Challenges of variational quantum optimization with measurement shot noise

Giuseppe Scriva [1,2,3,*] Nikita Astrakhantsev [4] Sebastiano Pilati [1,3] and Guglielmo Mazzola [2]

[1] *Physics Division, School of Science and Technology, University of Camerino, Via Madonna delle Carceri 9, I-62032 Camerino (MC), Italy*

[2] *Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*

[3] *INFN Sezione di Perugia, Via A. Pascoli, I-06123 Perugia, Italy*

[4] *Department of Physics, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*

(Received 3 October 2023; accepted 7 February 2024; published 11 March 2024)

Quantum enhanced optimization of classical cost functions is a central theme of quantum computing due to its high potential value in science and technology. The variational quantum eigensolver (VQE) and the quantum approximate optimization algorithm (QAOA) are popular variational approaches that are considered the most viable solutions in the noisy-intermediate scale quantum (NISQ) era. Here, we study the scaling of the quantum resources, defined as the required number of circuit repetitions, to reach a fixed success probability as the problem size increases, focusing on the role played by measurement shot noise, which is unavoidable in realistic implementations. Simple and reproducible problem instances are addressed, namely, the ferromagnetic and disordered Ising chains. Our results show that: (1) VQE with the standard heuristic *Ansatz* scales comparably to direct brute-force search when energy-based optimizers are employed. The performance improves at most quadratically using a gradient-based optimizer. (2) When the parameters are optimized from random guesses, also the scaling of QAOA implies problematically long absolute runtimes for large problem sizes. (3) QAOA becomes practical when supplemented with a physically inspired initialization of the parameters. Our results suggest that hybrid quantum-classical algorithms should possibly avoid a brute force classical outer loop, but focus on smart parameters initialization.

DOI: 10.1103/PhysRevA.109.032408

I. INTRODUCTION

Optimization is one of the most anticipated applications of quantum computers due to its commercial value and widespread use in scientific and technological applications [1]. The first argument supporting the benefit of quantum optimization is its ability to search through an exponentially large computational space, of size $N = 2^L$, using only L qubits. However, such memory compression alone is not sufficient, as the solution to a classical combinatorial optimization problem is represented by a single (or very few) L-bit string. This is in contrast with quantum algorithms for solving genuinely quantum mechanics problems, where the source of possible quantum advantage is easier to rationalize [2]. The quantum computational resource enabling the search is interference. The process begins with a simple, easy-to-prepare quantum state, which undergoes unitary evolution. Ideally, the result of this evolution is such that, when the state is measured, the desired bit string is observed with a high probability [3].

It is still unclear whether quantum optimization offers any advantage over the existing classical methods, such as simulated annealing [4]. Interestingly, optimization with quantum annealing has been the first application of commercial quantum devices [5,6], which mostly rely on incoherent tunneling events to escape the cost-function local minima [7]. However, it is not easy to prove systematic quantum speedups with analog quantum annealers [8,9], also because quantum Monte Carlo algorithms appear to be able to emulate their tunneling dynamics [10–13]. Yet, considerable effort is still ongoing in improving the architecture of these machines [14] and their coherence times [15].

As an alternative quantum optimization strategy, variational quantum algorithms, usually running on digital quantum devices, have gained attention in the quantum computing community due to their short-depth circuits [16,17]. In this approach, a long quantum state evolution is replaced by a series of short-depth quantum circuits connected through a classical feedback loop. Variational quantum computation features parametrized circuits that produce a trial state $|\psi(\boldsymbol{\theta})\rangle$. Its parameters $\boldsymbol{\theta}$ are adjusted at every step following an iterative classical procedure. The goal is to minimize a cost function C, which corresponds to the expectation value $\langle\psi_{\boldsymbol{\theta}}|\,\hat{H}_p\,|\psi_{\boldsymbol{\theta}}\rangle$ of the problem Hamiltonian $\hat{H}_p$, or a closely related measure. At the end of a successful optimization, $|\psi(\boldsymbol{\theta})\rangle$ should be peaked around the solution of the problem.

The two most popular variational algorithms for optimization are the quantum approximate optimization algorithm (QAOA) [18] and the variational quantum eigensolver (VQE) [16]. Both of them include a parametrized circuit, a classical feedback loop, and a measurement stage. The cost function is evaluated based on the measurement's outcome, and the parameters are adjusted to minimize the cost.

Let us also recall that for combinatorial optimization problems, like Ising spin glasses on general graphs [19], no polynomial-time algorithm can provably find the global minimum, and the resources to exactly solve these problems scale exponentially with problem size as $\sim 2^{kL}$. This is the type of

*giuseppe.scriva@unicam.it

speedup investigated in this article. While quantum algorithms are not expected to turn the exponential scaling into a polynomial one, the exponent k might be reduced, thus potentially realizing a substantial speedup over classical algorithms [7].

The QAOA method has been the subject of intense studies, including small- and medium-scale hardware experiments [20–23], numerical studies, and theoretical works [24–31]. Also, VQE optimization has been studied numerically and experimentally [32–38], and it has been applied to diverse combinatorial optimization problems from protein folding to finance [39–42]. However, these previous studies addressed small problem instances, without properly accounting for measurement shot noise. In fact, the latter is unavoidable in physical implementations of practically relevant problem sizes and it might affect the computational complexity of these algorithms. To the best of our knowledge, the scaling of the computational cost for a fixed target success probability, taking into account the measurement overhead to compute the cost function C, has not been exhaustively addressed yet.

The paper is organized as follows. In Sec. II, we define the testbed problems and the quantum circuits. In Sec. III, we introduce the metric to properly assess the computational scaling of the VQE and QAOA algorithms in realistic conditions. In Sec. IV A, it is shown that in the presence of quantum measurement noise, VQE displays a scaling not better than the direct space enumeration when energy-based optimizers are used. The situation improves using gradients, computed with the parameters shift rule (see Sec. IV B), but it remains scaling-wise impractical. In Sec. IV C, it is shown that, while showing some scaling improvements, QAOA remains impractical when a full optimization outer loop is required. In this case, the traditional energy-based and a gradient-based optimizer show consistent scalings. Finally, in Sec. IV E, we show that QAOA becomes competitive when the parameters are initialized to mimic an adiabatic process. In Sec. V, we draw conclusions and discuss realistic pathways toward quantum advantage in classical optimization problems.

II. OPTIMIZATION PROBLEMS AND QUANTUM CIRCUITS

The optimization problems we address correspond to the Ising models defined over L variables $(\sigma_1, \ldots, \sigma_L) = \boldsymbol{\sigma}$ with $\sigma_j = \pm 1$. Specifically, we consider the one-dimensional connectivity, nearest-neighbor interactions $J_{j,j+1}$, and local fields $\{h_j\}_{j=1}^{L}$. The energy of a spin configuration $\boldsymbol{\sigma}$ reads

$$E(\boldsymbol{\sigma}) = -\sum_{j=1}^{L-1} J_{j,j+1}\sigma_j\sigma_{j+1} - \sum_{j=1}^{L} h_j\sigma_j. \tag{1}$$

Representing a generic spin configuration $\boldsymbol{\sigma} \in \{1, -1\}^L$ as a binary string $\boldsymbol{x} \in \{0, 1\}^L$, and writing the energy as $E(\boldsymbol{\sigma}) \to f(\boldsymbol{x})$, we write the problem Hamiltonian $\hat{H}_{\mathrm{P}}$ as a diagonal operator

$$\hat{H}_{\mathrm{P}} = \sum_{x} f(x)|x\rangle\langle x|, \tag{2}$$

defined by its diagonal matrix elements $f : \{0, 1\}^L \to \mathbb{R}$. The classical spin variables σ_j are promoted to single-qubit Pauli operators $\hat{\sigma}_j^z$.

Most analyses reported in this article consider two problem Hamiltonians. The first is the *ferromagnetic* Hamiltonian defined by uniform couplings $J_{j,j+1} = J = 1$, and a (small) uniform local field $h_j = h = -0.05$ introduced to break the degeneracy between the two fully polarized configurations and obtain a single global minimum. Despite its simplicity, this model turns out to be hard for most of the considered algorithms. Its rugged energy surface $f(x)$ is shown in Fig. 1, where the bitstrings are sorted in the lexicographic order.

The second optimization problem we address is an ensemble of *disordered* Hamiltonians where the couplings and fields are sampled from a normal distribution with zero mean and unit variance: $J_{j,j+1}, h_j \sim \mathcal{N}(0, 1)$. In this case, 30 realizations of the disorder are simulated for each problem size L.

A. VQE with heuristic circuit

Parametrized quantum circuits [16,17] are the essential ingredients of any variational quantum algorithm. These circuits employ parametrized gates, including the single-qubit rotation gates, and multiqubit entangling gates such as the CNOT gate. The set of variational parameters $\boldsymbol{\theta}$ is optimized in a classical outer loop [16] to minimize a target cost function.

The most commonly studied heuristic circuit is made of d blocks built from a layer of single-qubit rotations $U_{\mathrm{R}}(\boldsymbol{\theta}^l)$ with $l = 1, \ldots, d+1$ and an entangling block U_{ent} that covers the whole qubit register (see Fig. 1). In this article, we consider the entangling block made of a ladder of CNOT gates with linear connectivity, such that the qubit q_{j-1} controls the target qubit q_j, and the latter controls the qubit q_{j+1}, obeying open boundary conditions. This choice is commonly used as it mimics the existing sparse qubit connectivity of the quantum hardware. The layer of single-qubit rotations $U_{\mathrm{R}}(\boldsymbol{\theta}^l)$ acts locally and it corresponds to a tensor product of single-qubit rotations:

$$U_{\mathrm{R}}(\boldsymbol{\theta}^l) = \bigotimes_{j=1}^{L} R_y\left(\theta_j^l\right), \tag{3}$$

where $R_y(\theta_j^l) = \exp\left(-i\theta_j^l\hat{\sigma}^y/2\right)$ is a rotation around the y axis of the Bloch sphere of the qubit q_j, and $l = 1, \ldots, d+1$. Here, $\boldsymbol{\theta}^l$ denotes an array of L angles. The full unitary circuit operation is described by

$$U_{\mathrm{R-CNOT}}(\boldsymbol{\theta}) = U_{\mathrm{R}}(\boldsymbol{\theta}^{d+1}) \overbrace{U_{\mathrm{ent}}U_{\mathrm{R}}(\boldsymbol{\theta}^d)\ldots U_{\mathrm{ent}}U_{\mathrm{R}}(\boldsymbol{\theta}^1)}^{\text{d-times}}, \tag{4}$$

and the final parametrized state reads

$$|\psi(\boldsymbol{\theta})\rangle = U_{\mathrm{R-CNOT}}(\boldsymbol{\theta})(|0\rangle^{\otimes L}). \tag{5}$$

The total number of variational parameters is $n_{\mathrm{par}} = L(d+1)$. Notice that we do not use symmetries nor prior knowledge of the optimization problem in building the circuit up.

B. The QAOA circuit

QAOA can be understood as a digitized version of quantum annealing [18] that requires variational optimization of circuit parameters. These parameters can be seen as the optimizable time steps that control the evolution of the state under the action of the *problem* and the *mixing* operators in

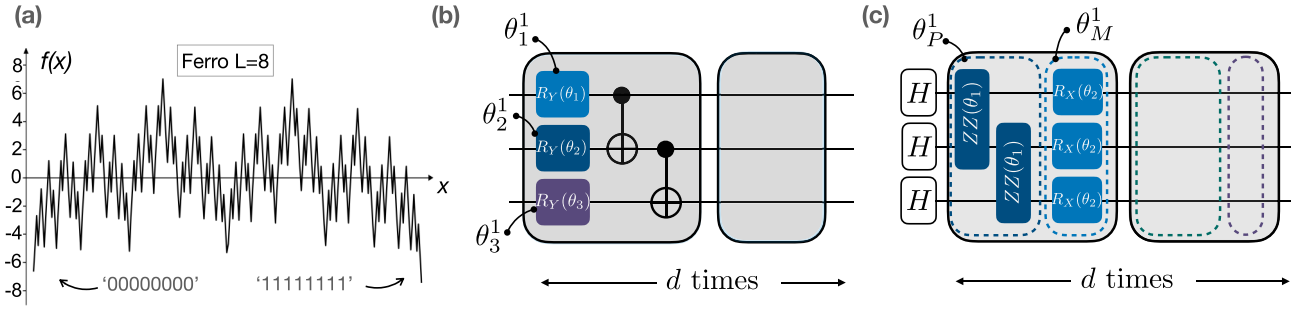

FIG. 1. (a) The energy landscape in the computational basis, where items are sorted in the lexicographical order for a ferromagnetic model with $L = 8$. The small uniform field breaks the degeneracy between the "00 . . . 0" and "11 . . . 1" bitstrings, with the latter being the global minimum. (b) Sketch of RY-CNOT circuits, i.e., a circuit consisting only of y rotations and CNOT gates, used in the VQE and commonly employed in related literature. (c) Sketch of QAOA circuit featuring the specific problem Hamiltonian. In both cases we only show the gate decomposition of the first block. The circuit features d blocks, which have the same structure but contain independent variational parameters.

a Trotterized fashion. Notice that the QAOA method precisely dictates the structure of the quantum circuits, while VQE can be implemented with any parametrized quantum circuit. In particular, the classical Hamiltonian (i.e., the cost function) explicitly appears in the QAOA circuit, while a VQE circuit may be completely heuristic, with the problem Hamiltonian informing the whole algorithm only through the evaluation of the cost function after the wave function collapses.

The unitary operator defining the *Ansatz* is made of d blocks, each of them being the product of two unitary operators $\hat{U}_{\rm P} = \exp(i\theta^l_{\rm P}\hat{H}_{\rm P})$, and $\hat{U}_{\rm M} = \exp(i\theta^l_{\rm M}\hat{H}_{\rm M})$, with $l = 1, \ldots, d$ and where $\hat{H}_{\rm P}$ is the problem Hamiltonian, and

$$\hat{H}_{\rm M} = \sum_{j=1}^{L} \hat{\sigma}^x_j \tag{6}$$

is the nondiagonal *mixing* operator.

The implementation of these unitary operators involves efficient single-qubit rotations along the x axis, denoted as $R_x(\theta) = \exp(i\theta\hat{\sigma}^x/2)$, and two-qubit parametrized gates, $R_{zz}(\theta) = \exp(i\theta\hat{\sigma}^z \otimes \hat{\sigma}^z/2)$. The structure of the QAOA *Ansatz* implies that all the local $\hat{\sigma}^z \otimes \hat{\sigma}^z$ interactions within the same block are "evolved" with the same time step θ^l_P, while all the x rotations within the block are parametrized by the same angle θ^l_M (see Fig. 1). The total number of parameters is $n_{\rm par} = 2d$, i.e., is independent of the problem size L, and the full unitary operator reads

$$U_{\rm QAOA}(\boldsymbol{\theta}) = \overbrace{\hat{U}_{\rm M}(\theta^d_M)\hat{U}_{\rm P}(\theta^d_P)\ldots\hat{U}_{\rm M}(\theta^1_M)\hat{U}_{\rm P}(\theta^1_P)}^{d\text{-times}}. \tag{7}$$

The final parametrized state is

$$|\psi(\boldsymbol{\theta})\rangle = U_{\rm QAOA}(\boldsymbol{\theta})\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes L}, \tag{8}$$

where the initial nonentangled state can be obtained from the state $|0\rangle^{\otimes L}$ by acting with one Hadamard gate on each qubit.

III. RESOURCE COUNTING AND SCALING ANALYSIS

A. Statistical noise in evaluating the cost function

The expectation value of $\hat{H}_{\rm P}$ over the prepared state is given by the sum of all spin configurations

$$\tilde{C} = \langle\psi_{\boldsymbol{\theta}}|\hat{H}_{\rm P}|\psi_{\boldsymbol{\theta}}\rangle = \sum_{x=0}^{2^L-1} |\psi_{\boldsymbol{\theta}}(x)|^2 f(x). \tag{9}$$

In a realistic setting, the full sum needs to be necessarily approximated using a finite sample of configurations

$$\tilde{C} \approx \frac{1}{M}\sum_{i=1}^{M} f(x_i), \tag{10}$$

where x_i are sampled from $|\psi_{\boldsymbol{\theta}}(x)|^2$. The precision of this estimate is affected by statistical noise induced by the finite number of quantum measurements M. The error in estimating $\tilde{C}$ scales as $1/\sqrt{M}$, following the law of large numbers. We denote this as quantum measurement noise. This noise is very different from hardware noise, produced by the qubit's imperfection, as it is rooted in the measurement process of wave functions. Each quantum measurement requires a circuit repetition.

In numerous studies, Eq. (9) is evaluated exactly, which is dubbed the *state-vector* simulation. Instead, in our analysis we account for the effects of the finite M.

It has been empirically shown that better performances for optimization problems can be obtained by considering the conditional value at risk (CVaR) estimator of Ref. [32], in which the cost function is evaluated by summing only over the best 25% of observed outcomes $f(x_i)$:

$$C = \frac{1}{M^*}\sum_{i=1}^{M^*} f(x_i). \tag{11}$$

Operatively, the M readouts are sorted in nondecreasing order following their output $f(x_i)$, and only $M^* = M/4$ samples corresponding to the 25% lowest values are retained. The value C represents the cost function that is optimized at each iteration. We can also keep track of the current minimum observed value $f_{\rm min}$, which is generally smaller than C. Its final value is compared with the exact global minimum of the optimization problem to determine the success rate of the

algorithm. Notice, however, that also when using CVaR one needs to draw M samples.

B. Optimal scaling

The time complexity of an optimization algorithm can be expressed as the number of function calls $f(x)$ necessary to find the optimum, aiming at a *fixed* success probability as the problem sizes increase. Each evaluation of the cost function requires M circuit repetitions.

The total number of function calls required for a full optimization run is therefore

$$n_{\text{calls}} = n_{\text{iter}} \times M, \tag{12}$$

where n_{iter} is the number of (classical) optimization steps. The total runtime of the algorithm is proportional to n_{calls}. A lower bound is given by $t_{\text{run}} = n_{\text{calls}} \times d \times t_{\text{gate}}$, where again d is the circuit depth, expressed as the number of repetitions of a minimal unit (called block) of quantum gates, and t_{gate} is the time to execute each block. The value of t_{gate} strongly depends on the hardware. In the noisy-intermediate scale quantum (NISQ) era, the gate times can be of order 10 ns (100 MHz) for superconducting hardware [43], while digital gate time is predicted to be about 0.1 ms (10 kHz) in the fault-tolerant regime [44]. These estimates neglect the qubit reset time, the classical communication, and the measurement time, so they clearly represent optimistic perspectives.

For each problem size, there exists a trade-off between the number of iterations n_{iter} needed to converge to the global minimum and the number M of measurements, which controls the accuracy in evaluating the cost function at each step. Large errors in C may imply slower convergence since the cost function landscape is not correctly reproduced, thus negatively affecting the performance of the classical optimization algorithm.

One of the merits of the present study is the systematic identification of the minimum number of calls, defined as n^*_{calls}, corresponding to the optimal combination of n_{iter} and M for each problem size L, thus enabling a proper scaling analysis. This concept is similar to the optimal time-to-solution metric developed in quantum annealing [9]. We point out that one must have $n^*_{\text{calls}} < 2^L$ to avoid quantum disadvantage [45], without even discussing the values of t_{gate}.

With the definitions given above, Eq. (12) can be used to compute the number of function calls only in the case of so-called energy-based optimizers. However, in this article, we also consider gradient-based methods (see Secs. IV B and IV D). In this case, one needs to compute a n_{par}-valued array of energy derivatives at each optimization step. For each parameter, two independent circuit runs need to be executed. This holds both for the parameters shift rule (in this case, when applicable, the gradients are exact) and the finite difference method. Therefore, the cost for a single iteration has to be computed as $M = 2n_{\text{par}}\tilde{M}$, where $\tilde{M}$ is the number of shots per single circuit execution.

IV. RESULTS

A. Impracticality of VQE

Here, we analyze the performance of the VQE method for the ferromagnetic problem. The circuit simulations are performed using the open-source QISKIT framework [46]. In evaluating the algorithm's efficiency, a run is considered successful when the absolute minimum is found at least once within the n_{iter} steps. This procedure is standard in benchmarking quantum devices, such as quantum annealers, versus classical optimizers [6,7,9]. The fraction F_{succ} of successful runs is estimated considering 1000 executions starting from different (random) initializations of the variational parameters. It is crucial to note that, within the VQE heuristic circuit, there is no *a priori* method for a smart initialization of the parameters. Therefore, we initialize the parameters using a random uniform distribution of $\boldsymbol{\theta}$. Moreover, optimized parameters are not transferable to different instances.

We first inspect how F_{succ} depends on the total number of function calls n_{calls}, for different problem sizes L. For each size L, several choices of shot numbers M and optimization steps n_{iter} are considered. Notice that these two parameters determine the number of function calls n_{calls} [see Eq. (12)]. Importantly, this analysis allows us to identify the minimal number n^*_{calls} for each target success rate F_{succ} and for each problem size L. This procedure is crucial to correctly assess the scaling of the computational cost with the problem size. Chiefly, it allows us to account for the role of measurement shot noise, which is enhanced for small measurement numbers M, while larger M imply a correspondingly larger computational cost for each iteration of the classical optimization algorithm.

In this section, the classical parameter optimization is performed using the constrained optimization by linear approximation (COBYLA) optimizer, a widely adopted energy-based algorithm for QAOA [32,47]. Let us also recall that the CVaR estimator of Eq. (10) is adopted. The gradient-based method, which uses the parameters shift rule, is discussed in Sec. IV B. The performance of the (COBYLA driven) VQE method, with circuit depths $d = 1,\ 2$, is shown in Fig. 2. First, we observe that, for all choices of M and n_{iter}, the value of n_{calls} required to reach a target F_{succ} is not better than the one corresponding to random search with replacement, see Fig. 2(a). Furthermore, as shown in Fig. 2(c), the minimal number of function calls n^*_{calls} displays a problematic scaling with the problem size, closely matching the exponential law $n^*_{\text{calls}} \sim 2^{kL}$ with $k \simeq 1$. This holds for all the thresholds of $0.25 \leqslant F_{\text{succ}} \leqslant 0.9$ considered in this study. Notably, VQE circuits with depths $d = 1$ and $d = 2$ display comparable scaling, suggesting that simply increasing the circuit depth does not help.

In Appendix A it is shown that hardware noise, which we simulate using a custom model in QISKIT [46], does not significantly affect this scaling.

B. Gradient-based VQE optimization

Next, we consider the VQE algorithm driven by a gradient-based algorithm, addressing again the ferromagnetic problem. The gradients are obtained using the parameter shift rule, which is applicable under certain conditions on the adopted

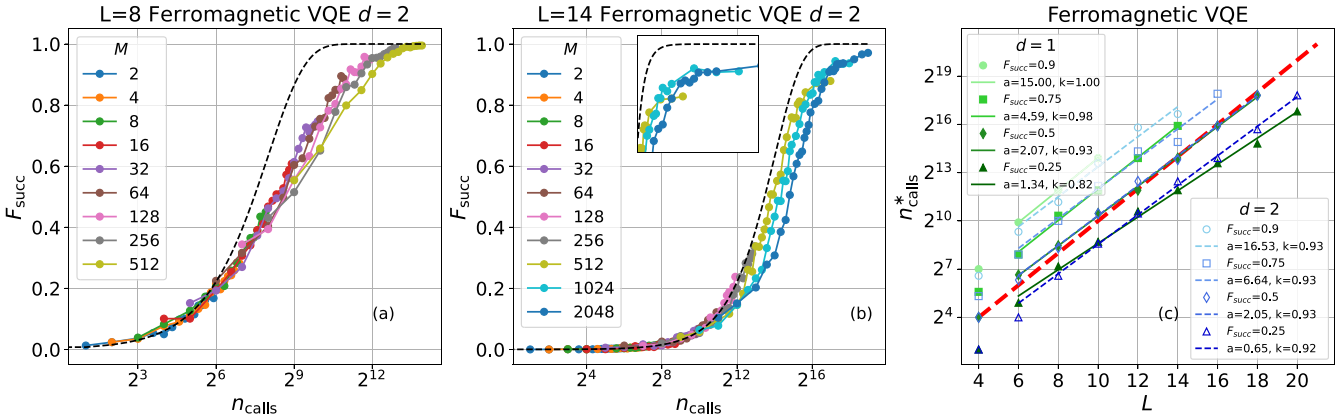

FIG. 2. Optimization of the ferromagnetic Ising chain using the VQE *Ansatz* with depth $d = 2$. (a and b): Success probability F_{succ} as a function of the total number of function calls $n_{\text{calls}} = M \times n_{\text{iter}}$, for (a) $L = 8$ spins and (b) $L = 14$ spins. Different colors correspond to the different combinations of measurement-shot number M and classical optimization steps n_{iter}. The dashed curve corresponds to the random search with replacement. The inset of (b) shows an example of the interplay between M and n_{iter}. To reach larger F_{succ} it is better to systematically increase M. For $F_{\text{succ}} \approx 0.8$, using $M = 512$ appears marginally better than $M = 1024$, which in turn becomes optimal at $F_{\text{succ}} \approx 0.9$, and so on. Crucially, each setup performs worse than the random search. (c) Minimal number of function calls n^*_{calls} as a function of the number of spins L for different F_{succ}. Circuits with $d = 1$ (full symbols) and $d = 2$ (empty symbols) blocks are considered. The thick dashed (red) line represents the scaling $n^*_{\text{calls}} \sim 2^L$ corresponding to full enumeration. Thin continuous and dashed lines represent fitting functions of the form $n^*_{\text{calls}} = a\, 2^{kL}$, and the fitting parameters a and k, obtained considering the large L regime, are given in the legend. All the quantities in this figure and the following are dimensionless.

gate set [48,49]. The n-th component of the gradient is computed as $n \in (1, \ldots, n_{\text{par}})$:

$$\frac{\partial \tilde{C}}{\partial \theta_n} = \frac{1}{2}[\langle \psi_{\boldsymbol{\theta}_n^+} | \hat{H}_{\text{P}} | \psi_{\boldsymbol{\theta}_n^+} \rangle - \langle \psi_{\boldsymbol{\theta}_n^-} | \hat{H}_{\text{P}} | \psi_{\boldsymbol{\theta}_n^-} \rangle], \tag{13}$$

where $\boldsymbol{\theta}_n^{\pm} = (\theta_1, \ldots, \theta_n \pm \pi/2, \ldots, \theta_{n_{\text{par}}})$. Notice that, in this case, the cost function is computed as in Eq. (10), rather than adopting the CVaR estimator.

At each iteration, the parameters $\boldsymbol{\theta}$ are updated as

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \eta \nabla \tilde{C}(\boldsymbol{\theta}), \tag{14}$$

where η is the learning rate. The value $\eta = 0.1$ is chosen, as it turns out to be reasonably close to optimal from a preliminary analysis on the problem size $L = 6$. As before, the optimal combination of M and n_{iter} is found, and the computational complexity is analyzed by observing the scaling of n^*_{calls} with the problem size (see Fig. 3). Interestingly, an approximately quadratic speedup compared with the COBYLA optimizer is found.

Concluding this subsection, it is worth mentioning that, in the context of quantum chemistry problems, a full quantum eigensolver has been introduced [50]. This algorithm implements gradient descent on the quantum device, avoiding the classical optimization step. Future work might focus on adapting this scheme to classical optimization problems.

C. QAOA with random parameters initialization

Here, the performance of QAOA is analyzed using the (energy-based) COBYLA optimizer. The first tests focus on the ferromagnetic model. We expect to observe a better performance compared to VQE, because QAOA features the problem Hamiltonian also in the circuit, not only in the cost function. To support this intuition, we perform a preliminary comparison, considering circuits with random variational parameters, i.e., avoiding any classical optimization iteration. Specifically, we prepare 1000 different QAOA circuits, and just as many for VQE, using uniformly distributed parameters, and sample $M = 16$ measurements from each of them. The probability F_{succ} of observing the exact solution at least once is then computed. As shown in Fig. 4, the QAOA *Ansatz* clearly outperforms VQE. We attribute this to its higher degree of localization around the correct solution, even when the parameters are random. Notice that in this analysis the choice of the optimizer is not relevant, allowing us to compare the circuits independently of the way they are optimized. One might also

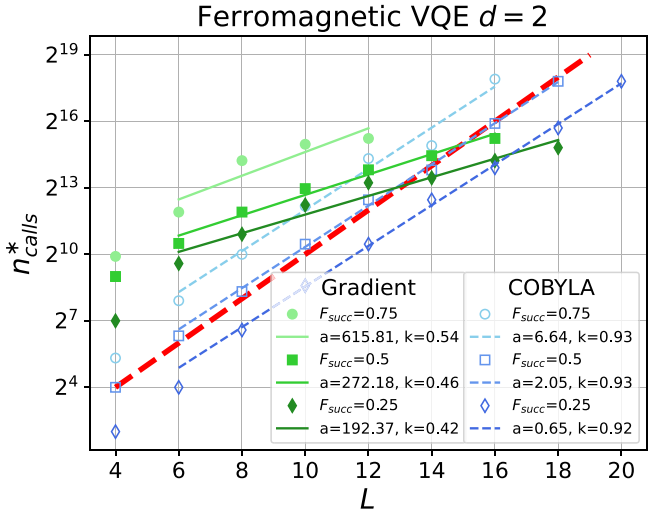

FIG. 3. Minimal number of function calls n^*_{calls} as a function of the number of spins L for different F_{succ}. Thin continuous and dashed lines represent fitting functions of the form $n^*_{\text{calls}} = a\, 2^{kL}$, and the fitting parameters a and k, obtained by fitting the large L data, are given in the keys.

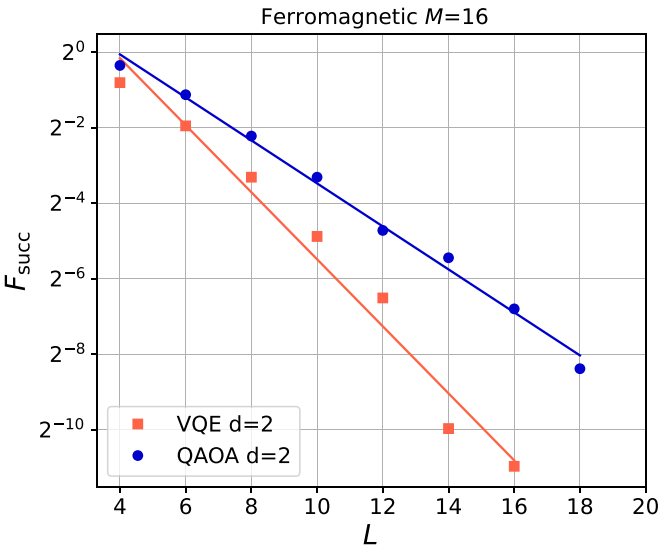

FIG. 4. Success probability F_{succ} as a function of the system size L before the classical optimization, for a fixed shot number $M = 16$. The QAOA (circles) and VQE (squares) *Ansätze* have the same depth $d = 2$ and randomly chosen parameters. The lines represent fits, obtained from the large L data, as a guide to the eye.

expect that, since the QAOA circuit features fewer parameters, it should be easier to optimize as compared to VQE [51].

To exhaustively assess the QAOA performance, we repeat the procedure described in Sec. IV A, exploring different combinations of n_{iter} and M. Again, this allows us to identify the optimal number of function calls n^*_{calls} for the target cumulative success probability F_{succ} and problem size L. For each choice of n_{iter} and M, 1000 circuit executions are performed starting from random uniformly distributed parameters. Notably, for all considered success probabilities F_{succ}, the number of function calls is well described by the exponential scaling law $n^*_{\text{calls}} \sim 2^{kL}$, with $k \simeq 0.4$. This corresponds to an approximately quadratic speedup as compared to the exact enumeration. Given that the choice of the classical optimizer may change the observed scaling, in Appendix B, we also test the simultaneous perturbation stochastic approximation (SPSA) algorithm [52]. We find that the SPSA and COBYLA results are compatible. We further test this finding on a more challenging system, namely, the disordered Ising model. The results are shown in Figs. 5(d)–5(f). Also, in this case they are averaged over 30 realizations of the random couplings and fields of the problem Hamiltonian. Similarly to the ferromagnetic case, we observe a profitable scaling, namely, $k \in [0.5, 0.8]$, to be compared with the full enumeration, corresponding to $k = 1$. However, in this case, extracting the scaling exponent is more difficult, because n_{calls} needs to be increased to reach large F_{succ}, leading to prohibitive computational times for large problem sizes. Notably, both for the ferromagnetic and the disordered problem Hamiltonians, increasing the circuit depth from $d = 2$ to 4 does not substantially affect the scaling.

To summarize the above findings, the observed QAOA scaling exponents are about $k \simeq 0.4$ for the ferromagnetic problem, and $0.5 \lesssim k \lesssim 0.8$ for the disordered models, using the COBYLA optimizer and random parameters initialization. This scaling is comparable to the one of VQE using gradients. While better than full enumeration, these scalings still determine unfeasible runtimes (see discussions in Sec. V) for problem instances of practical interest, i.e., featuring at least hundreds of spins.

D. Gradient-based QAOA optimization

Here, we benchmark the scalings of QAOA driven by the COBYLA optimizer against a gradient-based method. Contrary to the case described in Sec. IV B, the QAOA circuit does not satisfy the assumptions to apply the parameter shift rule [49]. While there are attempts to extend the parameter shift rule [53], here we adopt the finite difference approximation. The nth gradient component is computed as

$$\frac{\partial \tilde{C}}{\partial \theta_n} = \frac{1}{2\varepsilon}[\langle \psi_{\boldsymbol{\theta}_n^{+\varepsilon}} | \hat{H}_{\text{P}} | \psi_{\boldsymbol{\theta}_n^{+\varepsilon}} \rangle - \langle \psi_{\boldsymbol{\theta}_n^{-\varepsilon}} | \hat{H}_{\text{P}} | \psi_{\boldsymbol{\theta}_n^{-\varepsilon}} \rangle], \tag{15}$$

where $\boldsymbol{\theta}_n^{\pm\varepsilon} = (\theta_1, \ldots, \theta_n \pm \varepsilon, \ldots, \theta_{n_{\text{par}}})$ and $\varepsilon > 0$ is the increment. Small values reduce the finite-difference error, but they also enhance the random fluctuations due to the finite number of measurements $\tilde{M}$ used to estimate the expectation values in Eq. (15). To identify the optimal trade-off regime, we compare the estimated gradients with the exact results from state-vector simulations [see Fig. 6(a)]. For the typically optimal shot numbers $\tilde{M} \in [2, 16]$, the error is minimized for increments close to $\varepsilon = 0.5$. This value is adopted hereafter.

With the above setting, we analyze the scaling of n^*_{call} with the problem size L [see Fig. 6(b)]. Any benefit provided by the gradient turns out to be essentially compensated by its cost in terms of measurement shots. Recently, the detrimental cost of gradient estimation has been highlighted addressing the application of quantum computers for electronic structure [54]. The overall improvement compared to the scaling obtained with the COBYLA optimizer is not sizable. Furthermore, it is worth noticing that even for the larger size considered in this work, the gradient-based optimization requires larger n^*_{call}. This is the second main result of the paper: a naive textbook implementation of QAOA using random starting parameters is practically inefficient, even when gradient-based optimizers are used, despite showing an improved scaling with respect to random search.

E. QAOA with annealing-inspired parameters initialization

On one hand, the above findings indicate that QAOA is computationally unfeasible for problem sizes of practical interest. On the other hand, QAOA can be interpreted as a digitized version of quantum annealing, and previous studies have shown that the available quantum-annealing devices can already find solutions of large-scale spin-glass instances in a reasonable runtime [7] even for problem sizes as large as $L = 512$. To solve this apparent conundrum, we perform QAOA in its adiabatic limit. Formally, this limit is reached when $d \to \infty$. However, as pointed out in Sec. IV C, doubling the number of layers does not decisively change the computational scaling. In fact, the number of parameters increases with the circuit depth, and more parameters usually require more optimization iterations.

Still, the analogy with quantum annealing inspires a systematic way to effectively initialize the parameters. Indeed,

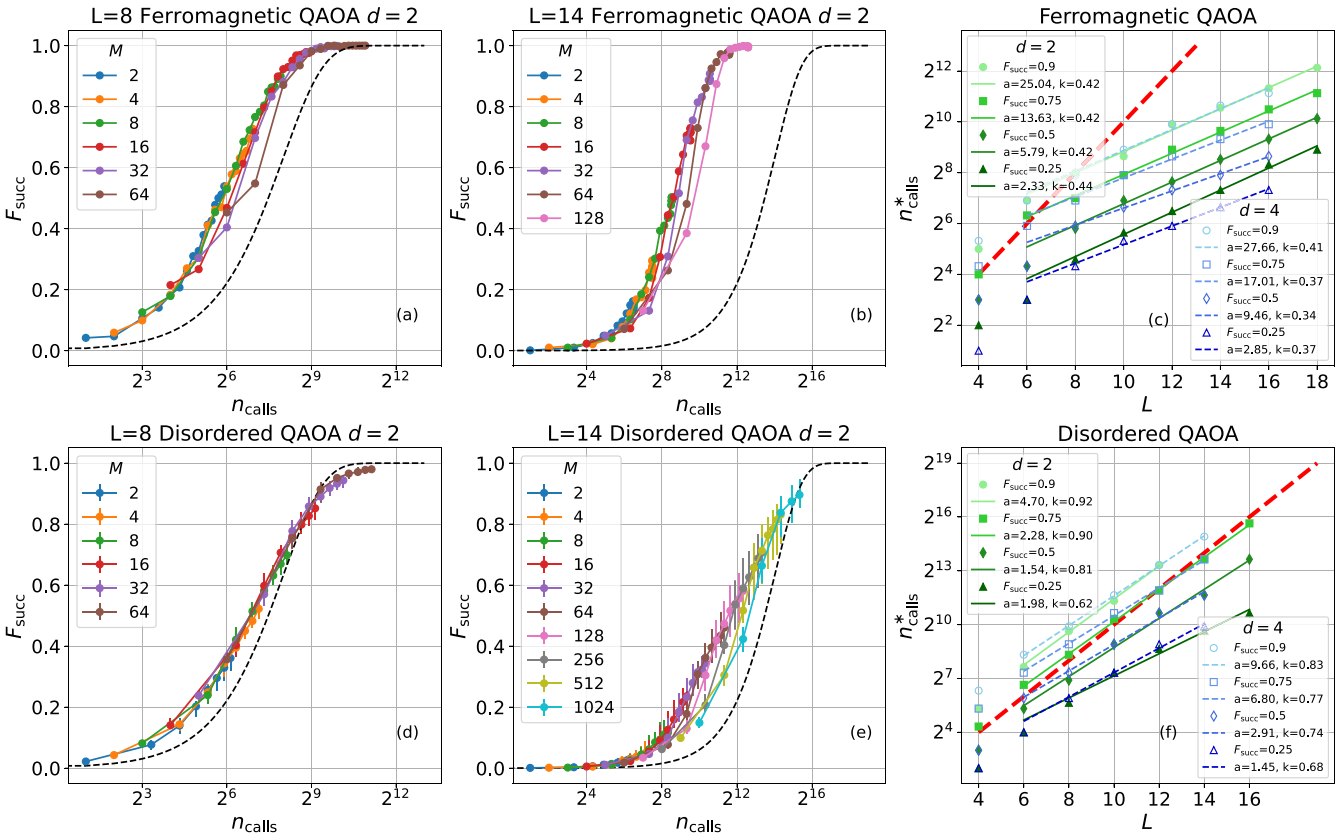

FIG. 5. Optimization of the ferromagnetic (first row) and the disordered (second row) Ising chains within the QAOA method. (a), (b), (d), and (e) Success probability $F_{\rm succ}$ as a function of the total number of function calls $n_{\rm calls}$ for (a and d) $L=8$ spins and (b and e) $L=14$ spins. The error bars indicate the 25th and the 75th percentiles. Different colors correspond to the different combinations of measurement budgets M and classical optimization-step counts $n_{\rm iter}$. The dashed curve corresponds to the random search with replacement. (c, f) The optimal number of function calls $n^*_{\rm calls}$ as a function of the number of spins L for different $F_{\rm succ}$. Circuits with $d=2$ blocks (full symbols) and with $d=4$ blocks (empty symbols) are considered. The thick dashed (red) line represents the scaling $n^*_{\rm calls}\sim 2^L$ corresponding to exact enumeration. Thin continuous and dashed lines represent fitting functions of the form $n^*_{\rm calls}=a\,2^{kL}$, and the fitting parameters a and k are obtained by fitting the large L data and are given in the keys.

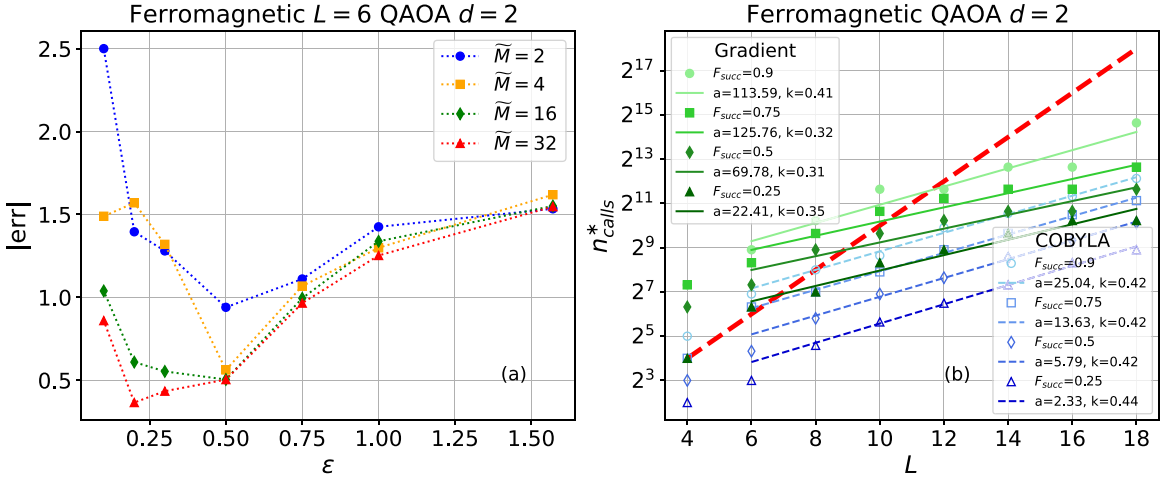

FIG. 6. (a) Absolute error |err| in estimating the gradient as a function of the step ε of the finite difference approximation. We compare results obtained with a different number of shots M. (b) The minimal number of function calls $n^*_{\rm calls}$ as a function of the number of spins L, for different success probabilities $F_{\rm succ}$. Circuits with $d=2$ blocks are considered, with gradient descent (full symbols) and with COBYLA optimizer (empty symbols). The thick dashed (red) line represents the scaling $n^*_{\rm calls}\sim 2^L$ corresponding to exact enumeration. Thin continuous and dashed lines represent fitting functions of the form $n^*_{\rm calls}=a\,2^{kL}$, and the fitting parameters a and k, obtained considering the large L data, are given in the keys.

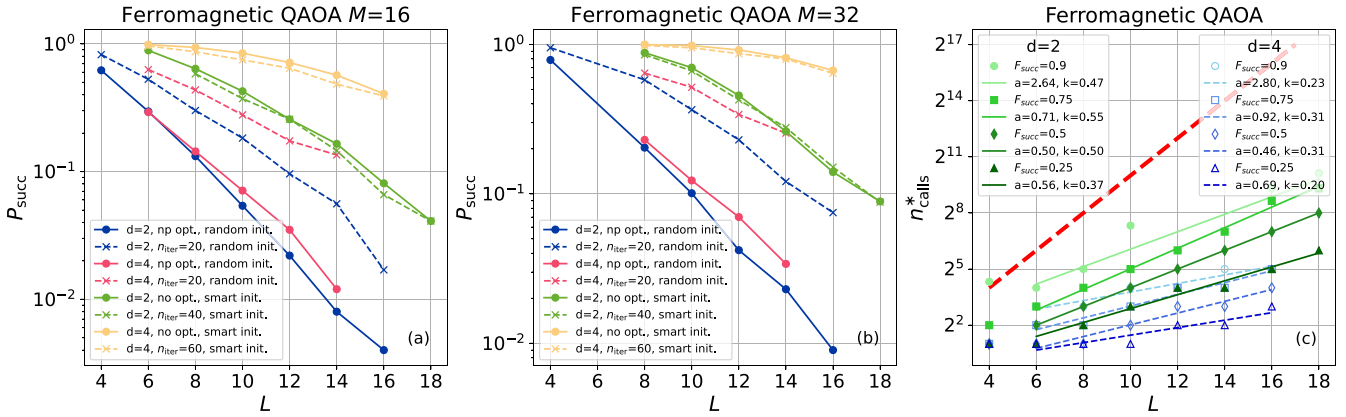

FIG. 7. Comparison of the optimizations starting from random (Sec. IV C) and annealing-inspired initializations (linear schedule, Sec. IV E) for the ferromagnetic model. (a and b) The success probability P_{succ} as a function of the number of spins L, obtained within QAOA before (continuous curves) and after (dashed curves) the classical optimization performed for n_{iter} steps. The number of shots (a) $M = 16$ and (b) $M = 32$ are considered, chosen so that $M \ll 2^L$. Note that P_{succ}, i.e., the probability to sample at least once the global minimum at the n-th iteration, is equal by definition to F_{succ} when $n_{\text{iter}} = 0$. (c) The minimal number of function calls n^*_{calls} as a function of L at fixed success probabilities F_{succ}, starting from the annealing-inspired initialization. Circuits with $d = 2$ blocks (full symbols) and with $d = 4$ blocks (empty symbols) are considered. The thick dashed (red) line represents the scaling $n^*_{\text{calls}} \sim 2^L$ corresponding to exact enumeration. The thin continuous and dashed lines represent fitting functions of the form $n^*_{\text{calls}} = a\, 2^{kL}$, and the fitting parameters a and k, obtained by fitting the large L data, are given in the key.

in Ref. [27] it was observed that, in state-vector simulations, the optimal parameters often follow a pattern similar to the quantum annealing prescription: the parameters controlling the mixing operator $\boldsymbol{\theta}_M$ decrease, while the parameters controlling the problem operator $\boldsymbol{\theta}_P$ increase with the layer index. Following this idea, we initialize the parameters using the simplest discretized linear schedule, as in Ref. [55]:

$$\theta^l_M = \left(1 - \frac{l}{d}\right)\Delta_t, \quad \theta^l_P = \frac{l}{d}\Delta_t, \tag{16}$$

where $l \in [1, \ldots, d]$. Notice that in most QAOA literature, the parameters that control the mixing operators are denoted with β, while the problem parameters are denoted with γ.

In Ref. [55], this initialization was found effective for MaxCut problems solved via state-vector simulations, i.e., eliminating the measurement shot noise. Here, we show that this initialization is not only an improvement to the QAOA textbook strategy, but it is also essential to make the algorithm practical in realistic conditions where measurement noise is accounted for. Notice that with the reparametrization in Eq. (16), the angles θ^l_M and θ^l_P depend only on one real degree of freedom Δ_t. More complex reparametrizations could also be possible [56].

To guide us in the choice of a suitable value for Δ_t, we perform a reasonably exhaustive search, using eight independent repetitions of state-vector simulation using $L = 4, 6, 8, 10$ and depths $d = 2, 4, 6, 8$. It is found that the value $\Delta_t \approx 0.80$ is the most frequent outcome of these optimization runs. Notably, a similar optimal value was found in Ref. [55] in the case of MaxCut instances on a random graph. These combined findings suggest that the quantum-annealing-inspired initialization is a general and robust procedure. This is further corroborated by the results for the disordered Hamiltonian, discussed below.

Hereafter, we first tackle the ferromagnetic problem, using the above prescription. The performance of the QAOA circuit with the annealing-inspired parameters is compared to the ones of the QAOA circuits with the parameters obtained after the fixed numbers of optimization iterations $n_{\text{iter}} = 20$ and $n_{\text{iter}} = 60$, starting from the same smart initialization. Notice that here the following definition of success probability P_{succ} is adopted: M measurements are performed on the prepared state (with $M = 16$ or $M = 32$), and the fraction of successful executions at a selected n_{iter} is recorded. This fraction differs from F_{succ}, which corresponds to the probability of observing the solution at least once during all optimization iterations, not only in the final state. The scaling of P_{succ} with problem size is shown in Figs. 7(a) and 7(b).

One observes that the QAOA *Ansatz* with the annealing-inspired linear initialization is already optimal, for both circuit depths $d = 2$ and $d = 4$. The optimization of the parameters does not yield better *Ansätz* to sample from. Two important observations are due: (1) the scaling exponent k is reduced compared to the random initialization case, and (2) k decreases with the circuit depth, as opposed to the case of random initialization displayed in Fig. 5(c). In Fig. 7(c), the scaling of the optimal number of calls n^*_{calls} is shown, following the procedure already discussed in Secs. IV A and IV C. This indicates that optimizations started from random parameters, performed with a finite budget of shots M, are not able to showcase the higher expressive power of deeper circuits. These numerical results suggest that, with a deep enough circuit, the exponent k can be reduced enough to reach practically useful performances for relevant problem sizes. This hypothesis is corroborated by the analysis reported at the end of this subsection.

As anticipated above, here we repeat the numerical experiment using ensembles of disordered Ising chains. It turns out that the precomputed value of $\Delta_t \approx 0.80$ is appropriate, in most instances, also in this setting. Importantly, in Fig. 8 we show that also in this case the smartly initialized *Ansatz* features almost converged parameters; indeed, the success

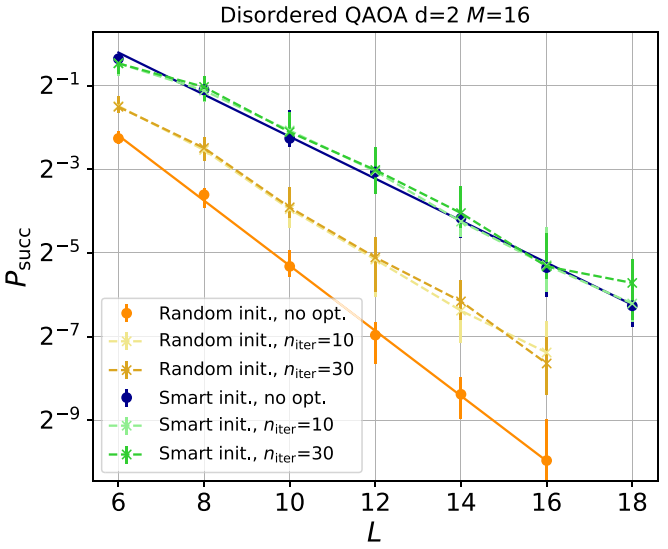

FIG. 8. Success probability P_{succ} as a function of the number of spins L, starting from random and from annealing-inspired initializations for the disordered Hamiltonian. We compare the success probability P_{succ} obtained via QAOA before (continuous curves) and after (dashed curves) the classical optimization performed for n_{iter} steps.

probability P_{succ} does not improve when the circuit is further optimized up to $n_{iter} = 30$ steps. Notice that P_{succ}, i.e., the probability to sample at least once the global minimum at the n-th iteration, is equal by definition to F_{succ} when $n_{iter} = 0$. The random-initialized circuit instead benefits from the optimization run, although it never reaches the success probability of the linearly initialized *Ansatz*. This result clearly shows that it is much better to use a clever parameters initialization without optimization, instead of randomly initializing the parameters $\boldsymbol{\theta}$ and performing the optimization.

Finally, we try to numerically demonstrate that, with the annealing-inspired initialization, sufficiently deep QAOA circuits can reach appealing performances, even without performing classical parameter optimizations. To this end, we determine the success probability F_{succ} as a function of the problem size L, for several circuit depths d at fixed shot numbers M (see Fig. 9). It is found that the performance systematically and rapidly increases with d, reaching $F_{succ} \simeq 1$ even for the largest considered size L, for sufficiently deep circuits. This evidence matches the intuition that QAOA reduces to quantum annealing when d is increased and the circuit parameters follow the pattern in Eq. (16) (although different schedules are possible) [30]. Notice that here, the number of shots is $M < 2^L$ for the sizes considered.

V. DISCUSSION

We critically analyze two popular quantum algorithms for optimization, VQE and QAOA, addressing controllable and reproducible testbed models, i.e., the ferromagnetic and the disordered Ising chains. On one hand, our results indicate that, in the practical regime where the number of measurements M per optimization step is much smaller than the Hilbert-space dimension 2^L, basic optimization strategies fail to identify suitable circuit parameters. On the other hand, appealing performances are achieved by deep QAOA circuits when a smart parameters initialization is adopted, as further discussed below. To reach the above conclusions, we track the total number of measurements n^*_{calls} to reach a fixed target success probability F_{succ} in the presence of measurement shot noise, and we analyze its scaling with the problem size L. As expected, we find exponential scalings in the form $n^*_{calls} \propto 2^{kL}$, and we determine the exponents k considering different setups, including energy-based versus gradient-based classical optimizers in both VQE and QAOA, different circuit depths d, as well as random and annealing-inspired parameter initializations in QAOA.

The first result of this article is that VQE shows a very poor scaling with problem size L. When an energy-based optimizer is adopted, the scaling is not better than direct enumeration of the whole computational space, which corresponds to $k = 1$. Introducing additional noise due to simulated hardware

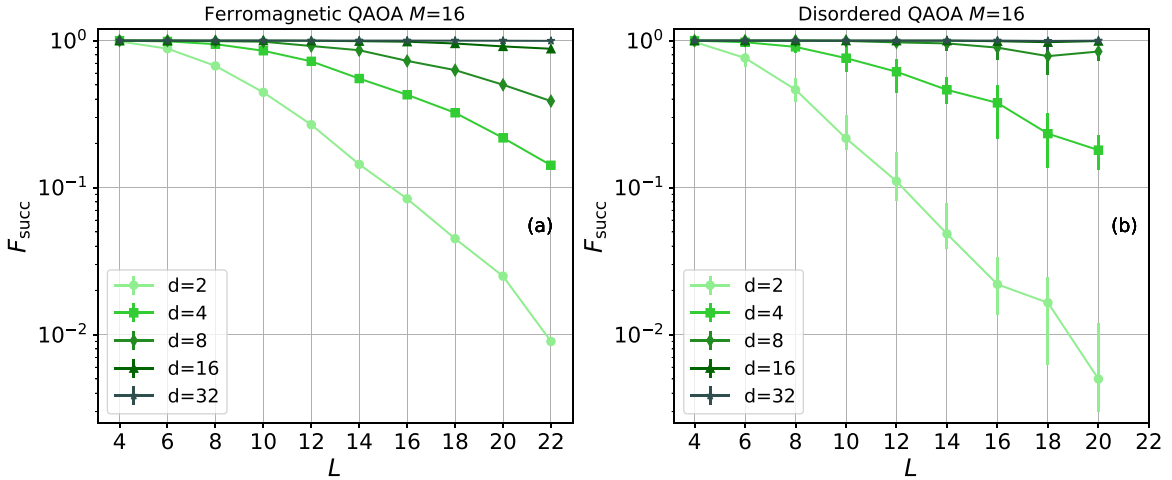

FIG. 9. Success probability F_{succ} as a function of the number of spins L without classical optimization. Using the smart linear initialization, F_{succ} grows when the circuit depth is increased, both for the (a) ferromagnetic model and (b) the disordered problems. In the latter, we consider 30 instances of disorder and the error bars indicate the 25th and the 75th percentiles.

errors does not significantly affect the scaling compared to the error-free case. Notice that our results are not in contrast with existing literature on the use of VQE for classical cost functions [32–35,39], since these studies report results in the regimes where $M \approx 2^L$ or $n_{\text{calls}} > 2^L$. We also find that a gradient-based optimization, which we implement in VQE via the parameters shift rule, is useful, leading up to a quadratic speedup compared to the energy-based optimizer COBYLA.

Then we consider QAOA: in contrast to most of the literature, we keep in consideration that the cost function needs to be stochastically evaluated, and in realistic conditions one can afford only $M \ll 2^L$ samples.

We first adopt a textbook version of QAOA, where we optimize the parameters from scratch, i.e., starting from random initial values. While, as expected, the total computation complexity is exponential, the exponent k is sizeably reduced compared to the full state enumeration. Notice that the performance degradation with the system size at fixed quantum resources n_{calls} is not due here to hardware noise [21], but to the intrinsic quantum measurement shot noise, a crucial ingredient which is often overlooked and usually leads to overoptimistic expectations for quantum algorithms [45,57]. Notice also that our numerical findings are compatible with Ref. [58], which discusses the query complexity of variational algorithms but only in the vicinity of the global minimum.

With the energy-based optimizer COBYLA, the QAOA scaling exponents turn out to be $k \simeq 0.4$ for the ferromagnetic problem, and in the range $0.5 < k < 0.8$ for the disordered models; the circuit depth does not significantly affect the scaling. As opposed to the VQE case, adopting a gradient-based optimization does not sizeably change k. Furthermore, a third optimizer, the SPSA algorithm, provides compatible results. These scaling exponents can be used to estimate the hypothetical runtimes required to execute the QAOA algorithm on physical quantum devices for realistic problem sizes. Assuming the best-observed scenario of the ferromagnetic case, $n_{\text{calls}} = 1 \times 2^{0.31L}$, some consequential bounds can be provided. For example, considering the circuit depth $d = 2$, gate execution time $t_{\text{gate}} = 10$ ns for the NISQ era (best case scenario here), one obtains runtimes of about tens of seconds for a hypothetical problem size $L = 100$, and a time much beyond the age of universe already for $L = 500$. These quotes need to be contrasted with tens of milliseconds of total CPU time of simulated annealing [59], or minutes for exact algorithms [6] for $L = 500$. To achieve a runtime of order 10 ms (resp. minutes), for $L = 500$, QAOA should achieve a scaling exponent of about $k = 0.04$ (resp. 0.07). We conclude that even the best-case scenario observed for the ferromagnetic model is insufficient to provide practical advantage relatively to classical methods or at least feasible absolute times.

Our numerical experiments are consistent with a very recent hardware assessment of QAOA versus quantum annealing, which shows that a $d = 2$ QAOA circuit, while better than random sampling, delivers worse performance than annealing [22]. However, it should be pointed out that, in that large-scale experiment, the performance metric cannot be defined in terms of success probability, since QAOA never provides the exact solution, nor approximate solutions qualitatively comparable with those of simulated or quantum annealing. This is again consistent with our picture. Moreover, our findings are not in contrast with other previous numerical [24] or experimental QAOA [20] studies, which are either presented in the $n_{\text{calls}} > 2^L$ regime or use a bootstrapping method to initialize the parameters.

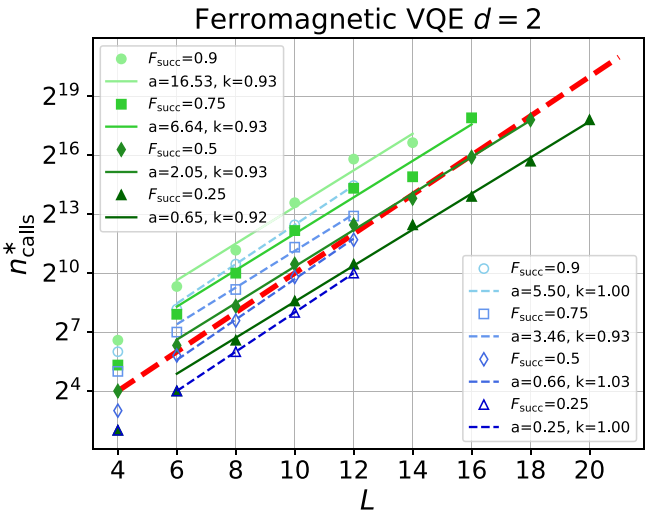

FIG. 10. Minimal number of function calls n^*_{calls} as a function of the problem size L, for different F_{succ}. VQE circuits with $d = 2$ blocks are considered, both with (empty symbols) and without (full symbols) simulated hardware errors, addressing ferromagnetic chains. The thick dashed (red) line represents the scaling $n^*_{\text{calls}} \sim 2^L$ corresponding to the exact enumeration. Thin continuous and dashed lines represent fitting functions of the form $n^*_{\text{calls}} = a\, 2^{kL}$, and the fitting parameters a and k, obtained considering the large L data, are given in the legend.

To recover an effective algorithm, it is crucial to use a smart initialization of the QAOA parameters. In fact, for the simple testbed models we consider, the parameter values given by the annealing-inspired schedule turn out to be very close to the optimal values, such that QAOA provides excellent success probabilities without the need for further parameter optimization. Interestingly, the same linear schedule proposed in Ref. [60] for MaxCut problems, based on noise-free simulations, turns out to be suitable also for our ferromagnetic and disordered Ising chains in the presence of measurement shot noise. While one cannot associate a specific scaling exponent to the smartly initialized QAOA algorithm, as, fortunately, in this case the scaling does improve with the circuit depth, it is quite plausible that sufficiently deep circuits can reach feasible computational times for practically relevant problem sizes.

It is worth remembering that the proposal to use a smart initialization of the QAOA parameter is not new [27,30,55,56]. However, our findings show that this choice should not only be considered as a good practice to marginally enhance the algorithm efficiency, but it is the only route to make the algorithm practical in the presence of shot noise. Indeed, if one performs (noise-free) state-vector emulations of the optimization run, good parameters can be recovered anyway, irrespective of the initialization [27,30,55].

Overall, we suggest that future implementation of QAOA should at least rethink the use of the outer optimization loop, focusing in particular on smart parameter initializations. While in this manuscript we adopt an annealing-inspired

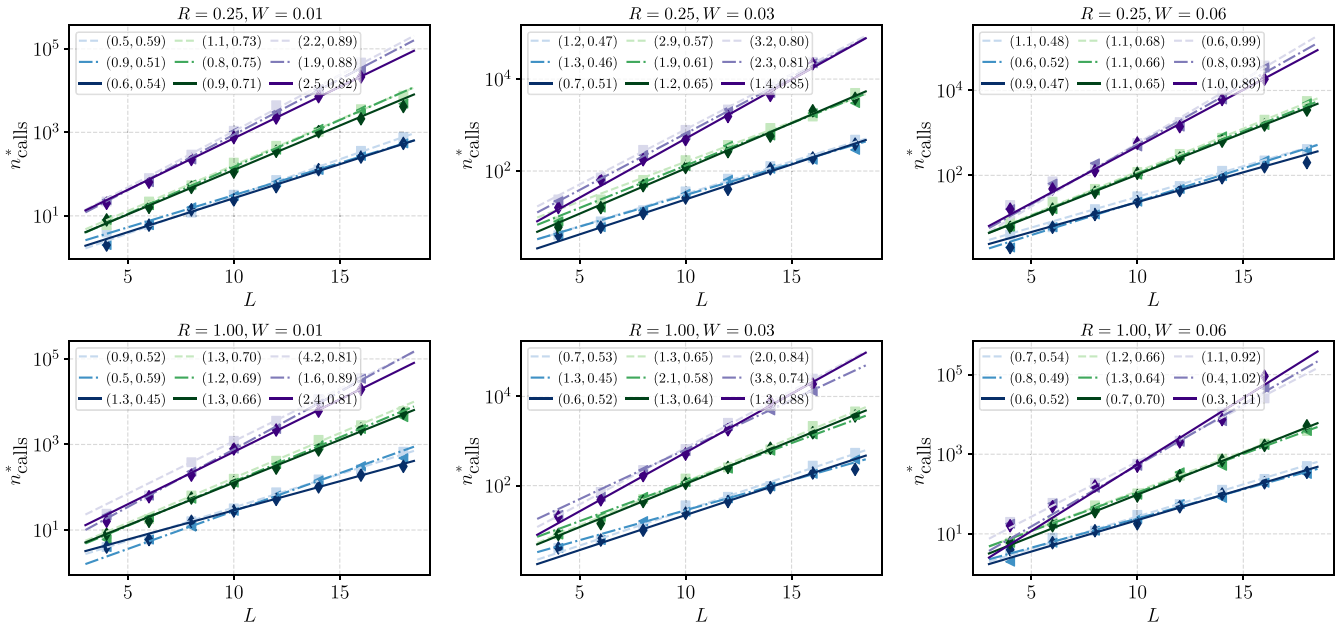

FIG. 11. Scaling of n^*_{calls} for various depths d, CVaR sample fractions R, and SPSA proposal lengths W. Three curves (from transparent to solid) show n^*_{calls} for $F_{\text{succ}} = \{0.25,\ 0.5,\ 0.75\}$, respectively. The colors indicate different depths $d = 2$ (blue, lowest data), $d = 3$ (green, intermediate data), and $d = 4$ (purple, highest data). The numbers (a, k) in the brackets in the legends give the optimal values of the exponential fit $n^*_{\text{calls}} = a\, 2^{kL}$ obtained by fitting the data in the regime $L \geqslant 8$.

initialization, more flexible solutions, suitable for shallower circuits, are possible. In general, the angle array can be reparametrized as $\boldsymbol{\theta} \to \boldsymbol{\theta}(\boldsymbol{\alpha})$, using a smaller number of optimizable parameters $\boldsymbol{\alpha}$. This might allow performing fewer optimization steps, similar to the Fourier reparametrization of Ref. [27]. The research concerning preoptimization is very active. For instance, the QAOA smart initialization has been studied for the Sherrington-Kirkpatrick model [61] and the MaxCut problem [29,62,63]. Moreover, the experimental results achieved in the studies cited above have been analytically confirmed in Ref. [64]. Note that the issue of shot noise is well known in VQE for genuine many-body quantum Hamiltonians, for example, in chemistry [57]. However, the fact that it also manifests so severely in the case of a classical cost function, which can be measured in a single basis, is important.

We expect our findings to apply in general to variational quantum algorithms strongly relying on a classical optimization loop, but not to other alternatives for quantum-enhanced optimization on digital hardware, including quantum-powered sampling [65–67], branch-and-bound algorithm [68], and quantum walks [69], to name a few proposals. On a methodological note, these results demonstrate the importance of simple and controllable models to analyze the scaling properties of quantum algorithms in realistic settings.

All data discussed in this article are freely available from Ref. [70].

ACKNOWLEDGMENTS

We acknowledge useful discussion with G. Carleo regarding the relation between QAOA and quantum annealing. We thank A. Miessen and C. Cozza for helping us set up the QISKIT simulation on the cluster. Computation for the work described in this paper was supported by the Science Cluster at the University of Zurich. G.S. acknowledges the hospitality of the Institute for Computational Science at the University of Zurich and useful discussions with E. Costa. G.M. acknowledges financial support from the Swiss National Science Foundation (Grant No. PCEFP2_203455). N.A. is funded by the Swiss National Science Foundation, Grant No. PP00P2_-176877. S.P. acknowledges PRACE for awarding access to the Fenix Infrastructure resources at Cineca, which are partially funded by the European Union's Horizon 2020 research and innovation program through the ICEI project under Grant Agreement No. 800858. This work was also supported by the PNRR MUR Project No. PE0000023-NQSTI and by the Italian Ministry of University and Research under the PRIN2022 project "Hybrid algorithms for quantum simulators," Project No. 2022H77XB7.

APPENDIX A: VQE WITH HARDWARE ERRORS

In this Appendix, we inspect the possible role of hardware errors. For this, a custom model of hardware noise is introduced, using the open-source QISKIT API [46]. A realistic model is obtained, e.g., considering the thermal relaxation due to the qubit environment. Each qubit is then parametrized by a thermal relaxation time constant $T_1 = 50\,\mu$s and a dephasing time constant $T_2 = 70\,\mu$s. The performance comparison against the error-free VQE circuits is shown in Fig. 10. Ferromagnetic chains are considered using the CVaR estimator. It turns out that the scaling of n^*_{calls} with L is not significantly affected by this simulated hardware noise.

APPENDIX B: QAOA WITH SPSA

In this Appendix, we analyze the computational scaling using an energy-based optimizer alternative to COBYLA, namely, the SPSA algorithm [52]. This is used to optimize QAOA circuits of depth in the range $2 \leqslant d \leqslant 4$. The testbed we consider here is the ferromagnetic Ising chain, corresponding to set $J_{j,j+1} = +1$ in Eq. (1).

In the SPSA algorithm, at each optimization step, a random uniformly distributed n_{par} parameters shift with a constrained length is applied: $\boldsymbol{\theta} \to \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$, with $\|\Delta\boldsymbol{\theta}\| \leqslant W$. We consider three values of this maximum norm, namely, $W = \{0.01, 0.03, 0.06\}$. The shift vector is generated as a random vector on a n_{par}-dimensional unit sphere, normalized to length W. We accept the new parameters if the cost function decreases. As the cost function, we use the CVaR with either 25% or 100% of the best-energy samples.

To obtain n^*_{calls}, we use a procedure similar to the one used in Sec. III. We consider optimization with M samples generated at each SPSA step, and optimize until F_{succ} reaches the target value. We then compute $n^*_{\text{calls}} = \min(M \times n_{\text{iter}})$. The initial parameters are uniform random values in the range $\theta_n \in (-1.0, 1.0)$, and the results for n^*_{calls} are obtained by averaging over 1000 simulations with random starting points. The results are presented in Fig. 11.

Notably, we observe that the scaling sizeably worsens as d increases from 2 to 4. This could be attributed to a more complex optimization landscape which requires a higher M or n_{calls} to approach the global minimum. At the same time, the *Ansatz* with $d = 2$ reaches the $k \simeq 0.5$ scaling, therefore it features a quadratic speedup compared with the scaling of the COBYLA-driven VQE optimization, as also found with the QAOA algorithm, driven either by COBYLA or by the gradient-based optimizer.

[1] A. Abbas, A. Ambainis, B. Augustino, A. Bärtschi, H. Buhrman, C. Coffrin, G. Cortiana, V. Dunjko, D. J. Egger, B. G. Elmegreen *et al.*, Quantum optimization: Potential, challenges, and the path forward, arXiv:2312.02279.

[2] R. P. Feynman, Simulating physics with computers, Int. J. Theor. Phys. **21**, 467 (1982).

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

[4] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Optimization by simulated annealing, Science **220**, 671 (1983).

[5] M. W. Johnson, M. H. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk *et al.*, Quantum annealing with manufactured spins, Nature (London) **473**, 194 (2011).

[6] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, Evidence for quantum annealing with more than one hundred qubits, Nat. Phys. **10**, 218 (2014).

[7] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, Defining and detecting quantum speedup, Science **345**, 420 (2014).

[8] V. S. Denchev, S. Boixo, S. V. Isakov, N. Ding, R. Babbush, V. Smelyanskiy, J. Martinis, and H. Neven, What is the computational value of finite-range tunneling? Phys. Rev. X **6**, 031015 (2016).

[9] T. Albash and D. A. Lidar, Demonstration of a scaling advantage for a quantum annealer over simulated annealing, Phys. Rev. X **8**, 031016 (2018).

[10] S. V. Isakov, G. Mazzola, V. N. Smelyanskiy, Z. Jiang, S. Boixo, H. Neven, and M. Troyer, Understanding quantum tunneling through quantum Monte Carlo simulations, Phys. Rev. Lett. **117**, 180402 (2016).

[11] G. Mazzola, V. N. Smelyanskiy, and M. Troyer, Quantum Monte Carlo tunneling from quantum chemistry to quantum annealing, Phys. Rev. B **96**, 134305 (2017).

[12] E. M. Inack, G. Giudici, T. Parolini, G. Santoro, and S. Pilati, Understanding quantum tunneling using diffusion Monte Carlo simulations, Phys. Rev. A **97**, 032307 (2018).

[13] T. Parolini, E. M. Inack, G. Giudici, and S. Pilati, Tunneling in projective quantum Monte Carlo simulations with guiding wave functions, Phys. Rev. B **100**, 214303 (2019).

[14] I. Ozfidan, C. Deng, A. Y. Smirnov, T. Lanting, R. Harris, L. Swenson, J. Whittaker, F. Altomare, M. Babcock, C. Baron *et al.*, Demonstration of a nonstoquastic Hamiltonian in coupled superconducting flux qubits, Phys. Rev. Applied **13**, 034037 (2020).

[15] A. D. King, S. Suzuki, J. Raymond, A. Zucca, T. Lanting, F. Altomare, A. J. Berkley, S. Ejtemaee, E. Hoskinson, S. Huang *et al.*, Coherent quantum annealing in a programmable 2,000 qubit Ising chain, Nat. Phys. **18**, 1324 (2022).

[16] 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).

[17] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[18] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[19] F. Barahona, On the computational complexity of Ising spin glass models, J. Phys. A: Math. Gen. **15**, 3241 (1982).

[20] G. Pagano, A. Bapat, P. Becker, K. S. Collins, A. De, P. W. Hess, H. B. Kaplan, A. Kyprianidis, W. L. Tan, C. Baldwin, L. T. Brady, A. Deshpande, F. Liu, S. Jordan, A. V. Gorshkov, and C. Monroe, Quantum approximate optimization of the long-range Ising model with a trapped-ion quantum simulator, Proc. Natl. Acad. Sci. **117**, 25396 (2020).

[21] M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya, J. Atalaya, J. C. Bardin, R. Barends, S. Boixo *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, Nat. Phys. **17**, 332 (2021).

[22] E. Pelofske, A. Bärtschi, and S. Eidenbenz, Quantum annealing vs. QAOA: 127 qubit higher-order Ising problems on NISQ computers, in *High Performance Computing*, edited by A. Bhatele, J. Hammond, M. Baboulin, and C. Kruse (Springer Nature Switzerland, Cham, 2023), pp. 240–258.

[23] G. Nannicini, Performance of hybrid quantum-classical variational heuristics for combinatorial optimization, Phys. Rev. E **99**, 013304 (2019).
[24] G. G. Guerreschi and M. Smelyanskiy, Practical optimization for hybrid quantum-classical algorithms, arXiv:1701.01450.
[25] S. Hadfield, Z. Wang, B. O'Gorman, E. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019).
[26] G. G. Guerreschi and A. Y. Matsuura, QAOA for Max-Cut requires hundreds of qubits for quantum speed-up, Sci. Rep. **9**, 6903 (2019).
[27] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices, Phys. Rev. X **10**, 021067 (2020).
[28] C. Moussa, H. Calandra, and V. Dunjko, To quantum or not to quantum: Towards algorithm selection in near-term quantum optimization, Quantum Sci. Technol. **5**, 044009 (2020).
[29] S. Boulebnane and A. Montanaro, Predicting parameters for the quantum approximate optimization algorithm for MAX-CUT from the infinite-size limit, arXiv:2110.10685.
[30] M. Willsch, D. Willsch, F. Jin, H. De Raedt, and K. Michielsen, Benchmarking the quantum approximate optimization algorithm, Quantum Inf. Process. **19**, 197 (2020).
[31] L. Binkowski, G. Koßmann, T. Ziegler, and R. Schwonnek, Elementary proof of QAOA convergence, arXiv:2302.04968 (2023).
[32] P. K. Barkoutsos, G. Nannicini, A. Robert, I. Tavernelli, and S. Woerner, Improving variational quantum optimization using CVaR, Quantum **4**, 256.
[33] P. Díez-Valle, D. Porras, and J. J. García-Ripoll, Quantum variational optimization: The role of entanglement and problem hardness, Phys. Rev. A **104**, 062426 (2021).
[34] D. Amaro, C. Modica, M. Rosenkranz, M. Fiorentini, M. Benedetti, and M. Lubasch, Filtering variational quantum algorithms for combinatorial optimization, Quantum Sci. Technol. **7**, 015021 (2022).
[35] C. Zoufal, R. V. Mishmash, N. Sharma, N. Kumar, A. Sheshadri, A. Deshmukh, N. Ibrahim, J. Gacon, and S. Woerner, Variational quantum algorithm for unconstrained black box binary optimization: Application to feature selection, Quantum **7**, 909 (2023).
[36] I. Kolotouros and P. Wallden, Evolving objective function for improved variational quantum optimization, Phys. Rev. Res. **4**, 023225 (2022).
[37] X. Liu, A. Angone, R. Shaydulin, I. Safro, Y. Alexeev, and L. Cincio, Layer VQE: A variational approach for combinatorial optimization on noisy quantum computers, IEEE Trans. Autom. Sci. Eng. **3**, 1 (2022).
[38] S. Chakrabarti, R. Krishnakumar, G. Mazzola, N. Stamatopoulos, S. Woerner, and W. J. Zeng, A threshold for quantum advantage in derivative pricing, Quantum **5**, 463 (2021).
[39] A. Robert, P. K. Barkoutsos, S. Woerner, and I. Tavernelli, Resource-efficient quantum algorithm for protein folding, npj Quantum Inf. **7**, 38 (2021).
[40] S. Harwood, C. Gambella, D. Trenev, A. Simonetto, D. Bernal, and D. Greenberg, Formulating and solving routing problems on quantum computers, IEEE Trans. Autom. Sci. Eng. **2**, 1 (2021).
[41] X. Wang, Z. Song, and Y. Wang, Variational quantum singular value decomposition, Quantum **5**, 483 (2021).
[42] L. Braine, D. J. Egger, J. Glick, and S. Woerner, Quantum algorithms for mixed binary optimization applied to transaction settlement, IEEE Trans. Autom. Sci. Eng. **2**, 1 (2021).
[43] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).
[44] C. Gidney and A. G. Fowler, Efficient magic state factories with a catalyzed $|CCZ\rangle$ to $2|T\rangle$ transformation, Quantum **3**, 135 (2019).
[45] G. Mazzola and G. Carleo, Exponential challenges in unbiasing quantum Monte Carlo algorithms with quantum computers, arXiv:2205.09203.
[46] M. Anis Sajid *et al.*, Qiskit: An open-source framework for quantum computing (2021).
[47] J. Weidenfeller, L. C. Valor, J. Gacon, C. Tornow, L. Bello, S. Woerner, and D. J. Egger, Scaling of the quantum approximate optimization algorithm on superconducting qubit based hardware, Quantum **6**, 870 (2022).
[48] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).
[49] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).
[50] S. Wei, H. Li, and G. Long, A full quantum eigensolver for quantum chemistry simulations, Res. **2020**, 2020/1486935 (2020).
[51] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).
[52] J. Spall, Implementation of the simultaneous perturbation algorithm for stochastic optimization, IEEE Trans. Aerosp. Electron. Syst. **34**, 817 (1998).
[53] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, General parameter-shift rules for quantum gradients, Quantum **6**, 677 (2022).
[54] H.-Y. Liu, Z.-Y. Chen, T.-P. Sun, C. Xue, Y.-C. Wu, and G.-P. Guo, Can variational quantum algorithms demonstrate quantum advantages? Time really matters, arXiv:2307.04089.
[55] S. H. Sack and M. Serbyn, Quantum annealing initialization of the quantum approximate optimization algorithm, Quantum **5**, 491 (2021).
[56] 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).
[57] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).
[58] A. W. Harrow and J. C. Napp, Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms, Phys. Rev. Lett. **126**, 140502 (2021).
[59] S. Isakov, I. Zintchenko, T. Rønnow, and M. Troyer, Optimised simulated annealing for Ising spin glasses, Comput. Phys. Commun. **192**, 265 (2015).

[60] F. G. S. L. Brandao, M. Broughton, E. Farhi, S. Gutmann, and H. Neven, For fixed control parameters the quantum approximate optimization algorithm's objective function value concentrates for typical instances, arXiv:1812.04170.

[61] E. Farhi, J. Goldstone, S. Gutmann, and L. Zhou, The quantum approximate optimization algorithm and the Sherrington-Kirkpatrick model at infinite size, Quantum **6**, 759 (2022).

[62] A. Galda, X. Liu, D. Lykov, Y. Alexeev, and I. Safro, Transferability of optimal QAOA parameters between random graphs, in *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE, Piscataway, NJ, 2021), pp. 171–180.

[63] D. J. Egger, J. Mareček, and S. Woerner, Warm-starting quantum optimization, Quantum **5**, 479 (2021).

[64] V. Akshay, D. Rabinovich, E. Campos, and J. Biamonte, Parameter concentrations in quantum approximate optimization, Phys. Rev. A **104**, L010401 (2021).

[65] G. Mazzola, Sampling, rates, and reaction currents through reverse stochastic quantization on quantum computers, Phys. Rev. A **104**, 022431 (2021).

[66] D. Layden, G. Mazzola, R. V. Mishmash, M. Motta, P. Wocjan, J.-S. Kim, and S. Sheldon, Quantum-enhanced Markov chain Monte Carlo, Nature (London) **619**, 282 (2023).

[67] G. Mazzola, Quantum computing for chemistry and physics applications from a Monte Carlo perspective, J. Chem. Phys. **160**, 010901 (2024).

[68] A. Montanaro, Quantum speedup of branch-and-bound algorithms, Phys. Rev. Res. **2**, 013056 (2020).

[69] A. Callison, N. Chancellor, F. Mintert, and V. Kendon, Finding spin glass ground states using quantum walks, New J. Phys. **21**, 123022 (2019).

[70] G. Scriva, N. Astrakhantsev, S. Pilati, and G. Mazzola, Data for "Challenges of variational quantum optimization with measurement shot noise," doi:10.5281/zenodo.8223528 (2023).