# Using gradient-based algorithms to determine ground-state energies on a quantum computer

Tomislav Piskor,[1,2,*] Jan-Michael Reiner,[1] Sebastian Zanker,[1] Nicolas Vogt,[1] Michael Marthaler,[1] Frank K. Wilhelm,[2,3] and Florian G. Eich[1]

[1]*HQS Quantum Simulations GmbH, Haid-und-Neu-Strasse 7, 76131 Karlsruhe, Germany*

[2]*Theoretical Physics, Saarland University, 66123 Saarbrücken, Germany*

[3]*Institute for Quantum Computing Analytics (PGI-12), Forschungszentrum Jülich, 52045 Jülich, Germany*

(Received 22 September 2021; revised 15 March 2022; accepted 18 May 2022; published 8 June 2022)

Variational algorithms are promising candidates to be implemented on near-term quantum computers. In our work, we investigate the variational Hamiltonian ansatz (VHA), where a parametrized trial state of the quantum mechanical wave function is optimized to obtain the ground-state energy. In the VHA, the trial state is given by a noninteracting reference state modified by unitary rotations using generators that are part of the Hamiltonian describing the system. The lowest energy is obtained by optimizing the angles of those unitary rotations. A standard procedure to optimize the variational parameters is to use gradient-based algorithms. However, shot noise and the intrinsic noise of the quantum device affect the evaluation of the required gradients. We study how different methods for obtaining the gradient, specifically the finite-difference and the parameter-shift rule, are affected by shot noise and the noise of the quantum computer. To this end, we simulate a simple quantum circuit, as well as the two-site and six-site Hubbard models.

DOI: 10.1103/PhysRevA.105.062415

## I. INTRODUCTION

Recent advances in quantum computing hardware [1–3] have led to an increasing interest in developing quantum algorithms for noisy intermediate scale quantum (NISQ) computers. One of the most promising applications for these near-term quantum computers is the simulation of fermionic systems, such as lattice systems or small molecules. The determination of the ground state and ground-state energy is one of the key aspects to gain information about the system. Hybrid quantum-classical algorithms, such as the Variational Quantum Eigensolver (VQE) [4] turn out to have low resource requirements making them suitable for NISQ computers. These algorithms determine the ground-state energy of a system via preparing a parametrized trial state on a quantum computer and optimizing the parameter set with a classical optimization routine.

The definition of the trial state depends on the choice of the VQE ansatz. One ansatz that is well suited for lattice systems is the variational Hamiltonian ansatz (VHA) [5–7], stating that for each term of a generic decomposition of the Hamiltonian a separate parameter is defined. This way, the number of parameters, and thus the circuit depth can be reduced compared to other VQE ansätze, such as the unitary coupled cluster with single and double excitations (uCCSD) [8]. In the uCCSD ansatz parameters are defined based on single and double excitations of electrons from occupied to unoccupied (molecular) orbitals, thus the number of parameters grows polynomially with the system size (with the excitation level determining the highest power of the polynomial scaling). By contrast, the VHA uses the structure of the Hamiltonian to determine the variational form. Therefore, periodic lattice Hamiltonians, which are typically characterized by a sparse Hamiltonian with few independent coupling constants yield an ansatz where the number of parameters does not scale with the system size.

The VHA is inspired by an adaptation of the so-called "adiabatic connection" from many-body perturbation theory [9,10]. Starting from the solution (ground state) of noninteracting electrons the state is evolved on the quantum computer using a sequence of unitary propagators constructed using parts of the fully interacting Hamiltonian. The variational parameters correspond to the propagation times of each unitary operator. Having opted for the VHA to provide the parametrized trial state, the parameter set is optimized such that the expectation value of the interacting Hamiltonian is minimized. There are several approaches to determine the optimal parameter set for a given cost function ranging from gradient-free [11,12] to gradient-based algorithms [13], which usually lead to faster convergence compared to the gradient-free alternatives. In the present work we focus on the analysis of the evaluation of gradients on NISQ computers, which means that we are less concerned about the specific gradient-based optimization algorithm. Hence, we use a simple steepest-descent approach for the optimization, where the parameters are updated using the gradient of the energy directly, employing a fixed learning rate (damping of the gradient).

The main question we address in this work is how measurement statistics and noise affect the optimization of the parameters appearing in the quantum circuit. Since we always have to perform measurements to get the expectation value of an observable, shot noise will always exist (even on error-

*tomislav.piskor@quantumsimulations.de

 

corrected quantum devices). However, when simulating the quantum device we can obtain the limit of an infinite number of measurements, providing us with the ideal reference result for the algorithm. Furthermore, due to adverse coupling of the qubits to the environment, another source of error has to be considered, namely the intrinsic depolarization of qubits. However, we find that the dominant error for the noise model chosen in our simulations stems from measurement statistics.

For the determination of the cost function's gradient we compare two procedures: (1) the finite-difference approximation to compute the gradients, which we also refer to as the "numerical" method and (2) the so-called parameter-shift rule, which we also refer to as the "analytical" method [14–18]. While calculating the gradient numerically, the outcome of the result is more susceptible towards noise effects, such as shot noise or depolarization effects occurring in qubits, because there is a competition for improving the numerical gradient by reducing the finite step size to evaluate the gradient versus resolving differences of the cost function evaluated for two nearby values of a given parameter. To bypass this hurdle, gradients can also be determined analytically via the parameter-shift rule, with the hope being to show a more resilient behavior [18] to noise effects because it relies on a *fixed* finite difference. However, there is an important caveat for using the parameter-shift rule for optimizing a VHA: Compared to the numerical case, where the number of additional circuit evaluations only grows with the number of parameters, the number of additional circuit evaluations grows with two times the number of parametrized gates for the parameter-shift rule. While the number of parameters solely depends on the ansatz, the number of parametrized gates depends on the implementation of the trial state in the quantum circuit. In general it has to be considered that one parameter might occur in multiple parametrized gates (especially in case of the VHA, see Sec. II) and thus the number of parameters might be much smaller than the number of parametrized gates.

In the following we compare both methods for determining the gradients, using a simple gradient-based optimization algorithm, by investigating a simple one-qubit circuit and a two- and six-site Hubbard model, mapped onto 4 and 12 qubits, respectively. In Sec. II we first lay down the theoretical model which we analyze in this work. We introduce the VHA, the Hubbard Hamiltonian, describe our gradient evaluation and parameter optimization routines, and explain our considered noise models; we also highlight certain subtleties that arise from the particular combination of ingredients that comprise our theoretical model. In Sec. III we then study our model through numerical simulation, where the results are shown for a simple quantum circuit in Sec. III A and for the one-dimensional Hubbard model in Sec. III B.

## II. MODEL

In this section, we expound the theoretical model that we analyze numerically below in Sec. III. In the following subsections we explain the ingredients of our model, motivate the different parts of the theoretical approach, and state again the main question we investigate in this work.

We are interested in near-term applications for NISQ devices. In this context, hybrid quantum-classical variational algorithms are widely discussed as a potential candidate; specifically for solving electronic structure problems [5,6,19].

In this work we chose to utilize the VHA as an ansatz for a variational quantum algorithm [5]. Compared to, e.g., a VQE-type uCCSD ansatz, where the number of parameters scales polynomially with the system size, the VHA has the benefit of having a constant number of parameters with respect to the number of orbitals. Yet, in contrast to often referred hardware-efficient ansätze [19], the VHA has a clear physical motivation behind it being inspired by an adiabatic evolution.

Particularly for highly structured systems like the Hubbard model, one can easily choose parametrizations that are deduced from physical reasoning (see below). In an earlier work of ours, the setup of using the VHA to simulate the Hubbard model proved to remain efficient under the influence of a specific type of noise [6], and (with some adaptations to the ansatz) enabled us to investigate the ground state, including explicit symmetry breaking through external fields [7].

To optimize the VHA we focus on gradient-based optimizers. With the application on NISQ devices in mind, we consider certain types of noise that make the evaluation of gradients more challenging. Specifically, we study the parameter-shift rule which was discussed to be more resilient to noise [18], and compare it to the "standard" gradient calculation by a finite difference method. The main question we address is: How does the parameter-shift rule perform under a realistic condition for a quantum simulation of an electronic structure problem on NISQ hardware under the influence of noise? As noise types we consider shot noise and depolarization.

We point out that this question is particularly interesting for the VHA, as the quantum circuits have few parameters to optimize, but a single parameter will appear in multiple parametrized gates in the circuit, which is a subtle but important difference to earlier works investigating the parameter-shift rule.

Following this motivation of the model the consequent subsections will treat the individual pieces of our model in more detail: We first describe the VHA (Sec. II A) and the Hubbard Hamiltonian (Sec. II B). Afterwards, we explain the gradient descent optimizer we use (Sec. II C) and our means of gradient evaluation through a finite-difference method and (more importantly) the parameter-shift rule (Sec. II D). We continue by formalizing our considered noise models of shot noise (Sec. II E) and depolarization (Sec. II F). Finally, we provide an analysis of our quantum circuits and highlight the difference between the number of individual parameters and parametrized gates and the connected complication for the parameter-shift rule (Sec. II G).

### A. Variational Hamiltonian ansatz

The VHA starts from a trial state given by

$$|\psi(\boldsymbol{\theta})\rangle = \hat{U}(\boldsymbol{\theta})\,|\psi_0\rangle\,, \tag{1}$$

with $\boldsymbol{\theta}$ defining the parameter set, $\hat{U}(\boldsymbol{\theta})$ being a parametrized unitary operator, which is implemented by a quantum circuit and $|\psi_0\rangle$ an initial state usually chosen to be a single Slater determinant of orbitals obtained from the mean-field solution. The energy of this trial state for a system with Hamiltonian $\hat{\mathcal{H}}$

can then be written as

$$E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,\hat{\mathcal{H}}\,|\psi(\boldsymbol{\theta})\rangle\,. \tag{2}$$

Minimizing the energy with respect to the parameter set leads to the optimal energy attainable with the ansatz and the corresponding parameter set $\boldsymbol{\theta}_{\text{opt}}$ satisfying

$$E(\boldsymbol{\theta}_{\text{opt}}) = \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}). \tag{3}$$

According to the variational principle [20,21], the exact ground state of the Hamiltonian $E_{\text{exact}}$ sets a lower bound for the variationally determined energy

$$E_{\text{exact}} \leqslant E(\boldsymbol{\theta}_{\text{opt}}). \tag{4}$$

Starting with an initial guess, $\boldsymbol{\theta}_0$, the parameter set is updated iteratively, using the procedure described in Secs. II C and II D, to approach the optimal solution.

The explicit form of the unitary operator $\hat{U}(\boldsymbol{\theta})$ is constructed by decomposing the Hamiltonian into $P$ separate terms, as stated in Eq. (5a). The partial contributions to the Hamiltonian are Hermitian operators, so we can use them as generators for unitary rotations. Since the partial contributions do not commute in general, the order in which the unitaries are applied matters. In practice we repeat the application of the unitaries $R$ times and allow for different parameters (or rotation angles) in each repetition [Eq. (5b)] yielding a total number of $n = RP$ free variational parameters

$$\hat{\mathcal{H}} = \sum_{\alpha=1}^{P} \hat{\mathcal{H}}_\alpha, \tag{5a}$$

$$\hat{U}(\boldsymbol{\theta}) = \prod_{k=1}^{R}\prod_{\alpha=1}^{P} e^{i\theta_{\alpha,k}\hat{\mathcal{H}}_\alpha}. \tag{5b}$$

### B. Hamiltonian

In this work we investigate the one-dimensional (1D) Hubbard model, which is described by the Hamiltonian

$$\begin{aligned}\hat{\mathcal{H}} &= \hat{\mathcal{T}} + \hat{\mathcal{W}} \\ &= -t\sum_i\sum_\sigma(\hat{c}^\dagger_{i,\sigma}\hat{c}_{i+1,\sigma} + \hat{c}^\dagger_{i+1,\sigma}\hat{c}_{i,\sigma}) \\ &\quad + U\sum_i\left(\hat{c}^\dagger_{i,\uparrow}\hat{c}_{i,\uparrow} - \frac{1}{2}\right)\left(\hat{c}^\dagger_{i,\downarrow}\hat{c}_{i,\downarrow} - \frac{1}{2}\right),\end{aligned} \tag{6}$$

with the first term describing nearest-neighbor hopping between sites $i$ and $i+1$ and spin $\sigma$, being either in the spin up $\uparrow$ or spin down $\downarrow$ state, with hopping amplitude $t$ (kinetic energy) and the second term describing the repulsion of two electrons on the same site with interaction strength $U$. We consider a system is at half-filling (an average of one electron per site). In our studies we focus on the case where the kinetic energy and the interaction have the same magnitude and use $t$ as our unit of energy, i.e., $U = t = 1$. Moreover, we employ periodic boundary conditions. The Hubbard model at half filling is a prototypical model describing a Mott insulator, i.e., an insulator where the fundamental gap is due to electron-electron interaction and any (static) mean-field theory would yield a metal (unless the symmetry is artificially broken).

Quite naturally Hamiltonian (6) can be split into two contributions, i.e., the kinetic energy $\hat{\mathcal{T}}$ and the interaction energy $\hat{\mathcal{W}}$. For system sizes larger than two sites the hopping operator is split further into a so-called "even" and "odd" contribution, $\hat{\mathcal{T}}_e$ and $\hat{\mathcal{T}}_o$. Note that in our case $\hat{\mathcal{T}}_e$, $\hat{\mathcal{T}}_o$, and $\hat{\mathcal{W}}$ correspond to the $H_\alpha$ of Eq. (5b). Furthermore, the individual contributions to each of the three parts of the Hamiltonian, $\hat{\mathcal{T}}_e$, $\hat{\mathcal{T}}_o$, and $\hat{\mathcal{W}}$ commute among each other, which implies that each exponential can be split and reordered easily without using the Baker-Campbell-Hausdorff formula.

### C. Gradient descent optimizer

In this work we are interested in investigating how noise affects the synthesis of gradients of the cost function [cf. Eq. (2)]. While there are several established algorithms, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [22–25], for gradient-based optimization, we decided to use a very simple approach, namely a (damped) steepest-descent optimizer, to reduce the complexity in the optimization process and focus on the gradient. Specifically, we update a parameter $\theta_i$ according to

$$\theta_i^{\tau+1} = \theta_i^\tau - \eta\partial_{\theta_i}E(\boldsymbol{\theta}^\tau), \tag{7}$$

where $\partial_{\theta_i}E(\boldsymbol{\theta})$ denotes the derivative of the cost function with respect to $\theta_i$, $\eta$ is a *fixed* learning rate (damping) controlling the step size towards the minimum of the cost function, and $\theta_i^\tau$ is the $i$th parameter at iteration step $\tau$. The parameter set at iteration $\tau+1$ is given by the parameter set at step $\tau$ modified by the cost function's gradient scaled by the learning rate $\eta$. Compared to other gradient-based optimizers [13], the steepest-descent optimizer is very simplistic in its form since it contains only one fixed hyperparameter. Since no information beyond the gradient, such as an approximate Hessian matrix, is used, algorithm (7) may converge slower than, for example, the aforementioned BFGS optimizer.

### D. Gradient evaluation

To determine the gradients for the optimization routine, two possibilities will be discussed throughout this work. These two possibilities will either be via a finite-difference method or via the parameter-shift rule.

First, gradients can be determined numerically with a finite-difference method in the following way:

$$\begin{aligned}\partial_{\theta_i}E(\boldsymbol{\theta}) \approx \frac{1}{\epsilon}(&E(\theta_1,\ldots,\theta_i+\epsilon,\ldots,\theta_n) \\ &- E(\theta_1,\ldots,\theta_i,\ldots,\theta_n)),\end{aligned} \tag{8}$$

with $\epsilon$ defining a small *but finite* step by which the parameter $\theta_i$ is shifted. This immediately implies that the number of additional circuit evaluations for obtaining the gradient using the finite-difference method corresponds to the number of parameters. However, this method may be more susceptible towards noise effects, such as statistical or depolarization noise due to the fact that the energy difference [cf. Eq. (8)] is getting smaller, and therefore harder to resolve, if we improve the accuracy of the gradient by making the step width $\epsilon$ smaller.

On the other hand we determine gradients analytically with the so-called parameter-shift rule [14]. At its core

the parameter-shift rule uses the fact that the cost function generally is represented by a quantum circuit composed of (parametrized) single-qubit rotations and *fixed* two-qubit gates. Note, that every unitary can be represented in such a form (e.g., using CNOT plus arbitrary single-qubit rotations as a universal gate set [26]) and that we use such decompositions in our simulations to implement the exponentials of the VHA in Eq. (5b). Note also, that for the decompositions of the VHA evolution a single parameter $\theta_{\alpha,k}$ will, in general, appear in multiple rotation gates; however, for the application of the parameter-shift rule they will need to be treated individually to evaluate the gradient. Focusing on the dependence of the energy on a single parameter $\theta_1$, which we assume to only control *one* single-qubit gate, the energy, Eq. (2), can be written as

$$E(\theta_1;\theta_2,\ldots,\theta_n) = A_1\cos(\omega\theta_1+\varphi_1)+\ldots \\ \ldots + A_\alpha\cos(\omega\theta_1+\varphi_\alpha)+\ldots+C, \quad (9)$$

with amplitudes $A_\alpha$, phase shifts $\varphi_\alpha$, and constant $C$, which depend on all other parameters $\theta_2,\ldots,\theta_n$ (see the Appendix A for a detailed derivation). The distance between the two eigenvalues of the generator for the single-qubit rotation is denoted by $\omega$ (for Pauli matrices we have $\omega = 2$). The main idea of the parameter-shift rule is to consider the derivative of a trigonometric function

$$\begin{aligned}\partial_\theta\cos(\omega\theta+\varphi) &= -\omega\sin(\omega\theta+\varphi)\\ &= \frac{\omega}{2}\Big[\cos\Big(\omega\Big[\theta+\frac{\pi}{2\omega}\Big]+\varphi\Big)\\ &\quad -\cos\Big(\omega\Big[\theta-\frac{\pi}{2\omega}\Big]+\varphi\Big)\Big], \quad (10)\end{aligned}$$

which highlights that the *exact* derivative with respect to the parameter $\theta$ is obtained by taking the difference of two quantum circuits, where the parameter is shifted by $\pm\frac{\pi}{2\omega}$, respectively. The parameter-shift rule, as presented above, is based on the assumption that a parameter $\theta_i$ *only* controls a single single-qubit gate (cf. the Appendix). However, the parameter $\theta_i$ can appear in more than one gate for the VHA aAnsatz. This is due to the fact that the parameters are defined in Eq. (5b) in reference to generators from the electronic Hamiltonian. In transforming the electronic Hamiltonian into the quantum circuit several single-qubit rotations are parametrized, in general, by the same parameter $\theta_i$. Hence the parameter-shift rule *cannot* be simply applied to the parameters defined in Eq. (5b). Use of the parameter-shift rule can be vindicated by defining a new set of parameters, $\boldsymbol{\mu}$, which assigns each parametrized single-qubit gate its own parameter $\mu_j$ ($j = 1,2,\ldots,m-1,m$, where $m \geqslant n$). Then we can compute the gradient with respect to the new set of parameters as

$$\begin{aligned}\partial_{\mu_j}E(\boldsymbol{\mu}) = r\Big[&E\Big(\mu_1,\ldots,\mu_j+\frac{\pi}{4r},\ldots,\mu_m\Big)\\ &-E\Big(\mu_1,\ldots,\mu_j-\frac{\pi}{4r},\ldots,\mu_m\Big)\Big]. \quad (11)\end{aligned}$$

Note that the parameter $r$, controlling the "step width" in principle depends on the explicit form of the unitary operator Eq. (5b), but we can always define a linear map from the parameter set $\boldsymbol{\theta}$ to $\boldsymbol{\mu}$, such that $r$ is the same for all parametrized gates ($r = \frac{1}{2}$ for standard Pauli rotations, see, e.g., Ref. [14] for details).

We emphasize that the parameter-shift rule yields, in principle, the exact gradient and not an approximation as the finite-difference method described above. Moreover, the parameter-shift rule is potentially more resilient towards noise and other effects, because (in spite of being exact) it is evaluated using a finite difference with a step width on the order of the spectral width of the generator (typically a scaled Pauli matrix). However, the number of circuit evaluations scales with the number of parametrized gates and thus, with increasing system size, requires more circuits than the finite-difference method. In our work we consider a structured ansatz, the VHA, for generating the variational quantum circuit, which requires discussing the difference between the number of parameters, $n$ (the number of components for $\boldsymbol{\theta}$), and the number of parametrized gates, $m$ (the number of components for $\boldsymbol{\mu}$). For a generic variational quantum circuit, which aims at optimizing all single-qubit gates independently, the number of parameters coincides trivially with the number of parametrized gates.

### E. Shot noise

The first type of noise we consider is shot noise which is inherent to any quantum device relying on a projective qubit measurement; even quantum error correction will not mitigate it. It is an important aspect when considering real applications for variational algorithms.

Shot noise stems from the fact that we use a finite number of projective measurements to evaluate the energy as in Eq. (2). The Hamiltonian is encoded into qubits as a linear combination of Pauli products, hence, to calculate the expectation value of the Hamiltonian one needs to estimate the expectation values of these Pauli products. However, the measurement process on a quantum computer is through the projective measurement of the qubits and a single measurement for a Pauli product yields either $-1$ or 1. Only through the statistics of many measurements an estimate of the expectation value be made. While for infinitely many measurement shots, the result will be exact, in reality the number will be finite. Thus, in reality we will always have a residual error due to shot noise.

Particularly for a variational algorithm the number of measurements per optimization loop has to be treated as a scarce resource since the energy has to be estimated very often to optimize the parameters of the circuit otherwise the total runtime of the algorithm will increase tremendously. Therefore, considering how robust the gradient evaluation and optimization procedure is against shot noise is highly important.

In the numerical results in Sec. III we faithfully simulate the projective measurement process. We always state exactly how many measurement shots were used for each simulation.

### F. Depolarization noise

In this work we consider the application of a variational algorithm, specifically on NISQ devices. To analyze this, we include an obstacle in our simulations that one will have to deal with on real hardware without complete quantum error

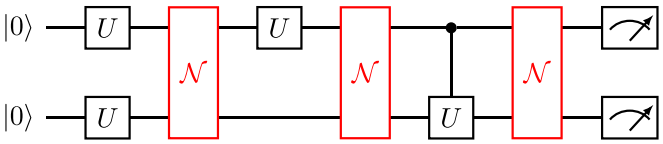


FIG. 1. Graphical illustration of a simple two-qubit circuit with one- and two-qubit unitary operations $U$ and three noise gates, labeled by $\mathcal{N}$, describing the effect of depolarization.

correction, namely decoherence. As the decoherence channel we take depolarization noise acting on the quantum device into account, which we simulate in the following way.

A quantum device can effectively be considered as an open quantum system, i.e., the qubits on the quantum chip interacting with the environment. We simulate the execution of the circuits as follows: the quantum circuit is arranged in such a way, that as many gates as possible are executed in parallel. After that the noise gate is applied to all qubits in the circuit, which can effectively be described as the application of Kraus operators to all qubits. This procedure is repeated until the measurement of observables is performed and is illustrated in Fig. 1.

The noise channel $\hat{\mathcal{N}}$, acting on a single qubit, can be described as [27]

$$\hat{\mathcal{N}}(\hat{\rho}) = \sum_{i=0}^{3} \hat{K}_i \hat{\rho} \hat{K}_i^\dagger, \tag{12}$$

with $\hat{\rho}$ being the density matrix representing the quantum system. The Kraus operators $\hat{K}_i$ are defined as follows:

$$\hat{K}_0 = \sqrt{1 - \frac{3}{4}\Gamma}\hat{I}, \tag{13a}$$

$$\hat{K}_i = \frac{\sqrt{\Gamma}}{2}\hat{\sigma}_i, i \in \{1, 2, 3\}, \tag{13b}$$

with $\Gamma = 1 - e^{-\gamma}$ defining the depolarizing term and $\hat{\sigma}_i$ being the Pauli operators. It can be easily seen that in the noiseless case, i.e., $\gamma = 0$, the depolarizing term $\Gamma$ vanishes and only the Kraus operator with the identity operator $\hat{I}$ is left, leading to a noiseless quantum operation $\hat{\mathcal{N}}(\hat{\rho}) = \hat{\rho}$. In the noisy case with $\gamma > 0$, not only $\hat{K}_0$, but also the three other Kraus operators $\hat{K}_i$ including the Pauli operators are taken into account, so that the quantum operation $\hat{\mathcal{N}}$ induces depolarization in the system. The parameter $\gamma$, determining the strength of the noise, is estimated as the ratio of the gate time $T_g$ and the coherence time $T_2$ of a qubit.

The gate time $T_g$ represents the time it takes to perform a given operation, for instance, the duration of a microwave pulse for controlling superconducting qubits. In Fig. 1 the gate time can be viewed as the width of a single gate column, if we interpret the horizontal direction of the algorithm as the time axis. An example for a gate operation is a qubit flip, where a qubit prepared in the $|0\rangle$ state is flipped to the $|1\rangle$ state after applying an $X$ gate on that qubit. For superconducting architectures, the usual single and two-qubit gate times range from 5–500 ns [28,29], whereas trapped-ion designs show single-qubits gate times in the microseconds and two-qubit gate times in the 10–100 $\mu$s regime [30,31]. The depolarization time $T_1$ determines the time it takes for an *undesired* qubit flip to occur due to the coupling of the qubit to its environment. Hence, if we would like to perform $N$ operations on a qubit, we require $NT_g \ll T_1$ or equivalently $\gamma \ll \frac{1}{N}$. In passing we note that the dephasing time $T_2$ is limited by the relaxation time $T_1$ of the qubit, i.e., $T_2 \leqslant 2 \cdot T_1$. While typical coherence times have values in the order of 100 $\mu$s for superconducting qubits [28,32–34], these values are several orders of magnitude higher for ion trap devices, which can be several seconds [30]. Considering these numbers, the ratio $\gamma$, which characterizes the noise strength, in our simulations is chosen within

$$\gamma = \frac{T_g}{T_2} \in [10^{-4}, 10^{-2}]. \tag{14}$$

TABLE I. Number of parameters and parametrized gates for 1D Hubbard chains with varying number of sites.

| # sites $M$ | Hilbert space size $2^{2M}$ | # parameters | # parametrized gates |
|---|---|---|---|
| 2 | 16 | 2 | 10 |
| 4 | 256 | 3 | 28 |
| 6 | 4 096 | 3 | 42 |
| 8 | 65 536 | 3 | 56 |
| 10 | 1 048 576 | 3 | 70 |
| 12 | 16 777 216 | 3 | 84 |
| 14 | 268 435 456 | 3 | 98 |
| 16 | 4 294 967 296 | 3 | 112 |

### G. Quantum circuit analysis

In this section, we highlight the difference between the number of parametrized gates $m$ and the number of variational parameters $n$ for the VHA mentioned earlier to expose the additional overhead for using the parameter-shift rule, i.e., the fact that $m \geqslant n$. An overview for the one-dimensional Hubbard model is given in Table I, where we compare the number of parameters and the number of parametrized gates for 1D Hubbard rings of various sizes (number of sites). Therefore a circuit was generated using a two-qubit decomposition [35,36], consisting of one-qubit rotation gates and a two-qubit controlled-$Z$ gate.

Table I shows the number of parameters and parametrized circuits for various sizes, determined by the number of sites $M$ in the 1D Hubbard model. In the case of $M = 2$ with one repetition for the ansatz made in Eq. (5b), we only have two parameters, namely one parameter for the hopping operator and one for the interaction operator. Increasing the site number to $M = 4$ or $M = 6$ yields one further parameter per repetition since we now also split the hopping term (kinetic energy) into two internally commuting contributions (labeled "even" and "odd"). The total number of circuits to evaluate for calculating energy and gradient with the finite-difference method is given by the number of parameters, i.e., one circuit for each parameter shifted by $\epsilon$, plus the circuit with no shift applied to any parameter. Considering the two-site Hubbard model with one repetition in the VHA, which implies two parameters, a total number of three circuits would be required

to evaluate the gradient and the energy. Increasing the number of repetitions from one to two, leading to four parameters, five circuits have to be evaluated to determine the energy and its gradient. In general, the total number of circuit evaluations required for gradient and energy determination is given by $N_{\mathrm{fd}} = RP + 1 = n + 1$ when using the finite-difference method, where $P$ defines the number of parameters and $R$ the number of repetitions in the VHA [cf. Eq. (5b)].

For the parameter-shift rule, the additional circuits for evaluating the gradient is proportional to twice the number of parametrized gates, leading to 21 circuits for the two-site Hubbard model with one repetition. Similarly to the finite-difference case, the number of parametrized gates scales linearly with the number of parameters and therefore also with the number of repetitions. In general, the number of circuits required using the parameter-shift rule can be written as $N_{\mathrm{ps}} = 2C(M)RP + 1 = 2m + 1$, where we introduce the coefficient $C(M)$, which translates the number of parameters to the number of parametrized gates. Considering the data in Table I and system sizes larger than two sites, we get $C(M) = \frac{7}{3}M$, which implies that the number of additional circuits scales linearly with the system size using the Jordan-Wigner transformation [37] to map fermionic sites to qubits. Note that this scaling is still superior to uCCSD, which scales quadratically at best [4].

## III. NUMERICAL RESULTS

In this section, we present numerical results for a simple quantum circuit and the 1D Hubbard model with two and six sites. In all cases, 50 000 measurements were performed to get the averaged result of the respective observables. Both the finite-difference and parameter-shift algorithm were performed five times to get the stochastic effect of the finite number of measurements. The shaded areas in the upcoming plots thus mark the worst (highest value) and the best (lowest value) energy for each iteration step and do not represent the standard deviations of the optimization runs. The solid lines mark runs without any effect of shot noise, representing the case where the number of measurements $N \to \infty$.

### A. Simple circuit

To get a better understanding of the gradient descent optimizer, a simple one-qubit circuit is investigated, where at first a Hadamard gate is applied to the prepared ground state $|0\rangle$ followed by a rotation $Z$ gate [see Eq. (15a)]. After this operation, a measurement of the Pauli $X$ operator is performed, leading to a trivial periodic function

$$\begin{aligned} |\psi(\theta)\rangle &= \hat{U}(\theta)\,|\psi_0\rangle = \hat{R}_z(\theta)\hat{H}\,|\psi_0\rangle \\ &= \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} e^{-i\theta/2} \\ e^{i\theta/2} \end{pmatrix}, \end{aligned} \tag{15a}$$

$$E(\theta) = \langle \hat{X} \rangle = \langle \psi(\theta)|\, \hat{X}\, |\psi(\theta)\rangle = \cos(\theta). \tag{15b}$$

The minimum of Eq. (15b) is attained for $\theta = (1 + 2p)\pi$ ($p \in \mathbb{Z}$) with a minimal value of $E_{\mathrm{exact}} = -1$.

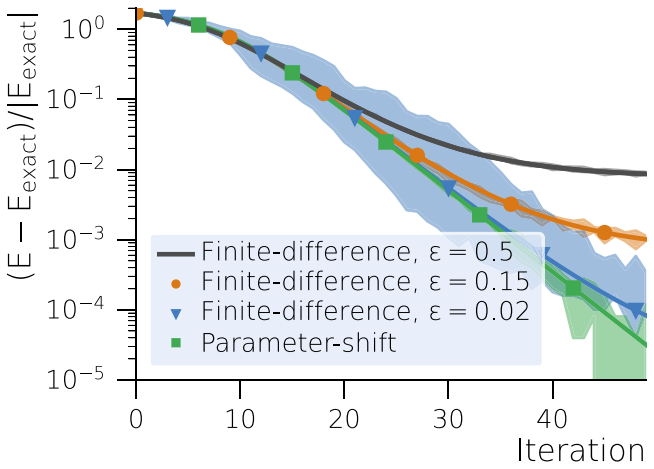


FIG. 2. Optimization runs using a simple gradient descent ansatz performed with the finite-difference method compared to the parameter-shift rule for the simple circuit and $N = 50\,000$. The solid lines indicate runs without shot noise, whereas the corresponding shaded region marks five optimization runs performed with shot noise. For all optimization runs, a learning rate of $\eta = 0.5$ is used.

#### 1. Shot noise

Figure 2 shows optimization runs for different values of $\epsilon$ for the finite-difference method, as well as runs performed with the parameter-shift rule with $N = 50\,000$ measurement shots. Note that in all plots, starting from Fig. 2, we use a logarithmic scale for the energy differences shown on the $y$ axis. The solid lines indicate simulations without any source of noise ($N \to \infty$) and the shaded areas with the corresponding color are obtained as the envelope of five runs performed with statistical errors ($N = 50\,000$), i.e., the upper line marks the maximum and the bottom line indicates the minimum of the relative deviation from the optimal value at a given iteration of the optimization, respectively. Focusing on the $N \to \infty$ results, it can be seen that with decreasing $\epsilon$ the accuracy gets better, which highlights the error introduced by approximating the derivative using a finite difference. However, if statistical effects are included it can also be seen that the fluctuations around the optimal curve increase with smaller $\epsilon$, clearly exposing that smaller $\epsilon$ are more prone to statistical errors. Considering the optimization runs performed with the parameter-shift rule it can be seen that, on the one hand, a more accurate result is achieved for the noiseless run ($N \to \infty$), demonstrating that the parameter-shift rule yields the exact gradient, and on the other hand, the fluctuations around the optimal curve are smaller compared to the finite-difference runs with $\epsilon = 0.02$.

#### 2. Shot and depolarization noise

Including the effect of depolarization noise, we observe that the deviation from the cost function's minimum increases. This is most pronounced for the parameter-shift rule and small values of $\epsilon$ for the finite-difference method. We can attribute this to the fact that the error introduced by using a larger $\epsilon$ in the finite difference dominates the overall error. This is depicted in Fig. 3, where for the two mentioned cases a small shift away from the minimum can be seen, compared

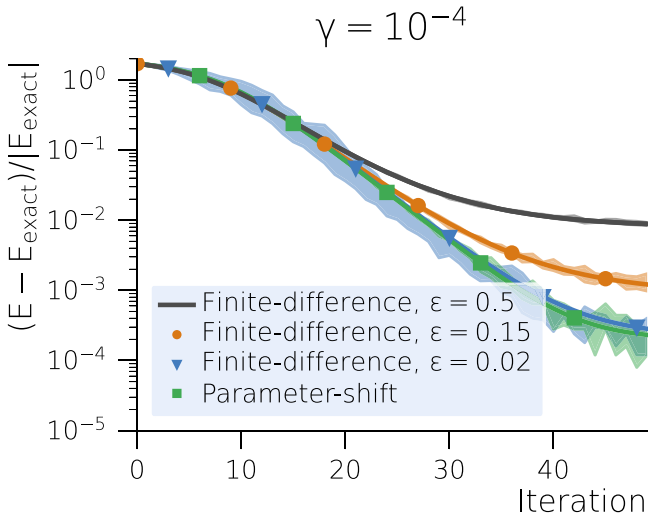


FIG. 3. Same plot as Fig. 2, but with depolarization noise, characterized by the rate $\gamma = 10^{-4}$ included in the simulation.

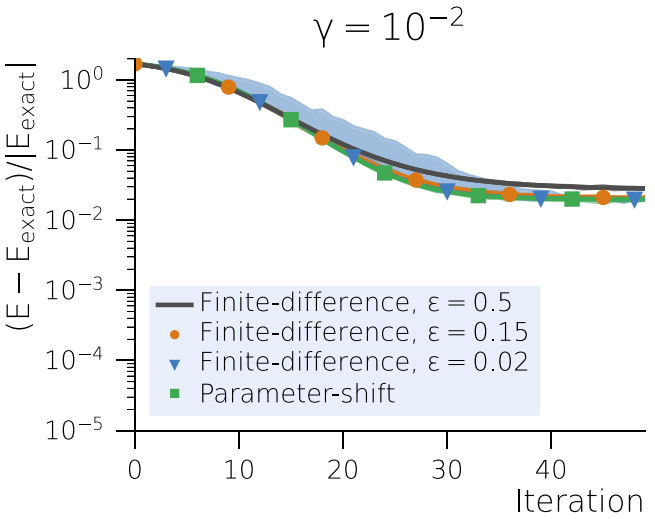


FIG. 5. Same plot as Fig. 3, but with a depolarization rate of $\gamma = 10^{-2}$.

to the noiseless run in Fig. 2. Increasing the depolarization rate further leads to a higher deviation from the minimum for all cases, as can be seen in Figs. 4 and 5. Especially for the case where $\gamma = 10^{-2}$, the finite-difference curves with the two smallest $\epsilon$ and the parameter-shift curve nearly show the same result for the cost function after 50 iteration steps, clearly showing that the depolarization error dominates the overall error.

### B. 1D Hubbard model

Next, we study the 1D Hubbard model by performing optimization runs with shot noise only and with shot and depolarization noise. The Hubbard Hamiltonian is defined in Eq. (6), where the initial state $|\psi_0\rangle$ is chosen as the ground-state Slater determinant for the noninteracting system.

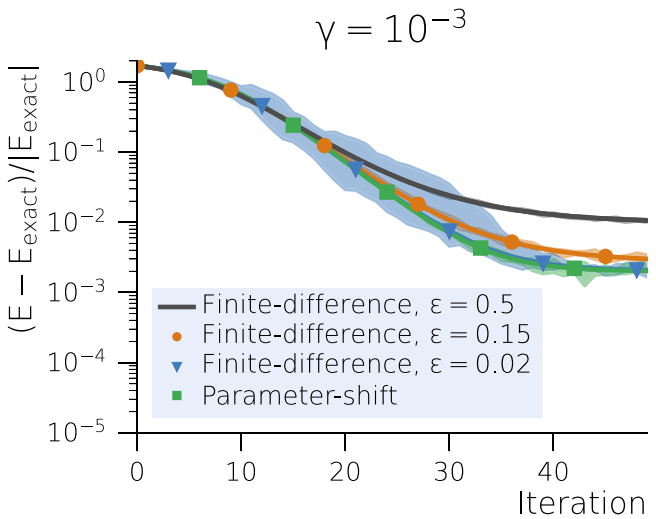


FIG. 4. Same plot as Fig. 3, but with a depolarization rate of $\gamma = 10^{-3}$.

### C. Results for two sites

For the two-site Hubbard model we decompose the Hamiltonian simply into kinetic energy and the interaction energy. Thus, the trial state for the two-site Hubbard model is given as

$$|\psi(\boldsymbol{\theta})\rangle = e^{i\theta_2\hat{\mathcal{T}}} e^{i\theta_1\hat{\mathcal{W}}} |\psi_0\rangle\,, \tag{16}$$

where we only use one repetition of the unitaries since this parametrization already encompasses the exact ground state for the two-site Hubbard model. As in the previous example the parameter set $\boldsymbol{\theta}$ was determined with the gradient-descent method explained in Sec. II C. The exact analytical result for the two-site Hubbard model [38] with Eq. (16) as the trial state has been taken as the reference value $E_{\text{exact}}$.

#### *1. Shot noise*

First, optimization runs including only shot noise were performed, where the number of measurement shots was set to $N = 50\,000$. As in the previous case, the solid lines in Fig. 6 show runs without shot noise, corresponding to the limit $N \to \infty$. The shaded areas represent the maximum and minimum of five optimization runs performed with shot noise, respectively. From the graph it can be seen that the shaded region for $\epsilon = 0.5$ is vanishingly small at the scale of the plot. Starting around iteration step 10 it can be observed, at the scale of the plot, that the fluctuations of the shaded regions increase with smaller $\epsilon$. The more accurate the gradients are in principle, the more relevant becomes the fact that, in practice, we always perform a finite amount of measurements. Figure 6 shows that the energy deviation exhibits a minimum for gradients obtained via finite difference, which can be attributed to the fact that the approximate derivatives do not even, in principle, correspond to the true derivatives of the cost function. This has to be contrasted with the energy deviation obtained by using the parameter-shift rule to synthesize the gradients for the optimization procedure. Here, the gradients are, in principle, $(N \to \infty)$ exact and the only error stems from the finite accuracy due to a finite number of measurements for the gradient. However, our results show that, for a fixed number of

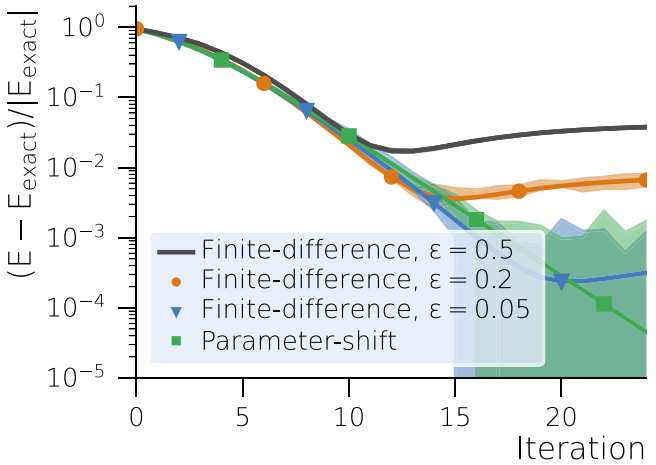


FIG. 6. Optimization runs using a simple gradient descent Ansatz performed with the finite-difference method compared to the parameter-shift rule for the two-site Hubbard model and $N = 50\,000$. The solid lines indicate runs without shot noise, whereas the corresponding shaded region marks five optimization runs performed with shot noise. For all optimization runs, a learning rate of $\eta = 0.1$ is used.

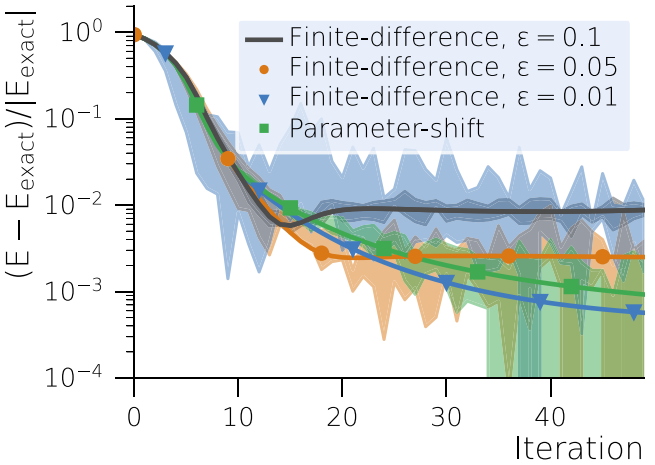


FIG. 8. Optimization runs using a simple gradient descent ansatz performed with the finite-difference method compared to the parameter-shift rule for the six-site Hubbard model and $N = 50\,000$. The solid lines indicate runs without shot noise, whereas the corresponding shaded region marks five optimization runs performed with shot noise. For all optimization runs, a learning rate of $\eta = 0.03$ is used.

measurements ($N = 50\,000$), the intrinsic error in calculating the gradients via finite difference using $\epsilon = 0.05$ is smaller than the statistical error due to the finite number of measurements. Considering the measurement overhead (cf. Table I and Sec. II G) in synthesizing the gradients via the parameter-shift rule favors the finite-difference gradients.

#### 2. *Shot and depolarization noise*

By adding depolarization noise with a rate of $\gamma = 10^{-4}$ to the previous simulations it can be seen that there are no significant differences for $\epsilon = 0.5$ and $\epsilon = 0.2$, except a minor increase in the deviation from the exact energy, as shown in Fig. 7. However, there is a larger offset for the parameter-shift run. For example, around the 20th iteration step, the $N \to \infty$ parameter-shift curve flattens out. As in the previous case, parameter-shift simulations show a similar behavior as the finite-difference runs with small $\epsilon$.

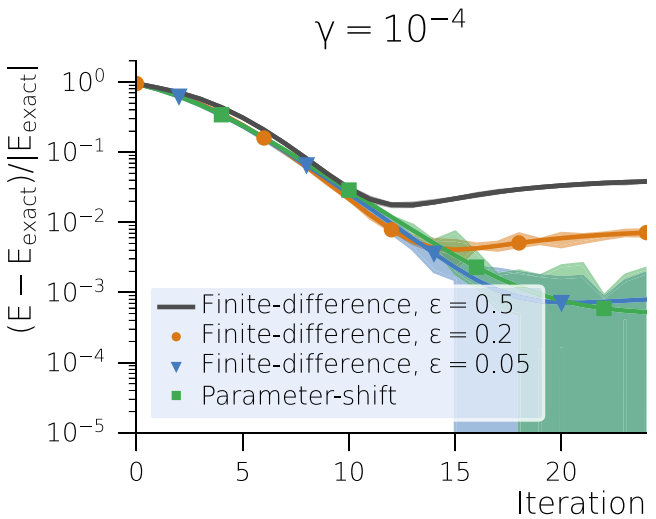


FIG. 7. Same as Fig. 6, but now including depolarization noise, characterized by the rate $\gamma = 10^{-4}$, in the simulations.

### D. Results for six sites

In contrast to the two-site Hubbard model, the number of repetitions has been set to two for the six-site Hubbard model, leading to a trial state given as

$$|\psi(\boldsymbol{\theta})\rangle = e^{i\theta_6\hat{\mathcal{T}}_o} e^{i\theta_5\hat{\mathcal{T}}_e} e^{i\theta_4\hat{\mathcal{V}}} e^{i\theta_3\hat{\mathcal{T}}_o} e^{i\theta_2\hat{\mathcal{T}}_e} e^{i\theta_1\hat{\mathcal{V}}} |\psi_0\rangle\,, \tag{17}$$

with $\hat{\mathcal{T}}_e$ and $\hat{\mathcal{T}}_o$ defined in Sec. II B.

By increasing the number of repetitions and thus the number of parameters in the ansatz, the variationally determined energy is closer to the true ground-state energy of the system. Note that we use the optimal energy achievable with the given ansatz (17) as the reference energy and denote it as $E_{\text{exact}}$. Due to the system size, runs with depolarization noise were not performed.

#### *Shot noise*

For the six-site Hubbard model we perform optimization runs for three different choices of $\epsilon$ and the parameter-shift rule with $N = 50\,000$. Taking a look at the solid lines in Fig. 8, representing noiseless runs, it can be seen that with decreasing $\epsilon$ the energy accuracy improves, which could also be observed in the two previous cases. Taking shot noise into account, some minor fluctuations around the optimal curve for $\epsilon = 0.1$ can be spotted. However, these fluctuations increase with decreasing $\epsilon$, besides that an offset seems to occur, especially for $\epsilon = 0.01$. While the noiseless run suggests an energy accuracy in the order of $10^{-3}$ after 50 iterations steps with the gradient-descent optimizer, this value roughly increases by an order of magnitude for the noisy runs.

The noiseless parameter-shift run shows a worse accuracy in energy after the first 50 iterations, however, it shows better results compared to the finite-difference method when shot noise is included. Fluctuations occur only around the noiseless run, whereas no offset, as in the case of $\epsilon = 0.01$ performed with the finite-difference method can be observed and thus showing an overall better performance compared to its numerical counterpart. It has to be emphasized that the total number of circuits is considerably higher for the parameter-shift rule (namely a factor of 24) for this particular system and consequently requires a lot more time to finish the 50 iterations compared to the finite-difference method. Therefore, a compromise between runtime and accuracy in energy is a plausible conclusion. Having a slightly worse accuracy but therefore requiring a smaller number of additional circuits, one might choose the finite-difference method with $\epsilon = 0.05$.

## IV. CONCLUSION

In this work, two possibilities for gradient determination were investigated and tested on a simple toy model and the 1D Hubbard model using the VHA and a simple gradient-descent algorithm for parameter optimization.

On the one hand, gradients were determined numerically with a simple finite-difference method, where the difference between the obtained and exact minimum of the cost function for noiseless runs gets smaller with decreasing step sizes $\epsilon$. The number of circuit evaluations scales only with the number of parameters defined in the parameter set and thus shows a fast runtime compared to the parameter-shift method.

On the other hand, analytical gradients, determined with the parameter-shift rule, lead to more accurate results and a more resilient behavior towards statistical noise. Especially in the case of the six-site Hubbard model, a better accuracy can be achieved with the parameter-shift rule when simulating the system taking into account shot noise. However, a major drawback of the parameter-shift rule, applied to the optimization of a VHA circuit, is its scaling. Since the VHA does not only scale with the number of parameters but also with the system size (which affects the number of parametrized gates), the number of required measurements increases linearly with the system size as detailed in Sec. II G. For example, in the six-site case the number of circuit evaluations is almost two orders of magnitude higher for the parameter-shift method compared to the numerical counterpart. Thus, a compromise between runtime and accuracy can be made for the finite-difference method, where the step size $\epsilon$ is chosen optimally to avoid statistical noise and provide the required accuracy. The obvious downside of this approach is that this optimal step width has to be determined and in general depends on the number of measurements and other hyperparameters of the optimization algorithm. Taking a look at the simple toy model and the two-site Hubbard model it can be concluded that the finite-difference method is the method of choice when weighing both runtime and accuracy. Even with shot and depolarization noise for both cases, there is no substantial difference between optimization runs performed with the finite-difference method choosing small step sizes $\epsilon$ and the parameter-shift rule.

We emphasize that the results presented in this work apply to the generic situation where the trial state is constructed with the goal of reducing the number of parameters optimized in the classical optimization loop of a quantum-classical hybrid optimization procedure, which leads to the important difference of number of parameters and number of parametrized gates when analyzing the measurement costs when considering the parameter-shift rule. One motivation for reducing the number of parameters in the trial state is given by the fact that the so-called Barren plateaus [39,40] hamper the optimization of functions defined in high-dimensional spaces using gradient-based algorithms. However, the trigonometric building blocks of any quantum circuit, highlighted in the discussion of the parameter-shift rule in Sec. II D, suggests to use alternative (gradient-free) optimization algorithms [41].

We stress, however, that our conclusions do not apply to *generic* variational quantum circuits, which recently gained considerable attention in the context of quantum machine learning [42–45]. These generic quantum circuits are directly composed using a universal gate set, e.g., CNOT and arbitrary single qubit rotations [26], so the distinction of number of parameters and number of parametrized gates does not apply (all parametrized gates are treated as independent, so the number of parametrized gates corresponds trivially to the number of variational parameters). The additional structure imposed by simulating fermions via the fermion-qubit mapping (Jordan-Wigner transformation etc.) and using parts of the Hamiltonian to generate the variational unitaries is responsible for the overhead of the parameter-shift rule, compared to the finite-difference method, in our studies.

Finally, this suggests another possibility to generate the trial state, i.e., using the VHA to generate the circuit, but optimizing each parametrized gate individually in the optimization procedure. An interesting question for future studies is to investigate whether treating each parametrized gate independently considerably improves the minimal achievable energy compared to the standard VHA.

## ACKNOWLEDGMENTS

This work was supported by the Federal Ministry of Economic Affairs and Energy, Germany, as part of the PlanQK project (01MK20005H) developing a platform and ecosystem for quantum-assisted artificial intelligence (see planqk.de).

## APPENDIX: DERIVATION OF THE PARAMETER-SHIFT RULE

In this Appendix we present an explicit derivation of Eq. (9). We start by choosing a specific one-qubit rotation $\hat{R}_{\theta_i}$ to decompose the parametrized unitary quantum circuit into

$$\hat{U}(\boldsymbol{\theta}) = \hat{V}\hat{R}_{\theta_i}\hat{W}, \tag{A1}$$

where $\hat{V}$ and $\hat{W}$ are unitary rotations parametrized by all angles $\theta_j$ except for the angle $\theta_i$, which *only* parametrizes the selected single-qubit gate. The single-qubit rotation $\hat{R}_{\theta_i}$ is explicitly given by

$$\hat{R}_{\theta_i} = e^{-i\theta_i \boldsymbol{n}\cdot\boldsymbol{\sigma}} = \cos(\theta_i)\mathbf{1} - i\sin(\theta_i)\boldsymbol{n}\cdot\boldsymbol{\sigma}, \tag{A2}$$

where $\boldsymbol{n}$ is a unit vector defining the rotation axis (e.g., Cartesian unit vectors would result in rotations around the $x$, $y$, or $z$ axis), $\theta_i$ provides the rotation angle, and $\boldsymbol{\sigma}$ denotes the vector

of elementary Pauli matrices. Note that the Pauli matrices are, in principle, also labeled by the qubit index on which they act, which here is suppressed to keep the notation concise.

The cost function (2) can be rewritten as

$$E(\boldsymbol{\theta}) = \sum_m \langle\psi_0|\,\hat{U}(\boldsymbol{\theta})^\dagger \hat{O}_m \hat{U}(\boldsymbol{\theta})\,|\psi_0\rangle\,, \tag{A3}$$

highlighting that the energy is synthesized by measuring several Hermitian operators $\hat{O}_m$. Defining $|\Psi\rangle = \hat{W}\,|\psi_0\rangle$ and $\hat{O}'_m = \hat{V}^\dagger \hat{O}_m \hat{V}$ a single term of the sum (A3) is given by

$$E_m = \langle\Psi|\,\hat{R}^\dagger_{\theta_i} \hat{O}'_m \hat{R}_{\theta_i}\,|\Psi\rangle\,. \tag{A4}$$

Using Eq. (A2) this leads to

$$\begin{aligned} E_m = {} & \langle\Psi|\,[\cos(\theta_i)\mathbf{1} + i\sin(\theta_i)\boldsymbol{n}\cdot\boldsymbol{\sigma}]\hat{O}'_m \\ & \times [\cos(\theta_i)\mathbf{1} - i\sin(\theta_i)\boldsymbol{n}\cdot\boldsymbol{\sigma}]\,|\Psi\rangle \\ = {} & \cos^2(\theta_i)\,\langle\Psi|\,\hat{O}'_m\,|\Psi\rangle \\ & + \sin(\theta_i)\cos(\theta_i)\,\langle\Psi|\,i[\boldsymbol{n}\cdot\boldsymbol{\sigma}, \hat{O}'_m]\,|\Psi\rangle \\ & + \sin^2(\theta_i)\,\langle\Psi|\,\boldsymbol{n}\cdot\boldsymbol{\sigma}\hat{O}'_m\boldsymbol{n}\cdot\boldsymbol{\sigma}\,|\Psi\rangle\,, \end{aligned}$$

where all three expectation values are given in terms of Hermitian operators. By virtue of standard trigonometric identities this can be expressed as

$$E_m = A_m \cos(2\theta_i + \varphi_m) + C_m. \tag{A5}$$

We stress that $A_m$, $\varphi_m$, and $C_m$ depend on all angles $\theta_j$ with $j \neq i$ since our initial assumption stated that $\theta_i$ does only appear in the explicitly selected single-qubit gate. The generalization to Pauli matrices scaled by a factor $s$ can be obtained by replacing $\theta_i \to s\theta_i$. Defining $\omega = 2s$ and summing all constant contributions $C_m$ leads to Eq. (9).


[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).

[2] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, P. Hu, X.-Y. Yang, W.-J. Zhang, H. Li, Y. Li, X. Jiang, L. Gan, G. Yang, L. You, Z. Wang *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[3] J. M. Pino, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, M. S. Allman, C. H. Baldwin, M. Foss-Feig, D. Hayes, K. Mayer, C. Ryan-Anderson, and B. Neyenhuis, Demonstration of the qccd trapped-ion quantum computer architecture, Nature (London) **592**, 209 (2021).

[4] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[5] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).

[6] J.-M. Reiner, F. Wilhelm-Mauch, G. Schön, and M. Marthaler, Finding the ground state of the Hubbard model by variational methods on a quantum computer with gate errors, Quantum Sci. Technol. **4**, 035005 (2019).

[7] N. Vogt, S. Zanker, J.-M. Reiner, M. Marthaler, T. Eckl, and A. Marusczyk, Preparing ground states with a broken symmetry with variational quantum algorithms, Quantum Sci. Technol. **6**, 035003 (2021).

[8] J. Romero, R. Babbush, J. R. McClean, C. Hempel, P. J. Love, and A. Aspuru-Guzik, Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz, Quantum Sci. Technol. **4**, 014008 (2018).

[9] A. L. Fetter and J. D. Valecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).

[10] G. F. Giuliani and G. Vignale, "*The perturbative calculation of linear response functions*," in *Quantum Theory of the Electron Liquid* (Cambridge University Press, Cambridge, England, 2005), Chap. 6, pp. 275–326.

[11] M. J. Powell, A view of algorithms for optimization without derivatives, Mathematics Today-Bulletin of the Institute of Mathematics and its Applications, Technical Report DAMTP 2007/NA03, CMS, University of Cambridge, Cambridge, UK, April, 2007.

[12] M. J. D. Powell, A direct search optimization method that models the objective and constraint functions by linear interpolation, in *Advances in Optimization and Numerical Analysis*, edited by S. Gomez and J.-P. Hennart (Springer Netherlands, Dordrecht, The Netherlands, 1994), pp. 51–67.

[13] J. Nocedal and S. Wright, *Numerical Optimization* (Springer Science & Business Media, New York, 2006).

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

[15] A. Mari, T. R. Bromley, and N. Killoran, Estimating the gradient and higher-order derivatives on quantum hardware, Phys. Rev. A **103**, 012405 (2021).

[16] G. E. Crooks, Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition, arXiv:1905.13311.

[17] L. Banchi and G. E. Crooks, Measuring analytic gradients of general quantum evolution with the stochastic parameter shift rule, Quantum **5**, 386 (2021).

[18] J. J. Meyer, J. Borregaard, and J. Eisert, A variational toolbox for quantum multi-parameter estimation, npj Quant. Inf. **7**, 89 (2021).

[19] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature (London) **549**, 242 (2017).

[20] J. J. S. Sakurai and J. J. Napolitano, *Modern Quantum Mechanics* (Pearson, London, 2014).

[21] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, Rev. Mod. Phys. **92**, 015003 (2020).

[22] C. G. Broyden, The convergence of a class of double-rank minimization algorithms 1. General considerations, IMA J. Appl. Math. **6**, 76 (1970).

[23] R. Fletcher, A new approach to variable metric algorithms, Comput. J. **13**, 317 (1970).

[24] D. Goldfarb, A family of variable-metric methods derived by variational means, Mathematics of Computation **24**, 23 (1970).

[25] D. F. Shanno, Conditioning of quasi-newton methods for function minimization, Mathematics of Computation **24**, 647 (1970).

[26] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, England, 2010).

[27] C. King, The capacity of the quantum depolarizing channel, IEEE Trans. Inf. Theory **49**, 221 (2003).

[28] M. H. Devoret and R. J. Schoelkopf, Superconducting circuits for quantum information: an outlook, Science **339**, 1169 (2013).

[29] R. Schutjens, F. A. Dagga, D. J. Egger, and F. K. Wilhelm, Single-qubit gates in frequency-crowded transmon systems, Phys. Rev. A **88**, 052330 (2013).

[30] C. D. Bruzewicz, J. Chiaverini, R. McConnell, and J. M. Sage, Trapped-ion quantum computing: Progress and challenges, Appl. Phys. Rev. **6**, 021314 (2019).

[31] N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Experimental comparison of two quantum computing architectures, Proc. Natl. Acad. Sci. USA **114**, 3305 (2017).

[32] C. Rigetti, J. M. Gambetta, S. Poletto, B. L. T. Plourde, J. M. Chow, A. D. Córcoles, J. A. Smolin, S. T. Merkel, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, Superconducting qubit in a waveguide cavity with a coherence time approaching 0.1 ms, Phys. Rev. B **86**, 100506(R) (2012).

[33] S. Novikov, J. E. Robinson, Z. K. Keane, B. Suri, F. C. Wellstood, and B. S. Palmer, Autler-townes splitting in a three-dimensional transmon superconducting qubit, Phys. Rev. B **88**, 060503(R) (2013).

[34] L. B. Nguyen, Y.-H. Lin, A. Somoroff, R. Mencia, N. Grabon, and V. E. Manucharyan, High-Coherence Fluxonium Qubit, Phys. Rev. X **9**, 041041 (2019).

[35] B. Bauer, D. Wecker, A. J. Millis, M. B. Hastings, and M. Troyer, Hybrid Quantum-Classical Approach to Correlated Materials, Phys. Rev. X **6**, 031045 (2016).

[36] F. Vatan and C. Williams, Optimal quantum circuits for general two-qubit gates, Phys. Rev. A **69**, 032315 (2004).

[37] P. Jordan and E. Wigner, Über das paulische äquivalenzverbot, Z. Phys. **47**, 631 (1928).

[38] M. Matlak, J. Aksamit, B. Grabiec, and W. Nolting, Hubbard hamiltonian in the dimer representation large-u case, Ann. Phys. (Leipzig) **12**, 304 (2003).

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

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

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

[42] J. Romero, J. P. Olson, and A. Aspuru-Guzik, Quantum autoencoders for efficient compression of quantum data, Quantum Sci. Technol. **2**, 045001 (2017).

[43] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, arXiv:1802.06002.

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

[45] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of Barren Plateaus in Quantum Convolutional Neural Networks, Phys. Rev. X **11**, 041011 (2021).