budget, we set

$$F_2^{\text{GD}} \approx \frac{N_{\text{surrogate}}^{\text{naive}}}{2mN_{\text{iters}}} = \frac{(2 + \frac{1}{m})F_2}{N_{\text{iters}}}. \quad (\text{D17})$$

b. Two-phase surrogate schedule

We adopt a two-phase training schedule in which the frequency of surrogate updates changes over the course of optimization. Empirically, this heuristic improves the performance of the surrogate. Concretely, we define

- (a) $N_{\text{surr}}^{(1)}$: number of surrogate refreshes in the first phase,
- (b) $N_{\text{iters}}^{(1)}$: number of classical gradient steps between surrogates in the first phase,

- (c) $N_{\text{surr}}^{(2)}$: number of surrogate refreshes in the second phase,
- (d) $N_{\text{iters}}^{(2)}$: number of classical gradient steps between surrogates in the second phase,

with the typical choice

$$N_{\text{iters}}^{(1)} > N_{\text{iters}}^{(2)}, \quad N_{\text{surr}}^{(1)} < N_{\text{surr}}^{(2)}.$$

This adaptive schedule balances exploration and local accuracy: larger patches in the early phase encourage movement out of flat regions, while smaller patches in the later phase ensure the Taylor surrogate remains accurate near convergence.

When comparing the surrogate strategy to standard GD under an equivalent total shot budget, we need to adapt Eq. (D17) to account for the two different values of N_{iters} . Specifically, we enforce

$$N_{\text{surrogate}}^{\text{naive}}(N_{\text{surr}}^{(1)} + N_{\text{surr}}^{(2)}) = 2mF_2^{\text{GD1}} \left(N_{\text{surr}}^{(1)}N_{\text{iters}}^{(1)} + N_{\text{surr}}^{(2)}N_{\text{iters}}^{(2)} \right). \quad (\text{D18})$$

Since this equation may yield a noninteger value for F_2^{GD} , we round it as

$$F_2^{\text{GD}} = \left\lfloor (2 + 1/m)F_2 \frac{N_{\text{surr}}^{(1)} + N_{\text{surr}}^{(2)}}{N_{\text{surr}}^{(1)}N_{\text{iters}}^{(1)} + N_{\text{surr}}^{(2)}N_{\text{iters}}^{(2)}} \right\rfloor. \quad (\text{D19})$$

This is equivalent to Eq. (D17) with N_{iters} replaced by an effective average over the simulation.

4. Analytical guarantees

We now analyze how shot noise propagates through the surrogate-based *gradient* estimation (not the shot noise analysis for the Taylor surrogate of the loss directly was covered in Appendix D2). To understand the impact of statistical uncertainty, we construct a Taylor expansion of the surrogate gradient $\tilde{\nabla} f(\alpha)$ around a chosen expansion point α^* ,

$$\tilde{\nabla} f(\alpha) = \nabla f(\alpha^*) + H(\alpha^*)(\alpha - \alpha^*), \quad (\text{D20})$$

where $H(\alpha^*)$ represents the Hessian of the loss function. Decomposing this into individual components for each parameter l , we obtain

$$\partial_l \tilde{f}(\alpha) = \partial_l f(\alpha^*) + \sum_{k=1}^m \partial_{l,k} f(\alpha^*) \delta_k, \quad \delta_k = \alpha_k - \alpha_k^*. \quad (\text{D21})$$

To estimate these derivatives on a quantum processor, we employ the parameter-shift rule. This approach allows us to compute exact gradients and second-order derivatives by evaluating the circuit at specific shifts,

$$\partial_l f(\alpha^*) = \frac{f(\alpha^* + \frac{\pi}{2} \mathbf{e}_l) - f(\alpha^* - \frac{\pi}{2} \mathbf{e}_l)}{2}, \quad (\text{D22})$$

$$\partial_{l,k} f(\alpha^*) = \frac{f(\alpha^* + \frac{\pi}{2}(\mathbf{e}_l + \mathbf{e}_k)) - f(\alpha^* + \frac{\pi}{2}(\mathbf{e}_l - \mathbf{e}_k)) - f(\alpha^* - \frac{\pi}{2}(\mathbf{e}_l - \mathbf{e}_k)) + f(\alpha^* - \frac{\pi}{2}(\mathbf{e}_l + \mathbf{e}_k))}{4}. \quad (\text{D23})$$

We define the shot noise error on the full gradient vector as the expected L_2 -norm distance between the empirical estimate and the exact surrogate gradient (its mean),

$$\mathbb{E} \left[\left\| \nabla \tilde{f}(\alpha) - \mathbb{E} [\nabla \tilde{f}(\alpha)] \right\|_2^2 \right] = \sum_{l=1}^m \text{Var}[\partial_l \tilde{f}(\alpha)]. \quad (\text{D24})$$

Assuming independent sampling noise for each evaluation, we propagate the variance through the parameter-shift expressions. Summing the contributions from the first-order shifts, the Hessian diagonals, and the off-diagonal cross-terms, we obtain

$$\begin{aligned} & \sum_{l=1}^m \text{Var}[\partial_l \tilde{f}(\alpha)] \\ &= \sum_{l=1}^m \frac{\text{Var}[f(\alpha^* + \frac{\pi}{2} \mathbf{e}_l)] + \text{Var}[f(\alpha^* - \frac{\pi}{2} \mathbf{e}_l)]}{4} \\ &+ \sum_{l=1}^m \sum_{k=1}^m \delta_k^2 \frac{\sum_{\pm, \pm} \text{Var}[f(\alpha^* \pm \frac{\pi}{2} \mathbf{e}_l \pm \frac{\pi}{2} \mathbf{e}_k)]}{16}. \end{aligned} \quad (\text{D25})$$

In the following, we perform the shot-noise analysis over these different scenarios:

- (a) Periodicity shot allocation for patch surrogate,
- (b) Naive shot allocation for patch surrogate, without periodicity,
- (c) Comparison to standard GDs.

a. Periodicity shot allocation for patch surrogate

Here, we analyze the shot allocation in Eq. (D25) using the periodicity condition. That is, by utilizing the periodicity of the loss, $f(\alpha + 2\pi \mathbf{e}_k) = f(\alpha)$, we can simplify the diagonal elements of the Hessian, reducing the number of required quantum circuit evaluations,

$$\partial_{l,l}^2 f(\alpha^*) = \frac{f(\alpha^* + \pi \mathbf{e}_l) - f(\alpha^*)}{2}. \quad (\text{D26})$$

Let $N(\delta\alpha)$ be the number of shots allocated for measuring $f(\alpha^* + \delta\alpha)$. The variance of the estimate is bounded

by the single-shot variance

$$\text{Var}[f(\alpha^* + \delta\alpha)] = \frac{\text{Var}_1[f(\alpha^* + \delta\alpha)]}{N(\delta\alpha)} \leq \frac{\|O\|_\infty^2}{N(\delta\alpha)}, \quad (\text{D27})$$

where $\text{Var}_1[\cdot]$ is upper bounded by the squared norm of the observable O ,

$$\text{Var}_1[f(\alpha)] = \langle \psi(\alpha) | O^2 | \psi(\alpha) \rangle - \langle \psi(\alpha) | O | \psi(\alpha) \rangle^2 \quad (\text{D28})$$

$$\leq \langle \psi(\alpha) | O^2 | \psi(\alpha) \rangle \quad (\text{D29})$$

$$\leq \|O\|_\infty^2. \quad (\text{D30})$$

By inserting Eq. (D27) and assuming a maximum patch radius $|\delta_l| \leq r$ into Eq. (D25), we find

$$\begin{aligned} & \frac{\sum_{l=1}^m \text{Var}[\partial_l \tilde{f}(\alpha)]}{\|O\|_\infty^2} \\ & \leq \sum_{l=1}^m \left(\frac{1}{4N(\frac{\pi}{2} \mathbf{e}_l)} + \frac{1}{4N(-\frac{\pi}{2} \mathbf{e}_l)} + \frac{r^2}{4N(\pi \mathbf{e}_l)} \right) + \frac{mr^2}{4N(0)} \\ & + \sum_{l=1}^m \sum_{k=l+1}^m \frac{r^2}{8} \left(\sum_{\pm, \pm} \frac{1}{N(\pm \frac{\pi}{2}(\mathbf{e}_l \pm \mathbf{e}_k))} \right). \end{aligned} \quad (\text{D31})$$

Applying the shot allocation (F_1, F_2, F_3) denoted in Appendix D 3 a, the expression simplifies to

$$\begin{aligned} & \frac{\sum_{l=1}^m \text{Var}[\partial_l \tilde{f}(\alpha)]}{\|O\|_\infty^2} \leq \frac{m}{2F_2} + \frac{m^2 r^2}{4F_3} + \frac{mr^2}{4F_1} \\ & = \frac{m}{2F_2} \left(1 + \frac{m^2 r^2}{2} + \frac{\sqrt{m} r^2}{2} \right). \end{aligned} \quad (\text{D32})$$

Considering a scaling of $r \in \mathcal{O}(1/m)$, the shot noise scales as

$$\sum_{l=1}^m \text{Var}[\partial_l \tilde{f}(\alpha)] \in \mathcal{O} \left(\frac{\|O\|_\infty^2 m}{F_2} \right) = \mathcal{O} \left(\frac{\|O\|_\infty^2 m^2}{N_{\text{surrogate}}^{\text{opti}}} \right). \quad (\text{D33})$$

b. Naive shot allocation for patch surrogate without periodicity

If we ignore periodicity and recompute the loss at the fixed point α^* whenever it is required, the diagonal Hessian

elements require four independent evaluations each. In this naive case, the total variance becomes

$$\begin{aligned} & \sum_{l=1}^m \text{Var}[\partial \tilde{f}(\alpha)] \\ &= \sum_{l=1}^m \frac{\text{Var}[f(\alpha^* + \frac{\pi}{2} \mathbf{e}_l)] + \text{Var}[f(\alpha^* - \frac{\pi}{2} \mathbf{e}_l)]}{4} \\ &+ \sum_{l=1}^m \sum_{k=1}^m \delta_k^2 \cdot \frac{\sum_{\pm} \sum_{\pm} \text{Var}[f(\alpha^* \pm \frac{\pi}{2} \mathbf{e}_l \pm \frac{\pi}{2} \mathbf{e}_k)]}{16}. \end{aligned} \quad (\text{D34})$$

Applying the same bounds and the allocation from Appendix D3 a, the error bound is

$$\frac{\sum_{l=1}^m \text{Var}[\partial \tilde{f}(\alpha)]}{\|O\|_{\infty}^2} \leq \frac{m}{2F_2} + \frac{m^2 r^2}{4F_3} = \frac{m}{2F_2} \left(1 + \frac{m^2 r^2}{2}\right). \quad (\text{D35})$$

While the bound is slightly lower, the total number of shots required is significantly higher. Thus, for $r \sim 1/m$, the shot noise maintains the same asymptotic scaling

$$\sum_{l=1}^m \text{Var}[\partial \tilde{f}(\alpha)] \in \mathcal{O} \left(\frac{\|O\|_{\infty}^2 m^2}{N_{\text{naive surrogate}}^{\text{naive}}} \right). \quad (\text{D36})$$

c. Comparison to standard GD

For standard GD, we assume that each of the $2m$ loss evaluations for the gradient is estimated using F_2^{GD} shots. The total variance is upper bounded by

$$\sum_{l=1}^m \text{Var}[\partial f(\alpha)] \leq \frac{\|O\|_{\infty}^2 m}{2F_2^{\text{GD}}}. \quad (\text{D37})$$

Recalling that $F_2^{\text{GD}} \approx (2/N_{\text{iters}})F_2$, we see that the variance for a standard gradient evaluation is approximately $N_{\text{iters}}/2$ times larger than that of the surrogate gradient

$$\sum_{l=1}^m \text{Var}[\partial f(\alpha)] \in \mathcal{O} \left(\frac{\|O\|_{\infty}^2 m N_{\text{iters}}}{F_2} \right). \quad (\text{D38})$$

APPENDIX E: SMALL ANGLE PAULI PROPAGATION

This appendix is devoted to the classical simulation and surrogation of quantum circuits in the near-Clifford regime.

1. The Pauli propagation framework

We start by briefly reviewing the Pauli-path formalism and how it can be employed for classically estimating expectation values of quantum circuits. For further details, see also Refs. [2,14,26,114] and references therein.

Let $U(\alpha)$ be a quantum circuit written as parameterized perturbations from Clifford circuits. That is, of the form

$$U(\alpha) = \prod_{i=1}^m C_i U_i(\alpha_i), \quad (\text{E1})$$

where $\{C_i\}_{i=1}^m$ are a set of fixed Clifford operators and $\{U_i(\alpha_i) = e^{-i\alpha_i P_i/2}\}_{i=1}^m$ are parameterized rotations generated by the Pauli operators $\{P_i\}_{i=1}^m$.

Given an observable $O = \sum_{P \in \mathcal{P}_n} a_P P$ and an initial state ρ , we want to approximate the following expectation value:

$$f(\alpha) = \text{Tr}[OU(\alpha)\rho U(\alpha)^{\dagger}]. \quad (\text{E2})$$

For the sake of simplicity, we will consider observables consisting of a single Pauli term. In the following lemma, we show that this assumption comes with no loss of generality.

Lemma E1 (Transferring guarantees from Pauli observables). Let $\mathcal{S} \subseteq \{I, X, Y, Z\}^{\otimes n}$ be a subset of Pauli operators and let $O = \sum_{P \in \mathcal{S}} a_P P$ be an observable.

- (1) **Worst-case error.** For all $P \in \mathcal{S}$ let $\tilde{f}^{(P)}(\alpha)$ be a function satisfying

$$|\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^{\dagger}]| \leq \epsilon. \quad (\text{E3})$$

Then the function $\tilde{f}(\alpha) = \sum_{P \in \mathcal{S}} a_P \tilde{f}^{(P)}(\alpha)$ satisfies the following:

$$|\tilde{f}(\alpha) - f(\alpha)| \leq \|a\|_1 \epsilon. \quad (\text{E4})$$

- (2) **Mean Square Error.** Given a distribution $\mathcal{D}$ over $[-\pi, \pi]^m$, for all $P \in \mathcal{S}$ let $\tilde{f}^{(P)}(\alpha)$ be a function satisfying

$$\mathbb{E}_{\alpha \sim \mathcal{D}} |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^{\dagger}]|^2 \leq \epsilon^2. \quad (\text{E5})$$

Then the function $\tilde{f}(\alpha) = \sum_{P \in \mathcal{S}} a_P \tilde{f}^{(P)}(\alpha)$ satisfies the following:

$$\mathbb{E}_{\alpha \sim \mathcal{D}} |\tilde{f}(\alpha) - f(\alpha)|^2 \leq \|a\|_1^2 \epsilon^2. \quad (\text{E6})$$

Proof. The first part of the lemma follows by a simple application of the triangle inequality,

$$|\tilde{f}(\alpha) - f(\alpha)| = \left| \sum_{P \in \mathcal{S}} a_P (\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^{\dagger}]) \right| \quad (\text{E7})$$

$$\leq \sum_{P \in \mathcal{S}} |a_P| |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^{\dagger}]| \leq \|a\|_1 \epsilon. \quad (\text{E8})$$

As for the second part, we have

$$|\tilde{f}(\alpha) - f(\alpha)|^2 = \left| \sum_{P \in \mathcal{S}} a_P (\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]) \right|^2 \quad (\text{E9})$$

$$\leq \left(\sum_{P \in \mathcal{S}} |a_P| |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]| \right)^2 \quad (\text{E10})$$

$$= \left(\sum_{P \in \mathcal{S}} \sqrt{|a_P|} \sqrt{|a_P|} |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]| \right)^2 \quad (\text{E11})$$

$$\leq \left(\sum_{P \in \mathcal{S}} |a_P| \right) \left(\sum_{P \in \mathcal{S}} |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]|^2 \right) \quad (\text{E12})$$

$$= \|\mathbf{a}\|_1 \left(\sum_{P \in \mathcal{S}} |a_P| |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]|^2 \right), \quad (\text{E13})$$

where in the second step we used the triangle inequality and in the fourth step we used the Cauchy-Schwarz inequality—i.e., $(\sum_i x_i y_i)^2 \leq (\sum_i x_i^2)(\sum_i y_i^2)$. We conclude the proof by taking the expectation over α ,

$$\mathbb{E}_{\alpha \sim \mathcal{D}} |\tilde{f}(\alpha) - f(\alpha)|^2 = \mathbb{E}_{\alpha \sim \mathcal{D}} \|\mathbf{a}\|_1 \left(\sum_{P \in \mathcal{S}} |a_P| |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]|^2 \right) \quad (\text{E14})$$

$$= \|\mathbf{a}\|_1 \left(\sum_{P \in \mathcal{S}} |a_P| \underbrace{\mathbb{E}_{\alpha \sim \mathcal{D}} |\tilde{f}^{(P)}(\alpha) - \text{Tr}[PU(\alpha)\rho U(\alpha)^\dagger]|^2}_{\leq \epsilon^2} \right) \quad (\text{E15})$$

$$\leq \|\mathbf{a}\|_1^2 \epsilon^2, \quad (\text{E16})$$

where in the second step we used the fact that the expectation is linear. ■

Thus, given a Pauli operator P , we consider the backpropagated (i.e., Heisenberg-evolved) observable $U(\alpha)^\dagger P U(\alpha)$ and express it in the Pauli-path formalism

$$U(\alpha)^\dagger P U(\alpha) = \sum_{\omega} \Phi_{\omega}(\alpha) P_{\omega}, \quad (\text{E17})$$

where $\omega \in \{-1, 0, 1\}^m$ is the so-called *frequency vector* or *path vector* that uniquely enumerates a Pauli path, and P_{ω} is the backpropagated Pauli along the path ω . If at the Pauli gate U_i in Eq. (5) the backpropagating Pauli operator commutes with P_i , then $\omega_i = 0$. If it anticommutes, $\omega_i = 1$ or $\omega_i = -1$ depending on whether the Pauli path picked up a cosine or sine coefficient, respectively. $\Phi_{\omega}(\alpha)$ thus represents the coefficient of the backpropagated Pauli path and is defined as

$$\Phi_{\omega}(\alpha) = \prod_{i=1}^m \phi_{\omega_i}(\alpha_i), \quad (\text{E18})$$

where

$$\phi_{\omega_i}(\alpha_i) = \begin{cases} 1, & \text{if } \omega_i = 0 \\ \cos(\alpha_i), & \text{if } \omega_i = 1 \\ \sin(\alpha_i), & \text{if } \omega_i = -1 \end{cases}. \quad (\text{E19})$$

Furthermore, we define

$$d_{\omega} := \text{Tr}[\rho P_{\omega}] \in [-1, 1] \quad (\text{E20})$$

as the overlap of the backpropagated Pauli along the path ω with the initial state, and we denote the expectation function as

$$f^{(P)}(\alpha) = \text{Tr}[PU(\alpha)\rho U^\dagger(\alpha)] = \sum_{\omega} \Phi_{\omega}(\alpha) d_{\omega}. \quad (\text{E21})$$

We additionally define

$$\begin{aligned} \mathcal{S}_{\kappa}^m &:= \{\text{paths } \omega \mid \omega \text{ contains more than } \kappa \text{ negative entries}\} \\ &\subseteq \{-1, 0, 1\}^m \end{aligned} \quad (\text{E22})$$

as the set of paths ω with more than κ entries of -1 , i.e., the set of paths ω with more than κ sine coefficients. For simplicity, we define

$$\#\sin(\omega) := \text{number of negative entries of } \omega, \quad (\text{E23})$$

which is also the number of sine coefficients of the path ω . Finally, we introduce the truncated function $\tilde{f}_\kappa^{(P)}(\alpha)$ which contains only the terms with at most κ sines

$$\tilde{f}_\kappa^{(P)}(\alpha) := \sum_{\omega \notin S_\kappa^m} \Phi_\omega(\alpha) d_\omega. \quad (\text{E24})$$

Similarly, we will use the notation $\tilde{P}(\alpha)$ to denote the truncated version of the backpropagated operator P given by

$$\tilde{P}(\alpha) := \sum_{\omega \notin S_\kappa^m} \Phi_\omega(\alpha) P_\omega. \quad (\text{E25})$$

In the following, we will consider two regimes:

- (a) **Average case.** When the parameter vector α is equipped with a suitable distribution $\mathcal{D}$ over the hypercube $\mathcal{V}_m(\mathbf{0}, r)$, we consider the mean square error

$$\begin{aligned} & \mathbb{E}_{\alpha \sim \mathcal{D}} \left[(f^{(P)}(\alpha) - \tilde{f}_\kappa^{(P)}(\alpha))^2 \right] \\ &= \mathbb{E}_{\alpha \sim \mathcal{D}} \left[\left(\sum_{\omega \in S_\kappa^m} \Phi_\omega(\alpha) d_\omega \right)^2 \right]. \end{aligned} \quad (\text{E26})$$

- (b) **Worst case.** When the parameter vector α is an arbitrarily chosen point in the hypercube $\mathcal{V}_m(\mathbf{0}, r)$, we consider the worst-case error

$$\begin{aligned} & \sup_{\alpha \in \mathcal{V}_m(\mathbf{0}, r)} |f^{(P)}(\alpha) - \tilde{f}_\kappa^{(P)}(\alpha)| \\ &= \sup_{\alpha \in \mathcal{V}_m(\mathbf{0}, r)} \left| \sum_{\omega \in S_\kappa^m} \Phi_\omega(\alpha) d_\omega \right|. \end{aligned} \quad (\text{E27})$$

2. Uncorrelated parameters

In this appendix we consider the case where all the parameters α_i are sampled independently from some probability distributions $\mathcal{D}_i$ supported in $[-\pi, \pi]$, and therefore the parameter vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$ is sampled from the product distribution $\mathcal{D} := \mathcal{D}_1 \times \mathcal{D}_2 \times \dots \times \mathcal{D}_m$. For all i , we make following key assumptions on the distribution $\mathcal{D}_i$.

- (a) **Symmetric distribution.** We assume $\mathcal{D}_i$ to be symmetric and centered around zero, i.e., for all

$$0 \leq a \leq b \leq \pi,$$

$$\Pr_{\alpha_i \sim \mathcal{D}_i} [\alpha_i \in [a, b]] = \Pr_{\alpha_i \sim \mathcal{D}_i} [\alpha_i \in [-b, -a]]. \quad (\text{E28})$$

This assumption ensures that the odd function $\cos(\alpha_i) \sin(\alpha_i)$ has mean zero

$$\mathbb{E}_{\alpha_i \sim \mathcal{D}_i} \cos(\alpha_i) \sin(\alpha_i) = 0. \quad (\text{E29})$$

- (b) **Variance bound.** The variance of the function $\sin(\alpha_i)$ is bounded by an appropriately chosen value σ^2

$$\mathbb{E}_{\alpha_i \sim \mathcal{D}_i} \sin^2(\alpha_i) \leq \sigma^2. \quad (\text{E30})$$

In particular, these two assumptions are satisfied by a truncated Gaussian distribution with mean 0 and variance σ^2 supported in $[-\pi, \pi]$, and by any symmetric distribution supported entirely in the range $[-\sqrt{3}\sigma, \sqrt{3}\sigma]$.

Crucially, the symmetry assumption ensures that the different paths are “orthogonal,” in the sense that the associated coefficients are uncorrelated

$$\omega' \neq \omega \implies \mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_\omega(\alpha) \Phi_{\omega'}(\alpha) = 0. \quad (\text{E31})$$

This occurs when for any i one has $\omega_i = 1, \omega'_i = -1$, or vice versa. Given that all paths stem from the same Pauli operator in the Pauli decomposition of the observable O , two different paths must have split off from another, collecting opposing coefficients. This implies that $\mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_\omega(\alpha) \Phi_{\omega'}(\alpha) = C_{\omega \setminus \omega_i} \cdot \mathbb{E}_{\alpha_i \sim \mathcal{D}_i} \cos(\alpha_i) \sin(\alpha_i) = 0$, where $C_{\omega \setminus \omega_i}$ are the integrals over all other independent parameters $\alpha_j, j \neq i$.

The above observation allows us to drastically simplify the expression of the mean square error,

$$\mathbb{E}_{\alpha \sim \mathcal{D}} \left(\sum_{\omega \in S_\kappa^m} \Phi_\omega(\alpha) d_\omega \right)^2 = \sum_{\omega \in S_\kappa^m} \mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_\omega^2(\alpha) d_\omega^2 \quad (\text{E32})$$

$$\leq \sum_{\omega \in S_\kappa^m} \mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_\omega^2(\alpha), \quad (\text{E33})$$

where we used the fact that $d_\omega^2 = \text{Tr}[P_\omega \rho] \leq 1$ for all ω . Leveraging the variance bound, we can upper bound the mean square error as follows.

Supplemental Theorem E1 (Mean square error for uncorrelated parameters). Let $\mathcal{D}$ satisfy the assumptions stated above with $\sigma^2 \leq \kappa/m$. Then we have

$$\mathbb{E}_{\alpha \sim \mathcal{D}} \left[(f^{(P)}(\alpha) - \tilde{f}_\kappa^{(P)}(\alpha))^2 \right] \leq \left(\frac{em\sigma^2}{\kappa} \right)^\kappa. \quad (\text{E34})$$

Proof. We start by observing that

$$\mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_{\omega}^2(\alpha) = \prod_{i=1}^m \begin{cases} 1 & \text{if } \omega_i = 0 \\ \mathbb{E}_{\alpha_i \sim \mathcal{D}_i} \cos(\alpha_i)^2 \leq 1 & \text{if } \omega_i = 1 \\ \mathbb{E}_{\alpha_i \sim \mathcal{D}_i} \sin(\alpha_i)^2 \leq \sigma^2 & \text{if } \omega_i = -1 \end{cases} \quad (\text{E35})$$

In order to upper bound the mean square error, we need to estimate how many sine terms are contained in each path ω . Unfortunately, this is highly problem-dependent. For a generic upper bound, we thus choose a maximally splitting problem, where any truncation induces the largest loss in probability mass. That is, we assume that the splitting tree yields 2^m paths with all combinations of sin and cos prefactors. We can thus write

$$\begin{aligned} \mathbb{E}_{\alpha \sim \mathcal{D}} \left[(f^{(P)}(\alpha) - \tilde{f}_{\kappa}^{(P)}(\alpha))^2 \right] \\ \leq \sum_{\omega \in S_{\kappa}^m} \mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_{\omega}^2(\alpha) \leq \sum_{\substack{\omega \in \{1, -1\}^m \\ \# \sin(\omega) \geq \kappa}} \mathbb{E}_{\alpha \sim \mathcal{D}} \Phi_{\omega}^2(\alpha) \quad (\text{E36}) \end{aligned}$$

$$\leq \sum_{i=\kappa}^m \binom{m}{i} (\sigma^2)^i \leq \left(\frac{em\sigma^2}{\kappa} \right)^{\kappa}, \quad (\text{E37})$$

where $\# \sin(\omega)$ is defined in Eq. (E23) and the last step follows from Lemma A4 since by assumption $\sigma^2 \leq \kappa/m$. ■

a. Uniformly sampled parameters. We now consider the special case of particular interest when all the parameters α_i are sampled independently from the uniform distribution over the symmetric interval $[-r, r]$, where $r > 0$ is a suitable small constant.

Averaging over the small angle range $[-r, r]$ leads to

$$\begin{aligned} \mathbb{E}_{\alpha_i \sim \mathcal{D}_i} \sin(\alpha_i)^2 &= \frac{1}{2r} \int_{-r}^r \sin(\alpha_i)^2 d\alpha_i \\ &= \frac{1}{2} \left(1 - \frac{\sin(2r)}{2r} \right) \leq \frac{r^2}{3}. \quad (\text{E38}) \end{aligned}$$

From the previous, we can upper bound the mean square error by setting $\sigma^2 = r^2/3$ and invoking Supplemental Theorem E1.

Corollary E1 (Mean square error for the uniform distribution). Let $\mathcal{D}$ be the uniform distribution over $[-r, r]^m$, where $r^2 \leq 3\kappa/m$. Then we have

$$\mathbb{E}_{\alpha \sim \mathcal{D}} (f^{(P)}(\alpha) - \tilde{f}_{\kappa}^{(P)}(\alpha))^2 \leq \left(\frac{emr^2}{3\kappa} \right)^{\kappa}. \quad (\text{E39})$$

In particular, for any $r \leq \sqrt{1/m}$, this bound decreases faster than exponentially with κ .

3. Worst-case analysis

In this appendix, we consider the parameter vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$ to be arbitrarily chosen—possibly adversarially—in the hypercube $[-r, r]^m$. Thus, we drop the “symmetry assumption” and we do not assume the different α_i to be independently sampled.

We consider the maximum error over the hypercube

$$\max_{\alpha \in [-r, r]^m} |f^{(P)}(\alpha) - \tilde{f}_{\kappa}^{(P)}(\alpha)|. \quad (\text{E40})$$

We provide the following upper bound.

Supplemental Theorem E2 (Worst-case error). Let $r \leq \kappa/m$. Then we have

$$\max_{\alpha \in [-r, r]^m} |f^{(P)}(\alpha) - \tilde{f}_{\kappa}^{(P)}(\alpha)| \leq \left(\frac{emr}{\kappa} \right)^{\kappa}. \quad (\text{E41})$$

Thus, for any $r \leq 1/m$, the absolute error decreases faster than exponentially in κ .

Proof. We start by upper bounding the error as follows:

$$|f^{(P)}(\alpha) - \tilde{f}_{\kappa}^{(P)}(\alpha)| = \left| \sum_{\omega \in S_{\kappa}^m} \Phi_{\omega}(\alpha) d_{\omega} \right| \leq \sum_{\omega \in S_{\kappa}^m} |\Phi_{\omega}(\alpha)|, \quad (\text{E42})$$

where we used the triangle inequality and the fact $|d_{\omega}| = \text{Tr}[P_{\omega}\rho] \leq 1$.

Proceeding as in the proof of Supplemental Theorem E1, we obtain

$$\sum_{\omega \in S_{\kappa}^m} |\Phi_{\omega}(\alpha)| \leq \sum_{\substack{\omega \in \{1, -1\}^m \\ \# \sin(\omega) \geq \kappa}} |\Phi_{\omega}(\alpha)| \quad (\text{E43})$$

$$\leq \sum_{\substack{\omega \in \{1, -1\}^m \\ \# \sin(\omega) \geq \kappa}} r^{\# \sin(\omega)} = \sum_{i=\kappa}^m \binom{m}{i} r^i \leq \left(\frac{emr}{\kappa} \right)^{\kappa}, \quad (\text{E44})$$

where we used the fact that $|\sin(\alpha_i)| \leq r$, $|\cos(\alpha_i)| \leq 1$ and we applied Lemma A4 in the last step. Recall that $\# \sin(\omega)$ is defined in Eq. (E23). Notice that the proof is also valid if the α_i 's are correlated and provides also a bound for the 1-norm of the coefficients of the paths. ■

4. Circuits with correlated parameters

Given the apparent difference in asymptotic scalings between the average case of parameters randomly sampled from a distribution and the worst case, we can ask the question: *How hard is it to simulate fundamentally interesting circuits?* Here we consider the example of Trotter circuits for simulating quantum dynamics. The scenario is that of perfectly correlated angles, as the same parameter is used

across all of the gates. This is the case when all coefficients in a Hamiltonian are equal. Clearly, our earlier worst-case error also bounds this case, but we will now consider the *average case of correlated parameters*.

Proposition E1 (Improved bound for correlated parameters). Let $\mathcal{D}$ the distribution over $\mathbf{V}_m(\boldsymbol{\alpha}^*, r)$ obtained by sampling a parameter α_1 uniformly at random over $[-r, r]$ and setting $\boldsymbol{\alpha} = (\alpha_1, \alpha_1, \dots, \alpha_1)$. Then the average error satisfies the following upper bound:

$$\mathbb{E}_{\boldsymbol{\alpha}} |f^{(P)}(\boldsymbol{\alpha}) - \tilde{f}_{\kappa}^{(P)}(\boldsymbol{\alpha})| \leq \frac{\left(\frac{emr}{\kappa}\right)^{\kappa}}{\kappa + 1}. \quad (\text{E45})$$

Proof. Here we assume a quantum circuit consisting of Clifford and Pauli gates, where the Pauli gates are all parametrized by the identical angle α_1 . The average error is

$$\begin{aligned} & \frac{1}{2r} \int_{-r}^r |f^{(P)}(\boldsymbol{\alpha}) - \tilde{f}_{\kappa}^{(P)}(\boldsymbol{\alpha})| d\alpha_1 \\ &= \frac{1}{2r} \int_{-r}^r \left| \sum_{\boldsymbol{\omega} \in \mathcal{S}_{\kappa}^m} \Phi_{\boldsymbol{\omega}}(\boldsymbol{\alpha}) d_{\boldsymbol{\omega}} \right| d\alpha_1 \end{aligned} \quad (\text{E46})$$

$$\leq \frac{1}{2r} \int_{-r}^r \sum_{\boldsymbol{\omega} \in \mathcal{S}_{\kappa}^m} |\Phi_{\boldsymbol{\omega}}(\boldsymbol{\alpha}) d_{\boldsymbol{\omega}}| d\alpha_1 \quad (\text{E47})$$

$$\leq \sum_{\boldsymbol{\omega} \in \mathcal{S}_{\kappa}^m} \frac{1}{2r} \int_{-r}^r |\alpha_1|^{\#\sin(\boldsymbol{\omega})} d\alpha_1 \quad (\text{E48})$$

$$= \sum_{\boldsymbol{\omega} \in \mathcal{S}_{\kappa}^m} \frac{1}{r} \int_0^r \alpha_1^{\#\sin(\boldsymbol{\omega})} d\alpha_1, \quad (\text{E49})$$

where we used the triangle inequality in the first inequality, and $|\sin(\alpha_1)| \leq |\alpha_1|$, $|\cos(\alpha_1)| \leq 1$, $d_{\boldsymbol{\omega}}^2 \leq 1$ for the second inequality with the same notation for the number of sines per path as in Eq. (E23). In the last equality we use the fact that the integral is symmetric over the absolute value α_1 , to compute the integral over a positive value range. Again assuming a maximally splitting tree behavior as the worst-case yields

$$\sum_{\substack{\boldsymbol{\omega} \in \{1, -1\}^m \\ \#\sin(\boldsymbol{\omega}) \geq \kappa}} \frac{1}{r} \int_0^r \alpha_1^{\#\sin(\boldsymbol{\omega})} d\alpha_1 \leq \sum_{i=\kappa}^m \binom{m}{i} \frac{1}{r} \int_0^r \alpha_1^i d\alpha_1 \quad (\text{E50})$$

$$= \sum_{i=\kappa}^m \binom{m}{i} \frac{r^i}{i+1} \quad (\text{E51})$$

$$\leq \sum_{i=\kappa}^m \binom{m}{i} \frac{r^i}{\kappa+1}, \quad (\text{E52})$$

where in the end we used that $1/(i+1) \leq 1/(\kappa+1)$ for $i \geq \kappa$. Finally, we can use Lemma A4 to get

$$\frac{1}{2r} \int_{-r}^r |f^{(P)}(\boldsymbol{\alpha}) - \tilde{f}_{\kappa}^{(P)}(\boldsymbol{\alpha})| d\alpha_1 \leq \frac{\left(\frac{emr}{\kappa}\right)^{\kappa}}{\kappa+1}, \quad (\text{E53})$$

which corresponds to the upper bound of the worst-case multiplied by a factor $1/(\kappa+1)$ (see Supplemental Theorem E2). ■

Looking ahead to Appendix E3, we see that the error scaling is the same for the worst case as in the average case with fully correlated angles. Whether this is a feature of our bounds or indicative of something more fundamental is left as an open question.

5. Time complexity of the Pauli Propagation algorithm

Recall that the circuit *Ansatz* defined in Eq. (5) is of the form

$$U(\boldsymbol{\alpha}) = \prod_{i=1}^m C_i U_i(\alpha_i). \quad (\text{E54})$$

We start by upper bounding the time complexity of the Pauli Propagation algorithm as a function of the system size n , the number of Pauli rotations (and Clifford gates) m and the truncation order κ .

Upper bound on the number of paths In a quantum circuit with m Pauli rotations as in Eq. (5), the total number of Pauli paths $n_{pth}^{(m)}$ is upper bounded by $n_{pth}^{(m)} \leq 2^m$. This is achieved when the backpropagating Pauli operators split into two at every gate. As used throughout this appendix, the distribution of the number of sines in the 2^m paths is given via

$$n_{pth}^{(m)} \leq \sum_{i=0}^m \binom{m}{i} = 2^m, \quad (\text{E55})$$

where the index i represents the number of sines.

Given that we truncate paths with more than κ sines, the number of paths $n_{pth}^{(\kappa)}$ that we need to compute is upper bounded like

$$n_{pth}^{(\kappa)} \leq \sum_{i=0}^{\kappa} \binom{m}{i} \leq \left(\frac{em}{\kappa}\right)^{\kappa}, \quad (\text{E56})$$

as stated in Eq. (A7). *Runtime T for updating the Pauli operators:* The runtime of the Pauli propagation algorithm is necessarily upper bounded by a number proportional to the number of paths $n_{pth}^{(\kappa)}$. We call this proportionality factor T , and it includes the time complexity of updating a propagating Pauli operator when a gate is applied to it. More concretely, let V be either a Clifford C_i or a Pauli rotation

$U_i(\alpha_i)$. Given a Pauli operator $P \in \mathcal{P}_n$ we need to upper bound the time complexity for computing $V^\dagger P V$.

Up until now, we have assumed $T \in \mathcal{O}(1)$ for simplicity, but we here include the explicit dependence on T for preciseness. Interestingly, the case of $T \in \mathcal{O}(1)$ can be accurate under certain implementation and simulation assumptions. For example, $T \in \mathcal{O}(1)$ is possible if V acts on $\mathcal{O}(1)$ qubits and the Pauli operator P can be updated in-place on a subset of qubits. If in-place manipulation is not done, then the full operator needs to be copied which takes $\mathcal{O}(n)$ time. In general, global Clifford gates can be decomposed into $\mathcal{O}(n^2/\log(n))$ local Clifford gates [88,115], each taking $\mathcal{O}(1)$ or $\mathcal{O}(n)$ time (in-place or not). Furthermore, one could define C_i as an arbitrary long concatenation of local Clifford gates without recompilation, which can take an arbitrary large amount of time to execute. However, even global Pauli rotations can be applied in time $\mathcal{O}(n)$. Thus, in summary, in the best case we have $T \in \mathcal{O}(1)$ and in the worst $T \in \mathcal{O}(n^3)$. Given that we assumed $m \in \mathcal{O}(\text{poly}(n))$, the extra factor of n^3 can be absorbed by big $\mathcal{O}$ into our final time complexity. This justifies why so far we take $T \in \mathcal{O}(1)$.

Runtime for a truncation order κ

The total number of gates applied to all the paths is

$$mn_{pth}^{(\kappa)} \leq m \left(\frac{em}{\kappa} \right)^\kappa. \quad (\text{E57})$$

As we assumed that the time for computing the Pauli terms is T , this leads to a total runtime t of

$$t \leq Tmn_{pth}^{(\kappa)} \leq Tm \left(\frac{em}{\kappa} \right)^\kappa. \quad (\text{E58})$$

In the following, we provide an upper bound on the runtime required to achieve an error of ϵ , considering both the cases of uncorrelated and worst-case parameters. Moreover, we will now consider more general observable $O = \sum_{P \in \mathcal{P}_n} a_P P$ by combining previous analysis for single Paulis together with Lemma E1.

Supplemental Theorem E3 (Time complexity of small-angle Pauli propagation, uncorrelated angles, Formal). Let $O = \sum_{P \in \mathcal{P}_n} a_P P$ be an observable containing N_{Paulis} different Pauli terms and satisfying $\|\mathbf{a}\|_1 = \mathcal{O}(1)$. Furthermore, assume that given some P , we can efficiently estimate $\text{Tr}[\rho P]$. Let T be the time necessary to apply a gate to a propagating Pauli operator. Then small-angle Pauli propagation can be used to simulate a quantum circuit of the form of Eq. (5) over uniformly sampled zero-centered hypercube of width $2r$, i.e., $\mathcal{D}_m(\mathbf{0}, r)$ with $r \in \mathcal{O}(1/\sqrt{m})$. To ensure a mean square error

$$\mathbb{E}_\alpha \left[(f(\alpha) - \tilde{f}_\kappa(\alpha))^2 \right] \leq \|\mathbf{a}\|_1^2 \left(\frac{emr^2}{3\kappa} \right)^\kappa = \epsilon^2, \quad (\text{E59})$$

it suffices to run the algorithm for time

$$t = N_{\text{Paulis}} Tm \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} Tm^{\log(1/\epsilon^2)} \right). \quad (\text{E60})$$

Proof. To approximate the expectation value of the observable O , we will run the Pauli Propagation algorithm for all the Pauli operators in the Pauli expansion of O . Let $\tilde{t}$ be the time required to estimate the expectation value of a single Pauli observable P up to mean square error $\tilde{\epsilon}^2 = \epsilon^2 / \|\mathbf{a}\|_1^2$. Then, invoking Lemma E1, we find that the time required to estimate the expectation value of O up to mean square error ϵ^2 is

$$\tilde{t} \cdot N_{\text{Paulis}}. \quad (\text{E61})$$

Given $r \in \mathcal{O}(1/\sqrt{m})$, we introduce a constant $c \in \mathcal{O}(1)$ such that $r = c/\sqrt{m}$ and consider the distribution α randomly sampled from $\mathcal{D}_m(\mathbf{0}, c/\sqrt{m})$. Then, as long as the truncation order satisfies

$$\kappa \geq \frac{c^2}{3}. \quad (\text{E62})$$

Corollary E1 implies that, for any given Pauli operator P , the Pauli Propagation algorithm returns an operator $\tilde{P}_\alpha$, which is a weighted sum of Pauli operators, satisfying

$$\mathbb{E}_{\alpha \sim \mathcal{D}_m(\mathbf{0}, r)} \left| \text{Tr}[PU(\alpha)\rho U^\dagger(\alpha) - \tilde{P}_\alpha \rho] \right|^2 \leq \left(\frac{ec^2}{3\kappa} \right)^\kappa. \quad (\text{E63})$$

In order to achieve the desired error, we set

$$\tilde{\epsilon}^2 := \left(\frac{ec^2}{3\kappa} \right)^\kappa \quad (\text{E64})$$

$$\implies \kappa \log \kappa \in \Theta(\log(1/\tilde{\epsilon}^2)). \quad (\text{E65})$$

As observed in Eq. (E56), the number of paths found by the Pauli Propagation algorithm is at most

$$n_{pth} \leq \left(\frac{em}{\kappa} \right)^\kappa = m^\kappa \cdot \tilde{\epsilon}^2 \cdot \left(\frac{3}{c^2} \right)^\kappa \in \mathcal{O}(m^\kappa \tilde{\epsilon}^2). \quad (\text{E66})$$

Therefore, the runtime for a single Pauli operator in Eq. (E58) is at most

$$\tilde{t} \leq Tm \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O}(Tm^{\kappa+1} \tilde{\epsilon}^2) = Tm^{\mathcal{O}(\kappa)}. \quad (\text{E67})$$

We further observe that, for sufficiently large constant κ ,

$$\begin{aligned} m^{\mathcal{O}(\kappa)} &= m^{\frac{1}{\log(\kappa)} \mathcal{O}(\kappa \log(\kappa))} \\ &= m^{\frac{\mathcal{O}(\log(\|\mathbf{a}\|_1^2/\epsilon^2))}{\log(\kappa)}} = \mathcal{O} \left(m^{\log(1/\epsilon^2)} \right), \end{aligned} \quad (\text{E68})$$

where we recall that $\|\mathbf{a}\|_1^2 \in \mathcal{O}(1)$. Thus, we conclude that the runtime of the Pauli Propagation algorithm for the

observable O can be upper bounded by

$$t \leq N_{\text{Paulis}} \cdot \tilde{t} \in \mathcal{O}\left(N_{\text{Paulis}} T m^{\log(1/\epsilon^2)}\right). \quad (\text{E69})$$

■

Supplemental Theorem E4 (Time complexity of small-angle Pauli propagation, worst-case). Let $O = \sum_{P \in \mathcal{P}_n} a_P P$ be an observable containing N_{Paulis} different Pauli terms and satisfying $\|a\|_1 = \mathcal{O}(1)$. Furthermore, assume that given some P , we can efficiently estimate $\text{Tr}[\rho P]$. Let T be the time necessary to apply a gate to a propagating Pauli operator. Then small-angle Pauli propagation can be used to both classically simulate *and* surrogate a quantum circuit of the form of Eq. (5) over sampled zero-centered hypercube of width $2r$, i.e., $\mathcal{V}_m(\mathbf{0}, r)$ with $r \in \mathcal{O}(1/m)$. In particular, running the algorithm for time

$$t = N_{\text{Paulis}} T m \left(\frac{em}{\kappa}\right)^\kappa \in \mathcal{O}\left(N_{\text{Paulis}} T m^{\log(1/\epsilon)}\right), \quad (\text{E70})$$

we can ensure that

$$\max_{\alpha \in \mathcal{V}_m(\mathbf{0}, r)} |f(\alpha) - \tilde{f}_\kappa(\alpha)| \leq \|a\|_1 \left(\frac{erm}{\kappa}\right)^\kappa = \epsilon. \quad (\text{E71})$$

Proof. This proof is similar to that of Supplemental Theorem E3, but here we are considering the worst-case error upper bound. Let us rewrite some part of previous proof for completeness. First, we fix $c \in \mathcal{O}(1)$ such that $r = c/m$. Let $\tilde{t}$ be the time required to estimate the expectation value of a single Pauli observable P up to error $\tilde{\epsilon} = \epsilon / \|a\|_1$ for any points $\alpha \in \mathcal{V}_m(\mathbf{0}, c/m)$. Then, invoking Lemma E1, we find that the time required to estimate the expectation value of O up to error ϵ is

$$\tilde{t} \cdot N_{\text{Paulis}}. \quad (\text{E72})$$

Then, as long as the truncation order satisfies

$$\kappa \geq c. \quad (\text{E73})$$

Supplemental Theorem E2 implies that, for any given Pauli operator P , the Pauli Propagation algorithm returns an operator $\tilde{P}_\alpha$ satisfying

$$|\text{Tr}[PU(\alpha)\rho U^\dagger(\alpha) - \tilde{P}_\alpha \rho]| \leq \left(\frac{ec}{\kappa}\right)^\kappa. \quad (\text{E74})$$

In order to achieve the desired error, we set

$$\tilde{\epsilon} := \left(\frac{ec}{\kappa}\right)^\kappa \quad (\text{E75})$$

$$\implies \kappa \log \kappa \in \Theta(\log(1/\tilde{\epsilon})). \quad (\text{E76})$$

As observed in Eq. (E56), the number of paths found by the Pauli Propagation algorithm is at most

$$n_{\text{pth}} \leq \left(\frac{em}{\kappa}\right)^\kappa = m^\kappa \cdot \tilde{\epsilon} \cdot \left(\frac{1}{c}\right)^\kappa \in \mathcal{O}(m^\kappa \tilde{\epsilon}). \quad (\text{E77})$$

And therefore the runtime for a single Pauli operator in Eq. (E58) is at most

$$\tilde{t} \leq T m \left(\frac{em}{\kappa}\right)^\kappa \in \mathcal{O}(T m^{\kappa+1} \tilde{\epsilon}) = T m^{\mathcal{O}(\kappa)}. \quad (\text{E78})$$

We further observe that, for sufficiently large constant κ ,

$$m^{\mathcal{O}(\kappa)} = m^{\frac{1}{\log(\kappa)} \mathcal{O}(\kappa \log(\kappa))} = m^{\frac{\mathcal{O}(\log(\|a\|_1/\epsilon))}{\log(\kappa)}} = \mathcal{O}(m^{\log(1/\epsilon)}). \quad (\text{E79})$$

Thus, we conclude that the runtime of the Pauli Propagation algorithm for the observable O can be upper bounded by

$$t \leq N_{\text{Paulis}} \cdot \tilde{t} \in \mathcal{O}\left(N_{\text{Paulis}} T m^{\log(1/\epsilon)}\right). \quad (\text{E80})$$

■

An interesting corollary of our theorem is that in the cases where the worst-case results hold, our simulation method can reduce Trotter error in quantum dynamics simulations with “only” a polynomial slow-down. To reduce Trotter error, one needs to reduce dt and in turn inversely proportionally increases the number of layers L to leave the total simulation time dtL unchanged. That is, if we reduce dt by a factor $1/c$ we need to increase the number of layers by a factor of c , which in turn increases the number of parameters by a factor c . The maximum dt we can handle with small-angle Pauli Propagation with worst-case error guarantees is determined by r . As the error for small-angle Pauli propagation scales in the worst-case simulation as mr (this is easiest to see from our formal statement in Theorem E4) we see that reducing Trotter error by reducing dt by a factor $1/c$ and increasing the total number of layers leaves the final simulation error unchanged. However, the total simulation time increases polynomially in m and so we obtain a polynomial overhead in time complexity of small-angle Pauli Propagation resulting from the total number of parameters increasing by a factor of c . More crudely speaking, if our algorithm can accurately simulate to time t with large time step dt in polynomial time, then it can simulate to time t with small time step dt in polynomial time.

6. Sample complexity of the Pauli Propagation algorithm

In the previous section in the appendix, we demonstrated that Pauli Propagation enables efficient simulation of quantum circuits composed of Clifford gates and Pauli rotations

within a small angle range. To upper bound the computational complexity, we crucially assumed that the input state ρ is classically simulable, meaning that for any Pauli operator P , we can efficiently estimate $\text{Tr}[P\rho]$.

In the most general case, ρ may be an unknown state that we can access through quantum measurements. In this appendix, we combine the Pauli Propagation guarantee from Appendix E with the randomized measurement protocol from Appendix B, achieving a polynomial sample complexity for arbitrarily small constant error, i.e., $\epsilon \in \Theta(1)$. We further show that, if the final backpropagated Pauli operators have low Pauli weight, then the number of required Pauli measurements scales logarithmically with the system size.

a. The truncated Heisenberg-evolved observable

Pauli observables. Given a Pauli operator $Q \in \mathcal{P}_n$ and a circuit $U(\alpha)$ as input, the Pauli-path simulation algorithm returns the observable

$$\tilde{Q}(\alpha) = \sum_{\omega \notin S_\kappa^m} \Phi_\omega(\alpha) Q_\omega = \sum_{P \in \mathcal{P}_n} c_P^{(Q)}(\alpha) P, \quad (\text{E81})$$

where we merged different Pauli paths by introducing the parametrized coefficients

$$c_P^{(Q)}(\alpha) = \sum_{\substack{\omega \notin S_\kappa^m \\ \text{s. t. } Q_\omega = P}} \Phi_\omega(\alpha). \quad (\text{E82})$$

Recall that Q_ω in previous expressions is the Pauli obtained after backpropagating Q along the path ω . For our sample complexity analysis, we need to upper bound the effective 1-norms $\|c^{(Q)}\|_{1,\text{avg}}$ and $\|c^{(Q)}\|_{1,\text{worst}}$ introduced in Definition B1. We start by providing a loose, yet polynomial in the number of gates, upper bound. First, we notice that the average-case effective 1-norm is always upper bounded by the worst-case effective 1-norm. Let $\mathcal{D}$ be a distribution over the set $\mathcal{A}$. We have

$$\begin{aligned} \|c^{(Q)}\|_{1,\text{avg}} &= \sum_P \sqrt{\mathbb{E}_{\alpha \sim \mathcal{D}} [c_P^{(Q)}(\alpha)^2]} \\ &\leq \sum_P \max_{\alpha \in \mathcal{A}} |c_P^{(Q)}(\alpha)| = \|c^{(Q)}\|_{1,\text{worst}}. \end{aligned} \quad (\text{E83})$$

Furthermore, we can upper bound $\|c^{(Q)}\|_{1,\text{worst}}$ as done in the proof of Supplemental Theorem E2,

$$\|c^{(Q)}\|_{1,\text{avg}} \leq \|c^{(Q)}\|_{1,\text{worst}} \leq \left(\frac{emr}{\kappa}\right)^\kappa = \Theta((rm)^\kappa). \quad (\text{E84})$$

Indeed, this is obtained by applying triangle inequality on $|c_P^{(Q)}(\alpha)|$ in Eq. (E82) and upper bounding each $|\Phi_\omega(\alpha)|$ by $r^{\#\sin(\omega)}$ as done in Eq. (E44).

General observables. Let us consider a general observable expressed in the Pauli basis

$$O = \sum_{P \in \mathcal{P}_n} a_P P, \quad (\text{E85})$$

with $\|a\|_1 \in \mathcal{O}(1)$. The final backpropagated observables equals

$$\begin{aligned} \tilde{O}(\alpha) &= \sum_{Q \in \{I,X,Y,Z\}^{\otimes n}} a_Q \tilde{Q}(\alpha) = \sum_{Q,P \in \{I,X,Y,Z\}^{\otimes n}} a_Q c_P^{(Q)}(\alpha) P \\ &:= \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} c_P(\alpha) P, \end{aligned} \quad (\text{E86})$$

where we introduced the parametrized coefficients

$$c_P(\alpha) = \sum_{Q \in \{I,X,Y,Z\}^{\otimes n}} a_Q c_P^{(Q)}(\alpha). \quad (\text{E87})$$

Then, we can upper bound the effective norms as follows:

$$\|c\|_{1,\text{avg}} \leq \|c\|_{1,\text{worst}} \leq \|a\|_1 \max_Q \|c^{(Q)}\|_{1,\text{worst}} \in \Theta((rm)^\kappa). \quad (\text{E88})$$

Proof. From Eq. (E87), we can use triangle inequality to get

$$\max_{\alpha \in \mathcal{A}} |c_P(\alpha)| \leq \max_{\alpha \in \mathcal{A}} \sum_{Q \in \mathcal{P}_n} |a_Q| |c_P^{(Q)}(\alpha)| \quad (\text{E89})$$

$$\leq \sum_{Q \in \mathcal{P}_n} |a_Q| \max_{\alpha \in \mathcal{A}} |c_P^{(Q)}(\alpha)|. \quad (\text{E90})$$

Now, we can use this result to upper bound $\|c\|_{1,\text{worst}}$ as follows:

$$\|c\|_{1,\text{worst}} \leq \sum_{P \in \mathcal{P}_n} \sum_{Q \in \mathcal{P}_n} |a_Q| \max_{\alpha \in \mathcal{A}} |c_P^{(Q)}(\alpha)| \quad (\text{E91})$$

$$= \sum_{Q \in \mathcal{P}_n} |a_Q| \|c^{(Q)}\|_{1,\text{worst}} \quad (\text{E92})$$

$$\leq \|a\|_1 \max_Q \|c^{(Q)}\|_{1,\text{worst}}, \quad (\text{E93})$$

where the last inequality is simply obtained by upper bounding each $\|c^{(Q)}\|_{1,\text{worst}}$ in the sum by $\max_Q \|c^{(Q)}\|_{1,\text{worst}}$. The average-case effective 1-norm is always upper bounded by the worst-case one. Therefore, we recover Eq. (E88) using Eq. (E84) and the assumption that $\|a\|_1 \in \mathcal{O}(1)$. ■

b. Uncorrelated parameters

Here, we assume that the each rotation parameter α_i is sampled independently from the uniform distribution over the interval $[-r, r]$. However, we remind the reader that this assumption can be relaxed, as the same results hold for any symmetric distribution satisfying the variance bound presented in Appendix E 2.

Supplemental Theorem E5 (Average-case Pauli surrogation guarantee). Consider an observable $O = \sum_{P \in \mathcal{P}_n} a_P P$, satisfying $\|a\|_1 \in \mathcal{O}(1)$ and containing N_{Paulis} distinct Pauli terms. Then running a classical Pauli Propagation algorithm in time

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} T m^{\log(1/\epsilon^2)} \right) \quad (\text{E94})$$

and performing

$$N_s = \left(\frac{3em}{\kappa} \right)^\kappa \in \mathcal{O} \left(m^{\log(1/\epsilon^2)} \right) \quad (\text{E95})$$

randomized Pauli measurements, it is possible to estimate a function $\tilde{f}_\kappa(\alpha)$ such that for all $r \in \mathcal{O}(1/\sqrt{m})$, we have

$$\mathbb{E}_{\mathcal{M}, \alpha} \left[\left(\tilde{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] \leq \epsilon^2, \quad (\text{E96})$$

where the expectation is both over the randomness of the parameters $\alpha \sim \mathcal{D}_m(\mathbf{0}, r)$ and the internal randomness of the measurement protocol denoted by $\mathcal{M}$.

Proof. First, let us bound the total mean square error by both contribution of the empirical error and the truncation error using that $(x + y)^2 \leq 2(x^2 + y^2)$ for any reals x and y . This leads to

$$\begin{aligned} \mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\tilde{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] &\leq 2\mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\tilde{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] \\ &\quad + 2\mathbb{E}_\alpha \left[\left(\tilde{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right]. \end{aligned} \quad (\text{E97})$$

Notice that the truncation error has no dependence on the measurement output. Therefore, the proof consists in two steps: first, we upper bound the runtime of Pauli Propagation for a given truncation error, and then we upper bound the number of required Pauli measurements for a given empirical error.

Pauli propagation step. Supplemental Theorem E3 ensures that the averaged truncation error in the Pauli

Propagation step is at most

$$\mathbb{E}_{\alpha \sim \mathcal{D}_m(\mathbf{0}, r)} \left[\left(\tilde{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] \leq \|a\|_1^2 \left(\frac{emr^2}{3\kappa} \right)^\kappa \quad (\text{E98})$$

with time

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa. \quad (\text{E99})$$

Surrogation step. It remains to upper bound the number of Pauli measurements. The observable of interest can be written as

$$\tilde{O}(\alpha) = \sum_P a_P \tilde{P}(\alpha). \quad (\text{E100})$$

The average-case effective 1-norm of each observable $\tilde{P}(\alpha)$ is upper bounded in Eq. (E84), and therefore from Eq. (E88) average-case effective 1-norm of $\tilde{O}(\alpha)$ is at most

$$\|a\|_1 \cdot \left(\frac{emr}{\kappa} \right)^\kappa. \quad (\text{E101})$$

Invoking Lemma B2, we find that

$$\mathbb{E}_{\alpha \sim \mathcal{D}_m(\mathbf{0}, r)} \left[\left(\tilde{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] \leq \frac{\|a\|_1^2}{N_s} \left(\frac{emr}{\kappa} \right)^{2\kappa}, \quad (\text{E102})$$

where N_s is the number of Pauli measurements (i.e., the number of measurement shots). Now, let us impose that mean square error upper bounds from both Eqs. (E98) and (E102) are equal to $(\epsilon/2)^2$ (to ensure a total mean square error ϵ^2),

$$\left(\frac{\epsilon}{2} \right)^2 = \frac{\|a\|_1^2}{N_s} \left(\frac{emr}{\kappa} \right)^{2\kappa} = \|a\|_1^2 \left(\frac{emr^2}{3\kappa} \right)^\kappa. \quad (\text{E103})$$

From this equation, we can solve for N_s , leading to

$$N_s = \left(\frac{3em}{\kappa} \right)^\kappa. \quad (\text{E104})$$

Finally, we follow the same analysis as in the proof of Supplemental Theorem E3 concerning the scaling of the quantity of interest. Therefore, for the time complexity, we have

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} T m^{\log(1/\epsilon^2)} \right). \quad (\text{E105})$$

For the sample complexity, we notice that $\mathcal{O}(N_s) \sim \mathcal{O}(t/(N_{\text{Paulis}} T))$ as

$$N_s = \left(\frac{3em}{\kappa} \right)^\kappa \in \mathcal{O} \left(m^{\log(1/\epsilon^2)} \right). \quad (\text{E106})$$

■

Next, we consider the special case of particular interest when the truncated, backpropagated observable contains only Pauli terms with low weight, i.e., Pauli terms acting nontrivially on few qubits. In this case, we can employ the machinery of classical shadow tomography [92,93] to obtain an alternative bound on the number of randomized Pauli measurements.

Supplemental Theorem E6 (Classical shadow tomography using randomized Pauli measurements, (Theorem 1 and Lemma 3 in Ref. [92])). Given an unknown n -qubit state ρ and a Pauli operator P of weight $|P| = k$, after N_s randomized Pauli measurements on copies of ρ satisfying

$$N_s \in \mathcal{O}\left(\frac{3^k \log(1/\delta)}{\epsilon^2}\right), \quad (\text{E107})$$

we can estimate $\text{Tr}(P\rho)$ to ϵ error with probability $1 - \delta$.

Supplemental Theorem E7 (Shadow Surrogation Guarantee, Formal). Under the same hypotheses of Supplemental Theorem E5, let us further assume that the truncated backpropagated observable contains only Pauli terms with constant Pauli weight, i.e.,

$$\tilde{O}(\alpha) = \sum_{\substack{P \in \mathcal{P}_n \\ |P| \in \mathcal{O}(1)}} c_P(\alpha) P. \quad (\text{E108})$$

Then the number of required Pauli measurements improves to

$$N_s \in \min(\log(n), \log(mN_{\text{Paulis}})) \cdot \mathcal{O}\left(\frac{\log(1/\delta)}{\epsilon}\right), \quad (\text{E109})$$

with probability $1 - \delta$ over the internal randomness of the measurement protocol.

Proof. In the first part of this proof, we restrict our attention to the case where the initial observable is a Pauli $O = Q$ with Pauli weight $|Q| \in \mathcal{O}(1)$. The final, backpropagated observable is

$$\tilde{Q}(\alpha) = \sum_{\omega \notin \mathcal{S}_\kappa^m} \Phi_\omega(\alpha) Q_\omega, \quad (\text{E110})$$

where for all $\omega \notin \mathcal{S}_\kappa^m$, the operator Q_ω satisfies $|Q_\omega| \in \mathcal{O}(1)$. We notice that the observable $\tilde{Q}(\alpha)$ contains at most

$$\min\{\text{poly}(n), \text{poly}(m)\} \quad (\text{E111})$$

different Paulis. This follows from the fact that it contains $\mathcal{O}(\text{poly}(m))$ different Pauli paths, and that, given a constant k , there exists only $\mathcal{O}(n^k)$ Paulis of weight at most k .

Then, we can use a classical shadows protocol for predicting some estimate $\hat{o}_\omega$ such that

$$\forall \omega \notin \mathcal{S}_\kappa^m : |\text{Tr}[Q_\omega \rho] - \hat{o}_\omega|^2 \leq \epsilon, \quad (\text{E112})$$

with probability at least $1 - \delta$. By Supplemental Theorem E6, this requires a number of Pauli measurement scaling as

$$N_s \in \mathcal{O}\left(\log\left(\frac{\log(\min(m, n))}{\delta}\right) \frac{1}{\epsilon}\right), \quad (\text{E113})$$

which in turn leads to

$$\begin{aligned} & \mathbb{E}_\alpha \left(\text{Tr}[\tilde{Q}(\alpha)\rho] - \sum_{\omega \notin \mathcal{S}_\kappa^m} \mathbb{E}_\alpha \Phi_\omega(\alpha) \hat{o}_\omega \right)^2 \\ &= \sum_{\omega \notin \mathcal{S}_\kappa^m} \mathbb{E}_\alpha \Phi_\omega^2(\alpha) (\text{Tr}[Q_\omega \rho] - \hat{o}_\omega)^2 \end{aligned} \quad (\text{E114})$$

$$\leq \max_{\omega \notin \mathcal{S}_\kappa^m} (\text{Tr}[Q_\omega \rho] - \hat{o}_\omega)^2 \mathbb{E}_\alpha \|\Phi(\alpha)\|_2^2 \leq \epsilon. \quad (\text{E115})$$

In the first step above we used the fact that different Fourier coefficients are uncorrelated, i.e. $\mathbb{E}_\alpha \Phi_\omega(\alpha) \Phi_{\omega'}(\alpha)$ for $\omega \neq \omega'$, and in the last step we used the fact that

$$\sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} \mathbb{E}_\alpha [\Phi_\omega(\alpha)^2] \leq \left(\frac{emr^2}{3\kappa}\right)^\kappa \leq 1, \quad (\text{E116})$$

which is provided by Supplemental Theorem E1 or more precisely Corollary E1.

We can upper bound the error for a general observable composed of N_{Paulis} different Pauli terms invoking Lemma E1. In particular, we need to rescale the error ϵ according to the 1-norm $\|a\|_1$, which we assumed to be $\mathcal{O}(1)$. Moreover, the number of different Pauli terms is at most

$$\min\{\text{poly}(n), N_{\text{Paulis}} \text{poly}(m)\}, \quad (\text{E117})$$

and therefore the required number of measurements is

$$N_s \in \mathcal{O}\left(\log\left(\frac{\min(mN_{\text{Paulis}}, n)}{\delta}\right) \frac{1}{\epsilon}\right). \quad (\text{E118})$$

■

c. Worst-case parameters

Supplemental Theorem E8 (Worst-case Pauli surrogation guarantee). Consider an observable $O = \sum_{P \in \{I, X, Y, Z\}^{\otimes n}} a_P P$, satisfying $\|a\|_1 \in \mathcal{O}(1)$ and $\|a\|_0 = N_{\text{Paulis}}$, i.e., O contains N_{Paulis} Pauli terms.

Then running a classical Pauli Propagation algorithm in time

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} T m^{\log(1/\epsilon)} \right) \quad (\text{E119})$$

and performing

$$N_s = 2 \log(2/\delta) \in \mathcal{O}(\log(1/\delta)) \quad (\text{E120})$$

randomized Pauli measurements, it is possible to estimate a function $\hat{f}_\kappa(\alpha)$ such that for all $r \in \mathcal{O}(1/m)$, we have

$$\left| \hat{f}_\kappa(\alpha) - f(\alpha) \right| \leq \epsilon, \quad (\text{E121})$$

with probability at least $1 - \delta$ over both the randomness of $\alpha \sim \mathcal{D}_m(\mathbf{0}, r)$ and the internal randomness of the measurement protocol.

Proof. The proof is analogous to the one of Supplemental Theorem E5. We start by split the total error as both truncation and empirical error contributions using triangle inequality, leading to

$$\left| \hat{f}_\kappa(\alpha) - f(\alpha) \right| \leq \left| \hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right| + \left| \tilde{f}_\kappa(\alpha) - f(\alpha) \right|. \quad (\text{E122})$$

Next, the proof consists of two steps. First, we upper bound the runtime of Pauli Propagation for a given truncation error, and second, we upper bound the number of required Pauli measurements for a given empirical error.

Pauli propagation step. Supplemental Theorem E4 ensures that the truncation error in the Pauli Propagation step is at most

$$\left| \tilde{f}_\kappa(\alpha) - f(\alpha) \right| \leq \|a\|_1 \left(\frac{emr}{\kappa} \right)^\kappa \quad (\text{E123})$$

with time

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa. \quad (\text{E124})$$

Surrogation step. It remains to upper bound the number of Pauli measurements. The observable of interest can be written as

$$\tilde{O}(\alpha) = \sum_P a_P \tilde{P}(\alpha). \quad (\text{E125})$$

The worst-case effective 1-norm related to each observable $\tilde{P}(\alpha)$ is upper bounded in Eq. (E84), and therefore from

Eq. (E88) worst-case effective 1-norm related to $\tilde{O}(\alpha)$ is at most

$$\|a\|_1 \cdot \left(\frac{emr}{\kappa} \right)^\kappa. \quad (\text{E126})$$

Invoking Lemma B3 ensures

$$\left| \hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right| \leq \sqrt{\frac{2 \log(2/\delta)}{N_s}} \|a\|_1 \left(\frac{emr}{\kappa} \right)^\kappa, \quad (\text{E127})$$

with probability at least $1 - \delta$ (over both parameters and measurement output), where N_s is the number of Pauli measurements (i.e., shots).

Now, let us assume that both errors are upper bounded by $\epsilon/2$ (to ensure a total error ϵ), so that

$$\frac{\epsilon}{2} = \sqrt{\frac{2 \log(2/\delta)}{N_s}} \|a\|_1 \left(\frac{emr}{\kappa} \right)^\kappa = \|a\|_1 \left(\frac{emr}{\kappa} \right)^\kappa. \quad (\text{E128})$$

From the previous, we can solve for N_s to get

$$N_s = 2 \log(2/\delta). \quad (\text{E129})$$

Finally, we follow the same analysis as in the proof of Supplemental Theorem E4 concerning the scaling of the quantity of interest. Therefore, for the time complexity, we have

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} T m^{\log(1/\epsilon)} \right). \quad (\text{E130})$$

Here, the sample complexity only depends on the probability δ . Indeed, both errors are similar given that $N_s \sim 2 \log(2/\delta)$, and $r \in \mathcal{O}(1/m)$ is small enough to ensure an arbitrary small error. ■

d. Even allocation of shots

Here, we provide a proof of the surrogate guarantee based on Supplemental Theorem E5 using even allocation of shots between the Paulis that needs to be measured instead of the effective 1-norm allocation strategy used in Supplemental Theorem E5. As opposed to the warm-up example presented in Appendix B 1, here we fix the number of shots n_s for each Pauli that needs to be measured in order to estimate the surrogate of the expectation function. Therefore, the total number of shots is at most $N_s = N_{\text{Paulis}} n_{\text{pth}}^{(\kappa)} n_s$ recalling that N_{Paulis} is the number of Paulis contained by O and $n_{\text{pth}}^{(\kappa)}$ is the number of path as mentioned previously.

Supplemental Theorem E9 (Surrogation via Direct Pauli Measurements, Formal). Consider an observable $O = \sum_{P \in \{I, X, Y, Z\}^{\otimes n}} a_P P$, satisfying $\|a\|_1 \in \mathcal{O}(1)$ and $\|a\|_0 = N_{\text{Paulis}}$, i.e., O contains N_{Paulis} Pauli terms.

Then running a classical Pauli Propagation algorithm in time

$$t \leq N_{\text{Paulis}} T m \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} T m^{\log(1/\epsilon^2)} \right) \quad (\text{E131})$$

and performing

$$N_s = N_{\text{Paulis}} \left(\frac{em}{\kappa} \right)^\kappa \in \mathcal{O} \left(N_{\text{Paulis}} m^{\log(1/\epsilon^2)} \right) \quad (\text{E132})$$

randomized Pauli measurements, it is possible to estimate a function $\hat{f}_\kappa(\alpha)$ such that for all $r \in \mathcal{O}(1/\sqrt{m})$, we have

$$\mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\hat{f}_\kappa(\alpha) - f(\alpha) \right)^2 \right] \leq \epsilon^2, \quad (\text{E133})$$

where the expectation is both over the randomness of $\alpha \sim \mathcal{D}_m(\mathbf{0}, r)$ and the internal randomness of the measurement protocol denoted by $\mathcal{M}$.

Notice that applying Markov's inequality together with Eq. (E133), we can ensure

$$\left| \hat{f}_\kappa(\alpha) - f(\alpha) \right| < \tilde{\epsilon}, \quad (\text{E134})$$

with probability at least $1 - \delta$ [for any $\delta \in \Theta(1)$] over both $\alpha \sim \mathcal{D}_m(\mathbf{0}, r)$ and the measurement output $\mathcal{M}$ where $\tilde{\epsilon} = \epsilon/\sqrt{\delta} \in \Theta(1)$ can be made arbitrarily small by increasing $\kappa \in \mathcal{O}(1)$. This leads to the formal version of Theorem 3 stated.

Proof. The proof here is the same as in Supplemental Theorem E5 except the surrogation step (i.e., empirical error estimation). In the following, we show that for uniform allocation of shots Eq. (E102) can be replaced by

$$\mathbb{E}_{\alpha, \mathcal{M}} \left[\left(\hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] \leq \frac{N_{\text{Paulis}} \|\mathbf{a}\|_2^2}{N_s} \left(\frac{emr}{\sqrt{3}\kappa} \right)^{2\kappa}. \quad (\text{E135})$$

The propagated observable can be written as

$$\tilde{O}(\alpha) = \sum_{\substack{P \in \mathcal{P}_n \\ a_P \neq 0}} \sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} a_P \Phi_\omega(\alpha) P_\omega, \quad (\text{E136})$$

where P_ω is the Pauli obtained after backpropagating P along the path ω . A worst-case scenario where each P_ω are different Paulis is considered here (that is, we don't group coefficients that corresponds to the same Pauli). So, for each Pauli P and each path ω considered (i.e., each P_ω), we will use n_s shots to estimate $\text{Tr}[\rho P_\omega]$. Let $\hat{x}_{P, \omega}^{(i)}$ denotes the i th measurement outcome obtained by measuring $\text{Tr}[\rho P_\omega]$ with the quantum computer. Then, our empirical estimate

of $\text{Tr}[\rho P_\omega]$ for each P_ω is simply the average over the n_s corresponding measurement outcomes

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \hat{x}_{P, \omega}^{(i)}. \quad (\text{E137})$$

Therefore, the empirical estimate for the surrogate of the expectation function is given by

$$\hat{f}_\kappa(\alpha) = \sum_{\substack{P \in \mathcal{P}_n \\ a_P \neq 0}} \sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} \sum_{i=1}^{n_s} \frac{a_P \Phi_\omega(\alpha)}{n_s} \hat{x}_{P, \omega}^{(i)}, \quad (\text{E138})$$

which is an unbiased estimator for $\tilde{f}_\kappa(\alpha)$. Moreover, notice that the $\hat{x}_{P, \omega}^{(i)}$ are uncorrelated (independent measurement). From the previous, the mean square error (over both the parameters α and the measurement process denoted by $\mathcal{M}$) is given by

$$\mathbb{E}_{\mathcal{M}, \alpha} \left[\left(\hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] = \mathbb{E}_\alpha \left[\text{Var}_{\mathcal{M}} \left[\hat{f}_\kappa(\alpha) \right] \right] \quad (\text{E139})$$

$$= \mathbb{E}_\alpha \left[\sum_{\substack{P \in \mathcal{P}_n \\ a_P \neq 0}} \sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} \sum_{i=1}^{n_s} \frac{a_P^2 \Phi_\omega(\alpha)^2}{n_s^2} \text{Var}_{\mathcal{M}} \left[\hat{x}_{P, \omega}^{(i)} \right] \right] \quad (\text{E140})$$

$$\leq \mathbb{E}_\alpha \left[\sum_{\substack{P \in \mathcal{P}_n \\ a_P \neq 0}} \sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} \sum_{i=1}^{n_s} \frac{a_P^2 \Phi_\omega(\alpha)^2}{n_s^2} \right] \quad (\text{E141})$$

$$= \frac{1}{n_s} \sum_{P \in \mathcal{P}_n} a_P^2 \sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} \mathbb{E}_\alpha [\Phi_\omega(\alpha)^2], \quad (\text{E142})$$

where the inequality is obtained by noticing that $\text{Var}_{\mathcal{M}} \left[\hat{x}_{P, \omega}^{(i)} \right] = 1 - \text{Tr}[\rho P_\omega]^2 \leq 1$. Moreover, from Supplemental Theorem E1 or more precisely from Corollary E1, we have

$$\sum_{\substack{\omega \\ \# \sin(\omega) > \kappa}} \mathbb{E}_\alpha [\Phi_\omega(\alpha)^2] \leq \left(\frac{emr^2}{3\kappa} \right)^\kappa. \quad (\text{E143})$$

The total number of shots is given by $N_s = n_s N_{\text{Paulis}} n_{\text{pth}}^{(\kappa)}$ where we recall that N_{Paulis} is the number of initial Paulis (i.e., the number of Pauli P such that $a_P \neq 0$) and $n_{\text{pth}}^{(\kappa)}$ is the number of Pauli path in the surrogate (the number of

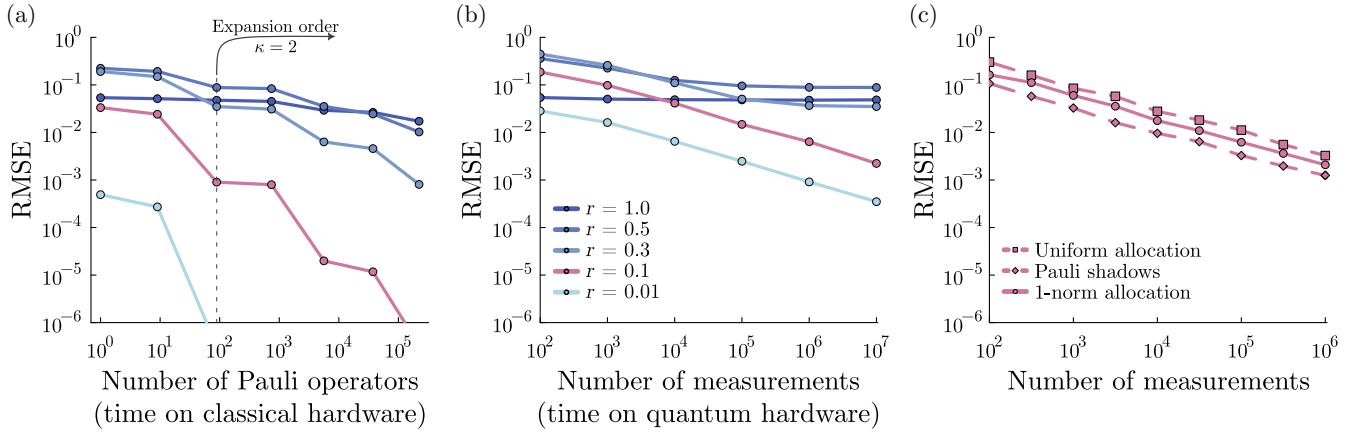


FIG. 9. Landscape and surrogation root mean square error as a function of classical and quantum resources with the Heisenberg Hamiltonian. We adapt the 16-qubit setup of Fig. 4 to quantum circuit representing a Heisenberg model Hamiltonian variational *Ansatz*, and the results in panels b and c are for order $\kappa = 2$. Other details remain the same. Overall, the Heisenberg Hamiltonian variational *Ansatz* generates more Pauli strings than the transverse field Ising, which puts more strain on both the classical and quantum components. Furthermore, the Pauli shadow estimation strategy now outperforms the 1-norm allocation.

path ω such that $\#\sin(\omega) > \kappa$). From Eq. (E56), we have

$$n_{pth}^{(\kappa)} \leq \left(\frac{em}{\kappa}\right)^\kappa. \quad (\text{E144})$$

Therefore, combining everything together leads to the following upper bound for the mean square empirical error

$$\mathbb{E} \left[\left(\hat{f}_\kappa(\alpha) - \tilde{f}_\kappa(\alpha) \right)^2 \right] \leq \frac{N_{\text{Paulis}} \|\mathbf{a}\|_2^2}{N_s} \left(\frac{emr}{\sqrt{3}\kappa} \right)^{2\kappa}. \quad (\text{E145})$$

Finally, substituting this in the proof of Supplemental Theorem E5, we get that the total number of shots required for even allocation is given by

$$N_s = \frac{\|\mathbf{a}\|_2^2 N_{\text{Paulis}}}{\|\mathbf{a}\|_1^2} \left(\frac{em}{\kappa} \right)^\kappa. \quad (\text{E146})$$

Therefore, by taking at least

$$N_s = N_{\text{Paulis}} \left(\frac{em}{\kappa} \right)^\kappa \quad (\text{E147})$$

is sufficient to get the same error guarantee as in Supplemental Theorem E5. ■

Notice here that with even allocation of shots we get a 2-norm instead of a 1-norm of the coefficients of the initial observable. However, we also get the additional factor N_{Paulis} . Let us consider the warm-up example presented in Appendix B 1 where the initial observable is given by

$$O = Z^{\otimes n} + \frac{\tau}{4^n - 1} \sum_{P \in \mathcal{P}_n \setminus \{Z^{\otimes n}\}} P, \quad (\text{E148})$$

where $\tau \in \mathcal{O}(1)$. Therefore, we have $\|\mathbf{a}\|_1^2 = (1 + \tau)^2 \in \mathcal{O}(1)$ and $\|\mathbf{a}\|_2^2 = 1 + (\tau^2/(4^n - 1)) \in \mathcal{O}(1)$, but $N_{\text{Paulis}} =$

4^n . Therefore, the number of shots required with even allocation (provided by the bound) is exponentially larger than the effective 1-norm allocation. In fact, this will highly depends on the number of Paulis in the initial observable. As we are mainly considering $\mathcal{O}(\text{poly}(n))$ Pauli in the observable, both bounds remains polynomial with n and m .

7. Additional numerical results

To better elucidate the system-dependence of the PP surrogation in Fig. 9, we here showcase the same numerical experiment with a Heisenberg model HVA instead of a transverse-field Ising (TFI) HVA. This *Ansatz* not only contains more Pauli rotations, but also more and larger groups of commuting operators. On the one hand, this leads to more Pauli strings being created during the classical simulation, in turn putting more strain on both the classical and quantum components. Generally speaking, the errors at the same resources are larger than in Fig. 4. However, we also see the Pauli shadow estimation strategy becoming superior to the 1-norm strategy, likely due to the increased structure and number of commuting terms. A notable difference between shadows and our 1-norm allocation is that the latter does not group commuting Pauli strings together for simultaneous estimation, which the shadow strategy implicitly does. This structure can be leveraged in sophisticated estimation strategies to design the best quantum-enhanced classical surrogate computation.

[1] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme

- et al.*, Evidence for the utility of quantum computing before fault tolerance, *Nature* **618**, 500 (2023).
- [2] M. S. Rudolph, E. Fontana, Z. Holmes, and L. Cincio, Classical surrogate simulation of quantum systems with LOWESA, [arXiv:2308.09109](#).
 - [3] J. Tindall, M. Fishman, E. M. Stoudenmire, and D. Sels, Efficient tensor network simulation of IBM's Eagle kicked Ising experiment, *PRX Quantum* **5**, 010308 (2024).
 - [4] T. Begušić, J. Gray, and G. K.-L. Chan, Fast and converged classical simulations of evidence for the utility of quantum computing before fault tolerance, *Sci. Adv.* **10**, eadk4321 (2024).
 - [5] E. G. D. Torre and M. M. Roses, Dissipative mean-field theory of IBM utility experiment, [arXiv:2308.01339](#).
 - [6] H.-J. Liao, K. Wang, Z.-S. Zhou, P. Zhang, and T. Xiang, Simulation of IBM's kicked Ising experiment with projected entangled pair operator, [arXiv:2308.03082](#).
 - [7] S. J. Thomson and J. Eisert, Unravelling quantum dynamics using flow equations, *Nat. Phys.* **20**, 1401 (2024).
 - [8] 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 (2025).
 - [9] N. Dowling, P. Kos, and X. Turkeshi, Magic resources of the Heisenberg picture, *Phys. Rev. Lett.* **135**, 050401 (2025).
 - [10] F. J. Schreiber, J. Eisert, and J. J. Meyer, Classical surrogates for quantum learning models, *Phys. Rev. Lett.* **131**, 100803 (2023).
 - [11] S. Jerbi, C. Gyurik, S. C. Marshall, R. Molteni, and V. Dunjko, Shadows of quantum machine learning, *Nat. Commun.* **15**, 5676 (2024).
 - [12] J. Landman, S. Thabet, C. Dalyac, H. Mhiri, and E. Kashefi, Classically approximating variational quantum machine learning with random Fourier features, [arXiv:2210.13200](#).
 - [13] A. Angrisani, Learning unitaries with quantum statistical queries, *Quantum* **9**, 1817 (2025).
 - [14] E. Fontana, M. S. Rudolph, R. Duncan, I. Rungger, and C. Cîrstoiu, Classical simulations of noisy variational quantum circuits, *npj Quantum Inf.* **11**, 1 (2025).
 - [15] 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).
 - [16] K. Bharti, T. Haug, V. Vedral, and L.-C. Kwek, Noisy intermediate-scale quantum algorithm for semidefinite programming, *Phys. Rev. A* **105**, 052445 (2022).
 - [17] A. Basheer, Y. Feng, C. Ferrie, and S. Li, Alternating layered variational quantum circuits can be classically optimized efficiently using classical shadows, [arXiv:2208.11623](#).
 - [18] S. Jerbi, J. Gibbs, M. S. Rudolph, M. C. Caro, P. J. Coles, H.-Y. Huang, and Z. Holmes, The power and limitations of learning quantum dynamics incoherently, [arXiv:2303.12834](#).
 - [19] A. Basheer, Y. Feng, C. Ferrie, and S. Li, Ansatz-agnostic exponential resource saving in variational quantum algorithms using shallow shadows, [arXiv:2309.04754](#).
 - [20] R. Shaffer, L. Kocia, and M. Sarovar, Surrogate-based optimization for variational quantum algorithms, *Phys. Rev. A* **107**, 032415 (2023).
 - [21] R. D. Somma, Quantum computation, complexity, and many-body physics, [arXiv:quant-ph/0512209](#).
 - [22] R. Somma, H. Barnum, G. Ortiz, and E. Knill, Efficient solvability of Hamiltonians and limits on the power of some quantum computational models, *Phys. Rev. Lett.* **97**, 190501 (2006).
 - [23] V. Galitski, Quantum-to-classical correspondence and Hubbard-Stratonovich dynamical systems: A Lie-algebraic approach, *Phys. Rev. A* **84**, 012118 (2011).
 - [24] M. L. Goh, M. Larocca, L. Cincio, M. Cerezo, and F. Sauvage, Lie-algebraic classical simulations for quantum computing, *Phys. Rev. Res.* **7**, 033266 (2025).
 - [25] E. R. Anschuetz, A. Bauer, B. T. Kiani, and S. Lloyd, Efficient classical algorithms for simulating symmetric quantum systems, *Quantum* **7**, 1189 (2023).
 - [26] 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).
 - [27] P. Bermejo, P. Braccia, M. S. Rudolph, Z. Holmes, L. Cincio, and M. Cerezo, Quantum convolutional neural networks are (effectively) classically simulable, *PRX Quantum* **7**, 020304 (2026).
 - [28] A. W. R. Smith, A. J. Paige, and M. S. Kim, Faster variational quantum algorithms with quantum kernel-based surrogate models, *Quantum Sci. Technol.* **8**, 045016 (2023).
 - [29] A. A. Mele, A. Angrisani, S. Ghosh, S. Khatri, J. Eisert, D. Stilck França, and Y. Quek, Noise-induced shallow circuits and absence of barren plateaus, [arXiv:2403.13927](#).
 - [30] R. Sweke, E. Recio, S. Jerbi, E. Gil-Fuster, B. Fuller, J. Eisert, and J. J. Meyer, Potential and limitations of random Fourier features for dequantizing quantum machine learning, *Quantum* **9**, 1640 (2025).
 - [31] B. Koczor and S. C. Benjamin, Quantum analytic descent, *Phys. Rev. Res.* **4**, 023017 (2022).
 - [32] E. J. Gustafson, J. Tihiainen, D. Chamaki, F. Sorourifar, J. W. Mullinax, A. C. Li, F. B. Maciejewski, N. P. Sawaya, J. T. Krogel, D. E. B. Neira *et al.*, Surrogate optimization of variational quantum circuits, [arXiv:2404.02951](#).
 - [33] K. Zhang, L. Liu, M.-H. Hsieh, and D. Tao, Escaping from the Barren Plateau via Gaussian Initializations in Deep Variational Quantum Circuits, in *Advances in Neural Information Processing Systems* (2022), Vol. 35, pp. 18612–18627.
 - [34] C.-Y. Park and N. Killoran, Hamiltonian variational ansatz without barren plateaus, *Quantum* **8**, 1239 (2024).
 - [35] 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).
 - [36] C.-Y. Park, M. Kang, and J. Huh, Hardware-efficient ansatz without barren plateaus in any depth, [arXiv:2403.04844](#).
 - [37] X. Shi and Y. Shang, Avoiding barren plateaus via Gaussian mixture model, [arXiv:2402.13501](#).

- [38] R. Puig, M. Drudis, S. Thanasilp, and Z. Holmes, Variational quantum simulation: A case study for understanding warm starts, *PRX Quantum* **6**, 010317 (2025).
- [39] H. Mhiri, L. Monbroussou, M. Herrero-Gonzalez, S. Thabet, E. Kashefi, and J. Landman, Constrained and vanishing expressivity of quantum Fourier models, *Quantum* **9**, 1847 (2025).
- [40] R. Puig, B. Casas, A. Cervera-Lierta, Z. Holmes, and A. Pérez-Salinas, Warm starts, cold states: Exploiting adiabaticity for variational ground-states, [arXiv:2602.06137](https://arxiv.org/abs/2602.06137).
- [41] H. Mhiri, R. Puig, S. Lerch, M. S. Rudolph, T. Chotibut, S. Thanasilp, and Z. Holmes, A general framework for analyzing warm starts in variational quantum computing (to be published).
- [42] S. Lerch, J. Bowles, R. Puig, E. Armengol, Z. Holmes, and S. Thanasilp, IQP born machines under data-dependent and agnostic initialization strategies, [arXiv:2603.14576](https://arxiv.org/abs/2603.14576).
- [43] C. Huerta Alderete, M. H. Gordon, F. Sauvage, A. Sone, A. T. Sornborger, P. J. Coles, and M. Cerezo, Inference-based quantum sensing, *Phys. Rev. Lett.* **129**, 190501 (2022).
- [44] A. Ijaz, C. H. Alderete, F. Sauvage, L. Cincio, M. Cerezo, and M. L. Goh, More buck-per-shot: Why learning trumps mitigation in noisy quantum sensing, *Mater. Today Quantum* **6**, 100042 (2025).
- [45] M. Cerezo, K. Sharma, A. Arrasmith, and P. J. Coles, Variational quantum state eigensolver, *npj Quantum Inf.* **8**, 1 (2022).
- [46] C. Cirstoiu, Z. Holmes, J. Iosue, L. Cincio, P. J. Coles, and A. Sornborger, Variational fast forwarding for quantum simulation beyond the coherence time, *npj Quantum Inf.* **6**, 1 (2020).
- [47] P. Rall, D. Liang, J. Cook, and W. Kretschmer, Simulation of qubit quantum circuits via pauli propagation, *Phys. Rev. A* **99**, 062337 (2019).
- [48] D. Aharonov, X. Gao, Z. Landau, Y. Liu, and U. Vazirani, A polynomial-time classical algorithm for noisy random circuit sampling, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (2023), p. 945.
- [49] N. A. Nemkov, E. O. Kiktenko, and A. K. Fedorov, Fourier expansion in variational quantum algorithms, *Phys. Rev. A* **108**, 032406 (2023).
- [50] T. Begušić, K. Hejazi, and G. K. Chan, Simulating quantum circuit expectation values by Clifford perturbation theory, *J. Chem. Phys.* **162**, 154110 (2025).
- [51] Y. Shao, F. Wei, S. Cheng, and Z. Liu, Simulating noisy variational quantum algorithms: A polynomial approach, *Phys. Rev. Lett.* **133**, 120603 (2024).
- [52] T. Schuster, C. Yin, X. Gao, and N. Y. Yao, A polynomial-time classical algorithm for noisy quantum circuits, *Phys. Rev. X* **15**, 041018 (2025).
- [53] G. González-García, J. I. Cirac, and R. Trivedi, Pauli path simulations of noisy quantum circuits beyond average case, *Quantum* **9**, 1730 (2025).
- [54] M. S. Rudolph, A. Angrisani, A. Wright, I. Sanderski, R. Puig, and Z. Holmes, Thermal state simulation with Pauli and Majorana propagation, [arXiv:2602.04878](https://arxiv.org/abs/2602.04878).
- [55] M. S. Rudolph, T. Jones, Y. Teng, A. Angrisani, and Z. Holmes, Pauli propagation: A computational framework for simulating quantum systems, [arXiv:2505.21606](https://arxiv.org/abs/2505.21606).
- [56] 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).
- [57] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nat. Commun.* **9**, 1 (2018).
- [58] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Equivalence of quantum barren plateaus to cost concentration and narrow gorges, *Quantum Sci. Technol.* **7**, 045015 (2022).
- [59] M. Ragone, B. N. Bakalov, F. Sauvage, A. F. Kemper, C. Ortiz Marrero, M. Larocca, and M. Cerezo, A lie algebraic theory of barren plateaus for deep parameterized quantum circuits, *Nat. Commun.* **15**, 7172 (2024).
- [60] E. Fontana, D. Herman, S. Chakrabarti, N. Kumar, R. Yalovetzky, J. Heredge, S. H. Sureshbabu, and M. Pistoia, Characterizing barren plateaus in quantum ansätze with the adjoint representation, *Nat. Commun.* **15**, 7171 (2024).
- [61] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-induced barren plateaus, *PRX Quantum* **2**, 040316 (2021).
- [62] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, Trainability of dissipative perceptron-based quantum neural networks, *Phys. Rev. Lett.* **128**, 180505 (2022).
- [63] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, *Phys. Rev. Res.* **3**, 033090 (2021).
- [64] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.* **12**, 1 (2021).
- [65] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, Diagnosing barren plateaus with tools from quantum optimal control, *Quantum* **6**, 824 (2022).
- [66] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, *PRX Quantum* **3**, 010313 (2022).
- [67] 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**, 1 (2021b).
- [68] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht, and A. T. Sornborger, Barren plateaus preclude learning scramblers, *Phys. Rev. Lett.* **126**, 190501 (2021).
- [69] E. C. Martín, K. Plekhanov, and M. Lubasch, Barren plateaus in quantum tensor network optimization, *Quantum* **7**, 974 (2023).
- [70] G. Crognaletti, M. Grossi, and A. Bassi, Estimates of loss function concentration in noisy parametrized quantum circuits, [arXiv:2410.01893](https://arxiv.org/abs/2410.01893).
- [71] E. R. Anschuetz, A Unified Theory of Quantum Neural Network Loss Landscapes, in *International Conference on Learning Representations*, edited by Y. Yue, A. Garg, N. Peng, F. Sha and R. Yu (2025), pp. 97859–97918, https://proceedings.iclr.cc/paper_files/paper/2025/file/f3680ca31afcb9076846950dbc72c7e3-Paper-Conference.pdf.

- [72] K. Srimahajariyapong, S. Thanasilp, and T. Chotibut, Connecting phases of matter to the flatness of the loss landscape in analog variational quantum algorithms, [arXiv:2506.13865](#).
- [73] M. S. Rudolph, S. Lerch, S. Thanasilp, O. Kiss, O. Shaya, S. Vallecorsa, M. Grossi, and Z. Holmes, Trainability barriers and opportunities in quantum generative modeling, *npj Quantum Inf.* **10**, 116 (2024).
- [74] M. Kieferova, O. M. Carlos, and N. Wiebe, Quantum generative training using Rényi divergences, [arXiv:2106.09567](#).
- [75] J. Tangpanitanon, S. Thanasilp, N. Dangniam, M.-A. Lemonde, and D. G. Angelakis, Expressibility and trainability of parametrized analog quantum systems for machine learning applications, *Phys. Rev. Res.* **2**, 043364 (2020).
- [76] S. Thanasilp, S. Wang, N. A. Nghiem, P. J. Coles, and M. Cerezo, Subtleties in the trainability of quantum machine learning models, *Quantum Mach. Intell.* **5**, 21 (2023).
- [77] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, Exponential concentration in quantum kernel methods, *Nat. Commun.* **15**, 5200 (2024).
- [78] A. Letcher, S. Woerner, and C. Zoufal, Tight and efficient gradient bounds for parameterized quantum circuits, *Quantum* **8**, 1484 (2024).
- [79] S. Y. Chang, S. Thanasilp, B. L. Saux, S. Vallecorsa, and M. Grossi, Latent style-based quantum GAN for high-quality image generation, [arXiv:2406.02668](#).
- [80] W. Xiong, G. Facelli, M. Sahebi, O. Agnel, T. Chotibut, S. Thanasilp, and Z. Holmes, On fundamental aspects of quantum extreme learning machines, *Quantum Mach. Intell.* **7**, 20 (2025).
- [81] W. Xiong, Z. Holmes, A. Angrisani, Y. Suzuki, T. Chotibut, and S. Thanasilp, Role of scrambling and noise in temporal information processing with quantum systems, [arXiv:2505.10080](#).
- [82] R. Aghaei Saem, B. Tafreshi, Z. Holmes, and S. Thanasilp, Pitfalls when tackling the exponential concentration of parameterized quantum models, *Quantum Sci. Technol.* **11**, 015049 (2026).
- [83] C. Cao, Y. Zhou, S. Tannu, N. Shannon, and R. Joynt, Exploiting many-body localization for scalable variational quantum simulation, *Quantum* **9**, 1942 (2025).
- [84] L. Xin and Z.-Q. Yin, Improve variational quantum eigensolver by many-body localization, *Front. Phys.* **20**, 023202 (2025).
- [85] A. A. Mele, G. B. Mbeng, G. E. Santoro, M. Collura, and P. Torta, Avoiding barren plateaus via transferability of smooth solutions in a Hamiltonian variational ansatz, *Phys. Rev. A* **106**, L060401 (2022).
- [86] A. Miller, J. Favre, Z. Holmes, Ö. Salehi, R. Chakraborty, A. Nykänen, Z. Zimboras, A. Glos, and G. García-Pérez, Simulation of fermionic circuits using majorana propagation, [arXiv:2503.18939](#).
- [87] A. Angrisani, A. A. Mele, M. S. Rudolph, M. Cerezo, and Z. Holmes, Simulating quantum circuits with arbitrary local noise using pauli propagation, [arXiv:2501.13101](#).
- [88] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, *Phys. Rev. A* **70**, 052328 (2004).
- [89] We note that surrogation beyond $r \in \mathcal{O}(1/\sqrt{m})$ is also achievable with low error, by extending the truncation order κ to be polynomial. This, in turn, increases the computational resources required to be super-polynomial or even exponential with m . Detailed derivations for various truncations are provided in Appendix C 1 g. Furthermore, these techniques could be adapted to other geometries. Although a rigorous generalization requires more technical work, Fig. 3 offers evidence that the results of Theorem 1 remain valid.
- [90] J. M. Arrazola, O. Di Matteo, N. Quesada, S. Jahangiri, A. Delgado, and N. Killoran, Universal quantum circuits for quantum chemistry, *Quantum* **6**, 742 (2022).
- [91] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, General parameter-shift rules for quantum gradients, *Quantum* **6**, 677 (2022).
- [92] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nat. Phys.* **16**, 1050 (2020).
- [93] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, The randomized measurement toolbox, *Nat. Rev. Phys.* **5**, 9 (2023).
- [94] M. West, A. A. Mele, M. Larocca, and M. Cerezo, Real classical shadows, [arXiv:2410.23481](#).
- [95] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, *Phys. Rev. A* **92**, 042303 (2015).
- [96] A. Miessen, D. J. Egger, I. Tavernelli, and G. Mazzola, Benchmarking digital quantum simulations above hundreds of qubits using quantum critical dynamics, *PRX Quantum* **5**, 040320 (2024).
- [97] W. H. Zurek, U. Dorner, and P. Zoller, Dynamics of a quantum phase transition, *Phys. Rev. Lett.* **95**, 105701 (2005).
- [98] T. Kibble, Some implications of a cosmological phase transition, *Phys. Rep.* **67**, 183 (1980).
- [99] W. H. Zurek, Cosmological experiments in superfluid helium? *Nature* **317**, 505 (1985).
- [100] A. J. Daley, I. Bloch, C. Kokail, S. Flannigan, N. Pearson, M. Troyer, and P. Zoller, Practical quantum advantage in quantum simulation, *Nature* **607**, 667 (2022).
- [101] N. Herrmann, D. Arya, M. W. Doherty, A. Mingare, J. C. Pillay, F. Preis, and S. Prestel, Quantum utility—definition and assessment of a practical quantum advantage, in *2023 IEEE International Conference on Quantum Software (QSW)* (IEEE, 2023), p. 162.
- [102] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to Global Optimization* (Springer Science & Business Media, Springer New York, NY, 2000).
- [103] B. Koczor and S. C. Benjamin, Quantum analytic descent, *Phys. Rev. Res.* **4**, 023017 (2022).
- [104] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms* (Cambridge University Press, Cambridge, UK, 2009).
- [105] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.* **58**, 13 (1963).
- [106] P. Billingsley, *Probability and Measure*, Wiley Series in Probability and Statistics (Wiley, 1995).

- [107] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Series in Electrical and Computer Engineering (McGraw-Hill, 2002).
- [108] M. Protter and C. Morrey, *Intermediate Calculus*, Undergraduate Texts in Mathematics (Springer, New York, 2012).
- [109] J. J. Duistermaat and J. A. C. Kolk, in *Distributions: Theory and Applications* (Birkhäuser Boston, Boston, 2010), p. 59.
- [110] A. Arrasmith, L. Cincio, R. D. Somma, and P. J. Coles, Operator sampling for shot-frugal optimization in variational algorithms, [arXiv:2004.06252](https://arxiv.org/abs/2004.06252).
- [111] N. C. Rubin, R. Babbush, and J. McClean, Application of fermionic marginal constraints to hybrid quantum algorithms, *New J. Phys.* **20**, 053020 (2018).
- [112] M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, *Quantum Sci. Technol.* **6**, 035006 (2021).
- [113] We do not rule out that others using a different proof strategy could not prove guarantees in a larger region [112], although we think this unlikely.
- [114] D. Aharonov, X. Gao, Z. Landau, Y. Liu, and U. Vazirani, A polynomial-time classical algorithm for noisy random circuit sampling, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (2023), pp. 945–957.
- [115] D. Gottesman, Group22: Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics, edited by S. P. Corney, R. Delbourgo, and P. D. Jarvis (Cambridge, MA, International Press, 1999), pp. 32–43.