Quantum approximate optimization algorithm with adaptive bias fields

Yunlong Yu,[1,2] Chenfeng Cao,[3] Carter Dewey,[4] Xiang-Bin Wang,[1] Nic Shannon,[5] and Robert Joynt[4,2]

[1]*State Key Laboratory of Low Dimensional Quantum Physics, Department of Physics, Tsinghua University, Beijing 100084, China*

[2]*Kavli Institute for Theoretical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China*

[3]*Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China*

[4]*Department of Physics, University of Wisconsin–Madison, 1150 University Avenue, Madison, Wisconsin 53706, USA*

[5]*Theory of Quantum Matter Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Okinawa 904-0412, Japan*

(Received 24 May 2021; revised 25 March 2022; accepted 12 May 2022; published 27 June 2022)

The quantum approximate optimization algorithm (QAOA) transforms a simple many-qubit wave function into one that encodes a solution to a difficult classical optimization problem. It does this by optimizing the schedule according to which two unitary operators are alternately applied to the qubits. In this paper, the QAOA is modified by updating the operators themselves to include local fields, using information from the measured wave function at the end of one iteration step to improve the operators at later steps. It is shown by numerical simulation on MaxCut problems that, for a fixed accuracy, this procedure decreases the runtime of QAOA very substantially. This improvement appears to increase with the problem size. Our method requires essentially the same number of quantum gates per optimization step as the standard QAOA, and no additional measurements. This modified algorithm enhances the prospects for quantum advantage for certain optimization problems.

DOI: 10.1103/PhysRevResearch.4.023249

I. INTRODUCTION

We are in the era of noisy intermediate-scale quantum (NISQ) devices [1]. This motivates the development of variational quantum algorithms (VQAs) that use a sequence of relatively short quantum circuits with parameters that are iteratively updated by a classical optimizer [2–4]. VQAs have been designed for a wide range of problems, such as ground state and excited state preparation [5–9], quantum state diagonalization [10,11], quantum data compression [12–14], quantum fidelity estimation [15,16], and quantum compiling [17].

The quantum approximate optimization algorithm (QAOA) is the leading example of a VQA for combinatorial optimization [18]. The repeated quantum evolution depends on classical parameters that are iteratively updated. The final result is a calculated value for the cost function and a corresponding quantum state that encodes an approximate solution to a classical optimization problem. The QAOA is considered to be a good candidate for an algorithm that will be superior to classical algorithms reasonably soon [19], so many studies have focused on the experimental demonstrations for the QAOA in different physical systems [20–26]. However, it is generally thought that the standard QAOA will not be competitive with established classical methods until a time when quantum machines are considerably larger than they are today [27,28]. Thus there is intense activity to improve the QAOA [29–36], which would bring this time closer. That is also the goal of the present work.

The QAOA starts the quantum computer in the ground state of the mixing Hamiltonian H_M^s and then alternately applies the unitary operators $\exp(-i\gamma_k H_C)$ and $\exp(-i\beta_k H_M^s)$, where H_C is the problem Hamiltonian whose ground state is sought [18]. At level p, $\{\gamma_k\}_{k=1}^p$ and $\{\beta_k\}_{k=1}^p$ are two sets of parameters that fix the schedule of the evolution. At each iteration, $\{\gamma_k\}_{k=1}^p$ and $\{\beta_k\}_{k=1}^p$ are improved by measuring H_C. (Henceforth we drop the subscripts and superscripts on $\{\gamma_k\}_{k=1}^p$ and the other parameter sets.)

Simulations on classical computers have shown some impressive results for the QAOA as applied to MaxCut [35,37]. The authors of Ref. [35] produced an efficient iterative scheme that runs in time $O(\text{poly } p)$ and that approached the known solutions with high accuracy. They demonstrated the superiority of QAOA over standard quantum annealing—the classical optimization effectively isolates the small gap events that plague annealers and the Quantum Adiabatic Algorithm and substantially neutralizes them, though it should be noted that modifications of quantum annealing can do this for certain special problems [38,39].

Here we introduce a method that can greatly accelerate convergence of the QAOA through the use of adaptive bias fields in the mixing Hamiltonian (ab-QAOA). Varying the ansatz operator in the VQA [40,41] or QAOA [36,42] has been proposed before, but the ab-QAOA approach has two critical differences from earlier protocols. The first is that previous modifications of the QAOA do not use all the information available at the end of a step. One can use the energy measurement in more than one way to guide the system toward its ground state. The second is that local fields are introduced in

H_M^{ab} as was done previously in quantum annealing [43] and the starting state is reinitialized accordingly. Some methods such as FALQON [44,45] also use measurements to update the operators, but the operators applied are completely different.

Mean field theory is the usual starting point for the investigation of ordered spin systems. It can also be very useful for certain Ising spin glass systems, which is the case of interest here. The prime example is the Parisi solution of the Sherrington-Kirkpatrick model [46]. Hence it is natural to include elements of mean field theory in any search for a ground state. Thus the overall philosophy of our approach is to make a marriage between mean field theory and an iterative variational procedure. There are four reasons to suppose this will improve the QAOA.

(i) The optimization is guided not only by the energy but also by the local magnetization, so additional information available from the measurements is used.

(ii) Mean field theory is often the best starting point for a variational calculation on a system with many degrees of freedom.

(iii) For any algorithm based fundamentally on the adiabatic theorem, the shorter the distance in Hilbert space from the starting wave function final correct ground state, the better the chances of success [43]. Our procedure includes a modification of the wave function at each stage of the iteration.

(iv) For any problem whose solution is one of the computational basis vectors (an Ising problem, in condensed matter theory language), a local field term in the z direction will steer the solution in a good candidate direction, due to the fact that the solution lies in the set of ground states of some local-field Hamiltonian.

Bias fields have previously been introduced in a quantum adiabatic algorithm to improve accuracy (defined below) [43], but in that reference the procedure was not adaptive. This leads us to call our method the "adaptive bias QAOA" (ab-QAOA). The use of adaptive bias fields improves both the accuracy of the solution and its fidelity, i.e., the overlap between the computed final state and the actual target state. There have been some adaptive QAOA methods, such as the operator pool method [36], in that the mixing Hamiltonian is updated, but reinitialization has not been employed in the past.

This paper is organized as follows. In Sec. II, we give a detailed description of the MaxCut problem, the QAOA, and the ab-QAOA. In Sec. III, the relative performances of the QAOA and the ab-QAOA are computed and analyzed. We also investigate in detail the effects of the bias fields in the ab-QAOA. The conclusion of the paper is given in Sec. IV.

II. ALGORITHM DETAILS

A. MaxCut problem

The performance of a heuristic algorithm must be judged against competitors. In what follows, we compare the ab-QAOA against the standard QAOA (henceforth referred to simply as QAOA) described above. The QAOA has already been compared to classical algorithms [37], so this way of proceeding indirectly benchmarks the QAOA against classical competitors as well. Following Ref. [18], we define the

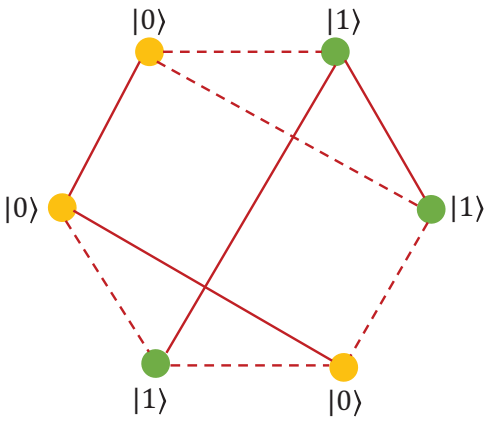

FIG. 1. MaxCut problem on an unweighted 3-regular 6-vertex graph. Different colors give the different states $|0\rangle$ and $|1\rangle$, and they represent the two different subsets V_1 and V_2 of the vertex set. The problem is to find the division of the vertices that maximizes the number of edges connecting the two subsets. The dashed edges in the figure represent the cut in this case.

accuracy as

$$r = \frac{E_{\text{opt}}(\psi_f)}{E_{\text{max}}(\psi_{\text{max}})}, \tag{1}$$

where $E_{\text{opt}}(\psi_f)$ is the expectation value of the problem Hamiltonian in the state ψ_f produced by the algorithm, and $E_{\text{max}}(\psi_{\text{max}})$ is the value in the optimum state ψ_{max}.

The problem we use for benchmarking ab-QAOA is MaxCut, a canonical problem in graph theory [47]. Let an undirected graph be denoted by $G(V, E)$, where V is the n-vertex set and E is the edge set. The edges may or may not be assigned weights. If they are, then the weights are chosen uniformly at random from the interval [0,1]. In the unweighted version, we wish to partition V into two subsets V_1, V_2 in such a way as to make the number of edges connecting V_1 and V_2 as large as possible. In the weighted version, the total weight of the partition is maximized.

We convert MaxCut to an n-vertex Ising model as follows. Define a Pauli matrix Z_j to act on the jth vertex and use the eigenstates $|0\rangle, |1\rangle$ of the Z_j to represent V_1 and V_2. Thus, in operator language, the MaxCut problem Hamiltonian for n qubits is $H = E_0 - H_C$, where

$$H_C = \sum_{\langle v_1 v_2\rangle \in E} \frac{\omega_{v_1 v_2}}{2} Z_{v_1} Z_{v_2}. \tag{2}$$

The constant $E_0 = \sum \omega_{v_1 v_2}/2$ plays no role in the partition of the graph, but it enters the calculations of the accuracy r as defined above. The ground state has an obvious $\mathbb{Z}_2$ symmetry. The ground state of H_C in Eq. (2) encodes the solutions to the original MaxCut problems. We consider weighted 3-regular graphs with $\omega_{v_1 v_2}$ chosen uniformly at random in [0,1] (w3r graphs), and unweighted 3-regular graphs with $\omega_{v_1 v_2} = 1$ (u3r graphs). An example of an unweighted graph is shown in Fig. 1.

The choice of problems is the same as in Ref. [35]. Classically, finding a solution where $r > 16/17 \approx 0.9412$ on all

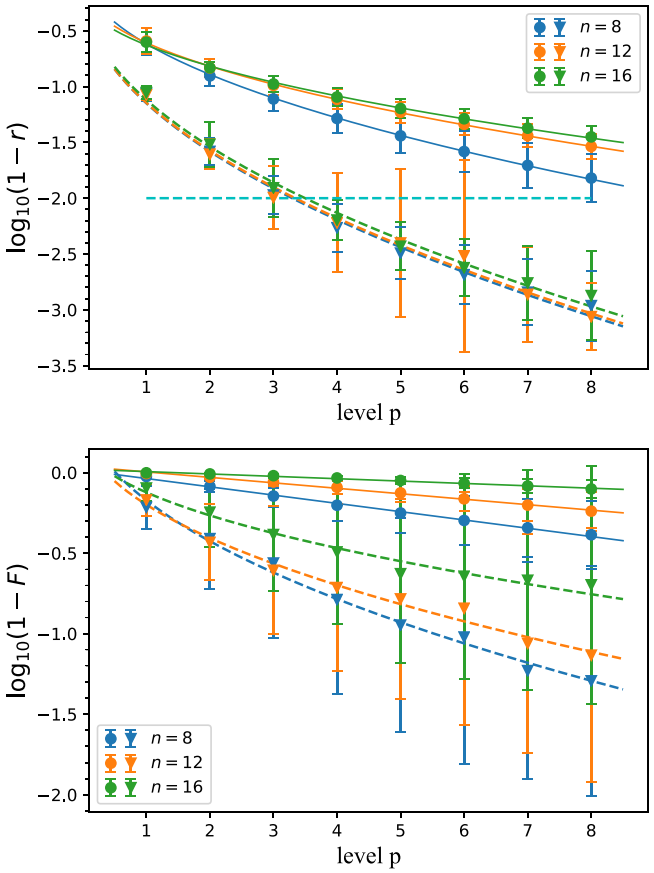

FIG. 2. Comparison of the accuracy (top panel) and infidelity (bottom panel) of the QAOA (solid lines) to the accuracy and infidelity of ab-QAOA (dashed lines) for $n = 8, 12, 16$ for w3r graphs. Each point is an average over 40 randomly chosen graphs. In (a) the accuracy is plotted as a function of the level p. The horizontal dashed line represents $r^* = 0.99$. Even for moderate values of p, the accuracy of ab-QAOA is an order of magnitude better than that of QAOA. In (b) the infidelity in QAOA and ab-QAOA is plotted as a function of p. Again, the improvement is nearly an order of magnitude at moderate p. The fits are described in the text. The error bars are standard deviations.

graphs is NP-hard [48,49], but there is a polynomial time classical algorithm that provably finds answers with $r = 0.8785$ [50].

B. QAOA and ab-QAOA

The quantum part of the standard QAOA is the repeated computation of a quantity $|\psi_f^{\rm s}\rangle$ according to

$$|\psi_f^{\rm s}\rangle = \prod_{k=1}^{p} e^{-i\beta_k H_M^{\rm s}} e^{-i\gamma_k H_C} |\psi_0^{\rm s}\rangle, \tag{3}$$

with $H_M^{\rm s} = \sum_j X_j$, where X_j is the Pauli X matrix that acts on the jth qubit. $|\psi_0^{\rm s}\rangle$ is the ground state of $H_M^{\rm s}$. The operators with subscript k are on the left of those with $k-1$ in $\prod_{k=1}^{p}\cdots$. The classical part is the iterative optimization of $\{\gamma_k\}$ and $\{\beta_k\}$.

The ab-QAOA algorithm modifies the QAOA algorithm in the following ways:

$$|\psi_f^{\rm ab}\rangle = \prod_{k=1}^{p} e^{-i\beta_k H_M^{\rm ab}(\{h_j\})} e^{-i\gamma_k H_C} |\psi_0^{\rm ab}(\{h_j\})\rangle, \tag{4}$$

where the mixing Hamiltonian is $H_M^{\rm ab} = \sum_j (X_j - h_j Z_j)$ and the starting wave function $|\psi_0^{\rm ab}\rangle$ is the ground state of the former. There are n additional parameters $\{h_j\}$ that comprise the local fields and enter both the $H_M^{\rm ab}$ and $|\psi_0^{\rm ab}\rangle$. They are *not* optimized, but rather updated according to the prescription

$$h_j \to h_j - \ell\big(h_j - \langle\psi_f^{\rm ab}|Z_j|\psi_f^{\rm ab}\rangle\big). \tag{5}$$

ℓ is the learning rate, which we took to be $\ell = 1.1$, and $\langle\psi_f^{\rm ab}|Z_j|\psi_f^{\rm ab}\rangle$ can be obtained from the measurement of ZZ terms in H_C. This is one step of the learning process. Thus both $H_M^{\rm ab}$ and $|\psi_0^{\rm ab}\rangle$ are updated (learned) along with the usual QAOA schedule parameters $\{\gamma_k\}$ and $\{\beta_k\}$ (which are optimized in the usual way at each iteration). The details of ab-QAOA in level p can be found in the following.

Algorithm: ab-QAOA in level *p*

- Initialization
  1. Initialize two p-element sets $\{u_l\}$ and $\{v_l\}$ that are used to update $\{\gamma_k\}$ and $\{\beta_k\}$.
  2. Initialize the n-element local field set $\{h_j\}$.
  3. Set a learning rate ℓ, a global parameter defined in Step 6 in the optimization procedure.
- Optimization
  1. Set $\{\gamma_k\}$ and $\{\beta_k\}$ according to the discrete Fourier transforms of $\{u_l\}$ and $\{v_l\}$.
  2. Construct the mixing Hamiltonian with bias fields: $H_M^{\rm ab}(\{h_j\}) = \sum_{j=1}^{n}(X_j - h_j Z_j)$.
  3. Prepare $|\psi_0^{\rm ab}\rangle$, the product ground state of $H_M^{\rm ab}(\{h_j\})$.
  4. Compute the final state for this step: $|\psi_f^{\rm ab}\rangle = \prod_{k=1}^{p} e^{-i\beta_k H_M^{\rm ab}(\{h_j\})} e^{-i\gamma_k H_C} |\psi_0^{\rm ab}\rangle$.
  5. Using projective measurements, obtain the gradients of the energy: $\frac{\partial\langle\psi_f^{\rm ab}|H_C|\psi_f^{\rm ab}\rangle}{\partial\vec{u}}$ and $\frac{\partial\langle\psi_f^{\rm ab}|H_C|\psi_f^{\rm ab}\rangle}{\partial\vec{v}}$ and the quantity $\delta h_j = h_j - \langle\psi_f^{\rm ab}|Z_j|\psi_f^{\rm ab}\rangle$.
  6. Update $\{v_l\}$, $\{u_l\}$ using the Adam gradient-based stochastic optimization algorithm [51]. Update $\{h_j\}$ with learning rate ℓ according to $h_j \to h_j - \ell\delta h_j$. The update of $\{h_j\}$ feeds back into both the mixing Hamiltonian in Step 2 and the wave function in Step 3.
  7. Measure the expectation value of the energy/cost function $E(\{u_l\},\{v_l\},\{h_j\}) = \langle\psi_f^{\rm ab}|H_C|\psi_f^{\rm ab}\rangle$.
  8. Repeat steps 1–7 until convergence with a fixed tolerance. Output the final energy $E_f(\{u_l\},\{v_l\},\{h_j\})$, and a measurement of $|\psi_f^{\rm ab}\rangle$ in the computational basis. Allowing for the constant term, the optimized energy is $E_p^{\rm opt} = E_0 - E_f$.

In addition to the optimization of $\{\gamma_k\}$ and $\{\beta_k\}$, another important issue is the choice of the initial $\{\gamma_k\}$ and $\{\beta_k\}$ at the beginning of the optimization. For this, we adapt the Fourier

strategy [35], as described in Appendix A. The main idea is not to directly optimize $\{\gamma_k\}$ and $\{\beta_k\}$, but rather to optimize their Fourier components $\{u_l\}$ and $\{v_l\}$, given by

$$\gamma_k = \sum_{l=1}^{p} u_l \sin\left[\left(l-\frac{1}{2}\right)\left(k-\frac{1}{2}\right)\frac{\pi}{p}\right],$$
$$\beta_k = \sum_{l=1}^{p} v_l \cos\left[\left(l-\frac{1}{2}\right)\left(k-\frac{1}{2}\right)\frac{\pi}{p}\right]. \quad (6)$$

Then the starting point in level p can be constructed from the optimized point in level $p-1$. We note that since QAOA is the $\ell \to 0$ and $h_j \to 0$ limit of the ab-QAOA, performance guarantees for the QAOA [18,37,52] apply also to the ab-QAOA.

III. NUMERICAL RESULTS

A. Comparison between QAOA and ab-QAOA

Our primary figure of merit is the time taken to reach a given accuracy r^*. The choice of a target accuracy r^* is to some extent arbitrary. We will take $r^* = 0.99$ as a value that is attainable in numerical simulations at moderate system sizes for the ab-QAOA and for the QAOA with reasonable extrapolations. This value of r^* also sets a goal that may be practical for future quantum computers in the medium term, and it exceeds the NP-hard threshold quoted above. The ratio of computation times for the QAOA and the ab-QAOA to reach the accuracy r^* is then our measurement of the improvement in the algorithm. We define p^* as the value of the level at which r^* is achieved.

To understand the dependence of the runtime on the level p, consider the optimization over a p-level output state from either QAOA or ab-QAOA. In the gradient-based classical optimization algorithm, $O(p)$ gradients are necessary, and for the calculation of each gradient, a p-level output state needs to be prepared. Intuitively, the total quantum computation time is $O(p^2)$, which will be analyzed rigorously in Appendix B.

Crucially, there is *no* additional quantum overhead in the ab-QAOA since *no* additional measurements are needed, and the number of gates for the state preparation *is the same* as in the QAOA. There is classical overhead due to the larger number of parameters. However, this cost turns out to be very small due to the fact that the only really important additional parameter is the bias field. This field is not optimized over, but rather simply fed back at each iteration, and the total number of these fields is only n, the number of qubits. These cost issues are treated in more detail in Appendix B. Given these considerations, the speedup is best defined as $S = (p^*_{\mathrm{QAOA}}/p^*_{\mathrm{ab-QAOA}})^2$. We will also plot the infidelity $1-F = 1-\sum_\alpha |\langle\psi_f^{s(\mathrm{ab})}|\psi_{\max}^\alpha\rangle|^2$ to compare the two methods (where α labels the degeneracy), since this quantity gives additional physical insight.

The results of the comparison of the ab-QAOA and QAOA algorithms for w3r graphs with $n = 8, 12, 16$ are shown in log-linear plots in Figs. 2(a) and 2(b) for $1-r$ and $1-F$, respectively, while the results for $n = 10, 14, 18$ are given in Appendix C. The convergence to the solution is much better in the ab-QAOA overall, in some cases by more than an order of magnitude. The improvement at small p is particularly striking. This is important, since only rather small values of p are likely to be accessible in near-term quantum machines [20–26].

It is surprising at first glance that the computed accuracy is not always significantly better for smaller graphs, as seen in Fig. 2(a) for $p = 7, 8$. This is due to the larger error bars at larger p, but the bars are magnified by the log scale. This inversion is discussed in more detail in Appendix C, where more extensive calculations are also presented.

To calculate the speedup, we need p^*. However, in the QAOA for larger n values, the algorithm does not achieve the desired accuracy r^* for $p \leqslant 8$. Thus some extrapolation is required, and this means choosing some fitting functions for $r(p)$, choosing the point where the curve intersects r^*, and rounding p at that point to the nearest integer. We fit both the w3r results and the u3r results in the standard QAOA using the purely empirical forms in [35].

The fits are surprisingly good. We have no good explanation for this at this point, but high-quality empirical fits often lead to later insights. We have also performed a scaling analysis, given in Appendix D, which shows that the points collapse onto a straight line in a rescaled plot. For the QAOA and w3r graphs, the fitting functions are

$$1-r = \exp(-\sqrt{p/p_0}+c),$$
$$1-F = \exp(-p/p_0+c). \quad (7)$$

The forms for the ab-QAOA for w3r graphs are slightly different, though we do not know at this point if the difference in the forms has any fundamental significance. The functions are

$$1-r = \exp(-\sqrt{p/p_0}+c),$$
$$1-F = \exp(-\sqrt{p/p_0}+c). \quad (8)$$

The fitting parameters p_0, c and the fitting errors are tabulated in Appendix D.

For the w3r results, the fitting functions work very well, as can be seen in Fig. 2. The upward curvature in the ab-QAOA fits is due to the fact that at higher p we are close to converging to the actual solution. It is notable that for the ab-QAOA, the curvature does not increase very rapidly with p, indicating that even when the ab-QAOA is quite close to the actual result, improvement still continues. The results for the relative infidelity of the QAOA and the ab-QAOA are nearly as impressive as those for the accuracy; the gap between the two methods is still clearly evident. In the ab-QAOA, $1-r$ is nearly independent of n, while $1-F$ changes noticeably. This is an indication that the energy spectrum of weighted graphs differs from that of unweighted graphs: the ground state for weighted graphs is more likely to be nearly degenerate with the low-lying excited states. This is shown by numerical calculation of the gap between the ground state and the first excited state, as shown in Fig. 3 for both types of graph.

The results for the u3r graphs with $n = 8, 12, 16$ are shown in Fig. 4 and the results for $n = 10, 14, 18$ are given in Appendix C. Again, the gap between the QAOA and the ab-QAOA is clearly evident. The initial convergence at small p is very fast for the ab-QAOA. Indeed, if the figure of merit

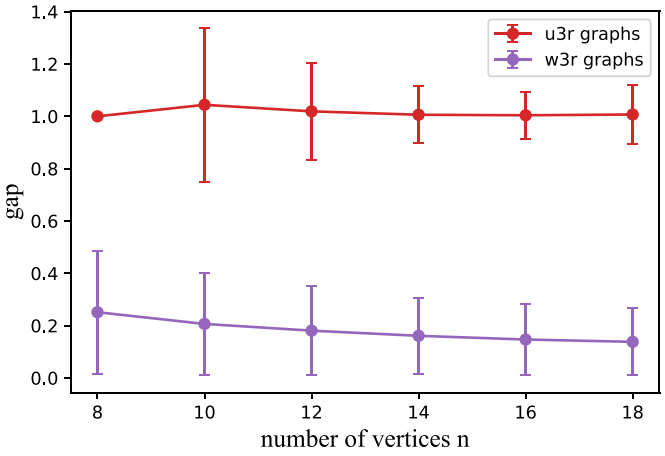

FIG. 3. The gap between the ground state and the first excited state of unweighted or weighted graphs for $n = 8, 10, 12, 14, 16, 18$. Each point is the average over 1000 graphs except $n = 8$ u3r graphs, where there are only five different nonisomorphic graphs. The error bars are standard deviations.

for the algorithms is taken as the accuracy at some fixed small p, the difference in performance for u3r graphs would exceed that for w3r graphs.

For u3r graphs, the fitting functions for QAOA are straight lines on the log-linear plots in Figs. 4(a) and 4(b):

$$1 - r = \exp(-p/p_0 + c),$$
$$1 - F = \exp(-p/p_0 + c), \tag{9}$$

while for ab-QAOA, the fitting functions are the same as those in w3r graphs, those in Eq. (8). Again, p_0 and c are fitting parameters that are given in Appendix D with fitting errors. The fits are generally good, with two slight exceptions. The first is when $n = 8$, for which there are few graphs so that little averaging can be performed. The second is the ab-QAOA at large p, where there is additional curvature that is not captured by the fits. In this region, the accuracy is so high that the curves must flatten out, and this introduces some finite-size error. Interestingly, the convergence rate of the infidelity of the ab-QAOA for u3r wave functions is considerably faster than that for the w3r case, and $1 - F$ does not depend so strongly on n. This is the main difference in our results for the w3r and u3r graphs. We believe that this is due to the fact that for the u3r graphs, the ground state is well-separated in energy from the low-lying excited states relative to the w3r graphs. This follows from the fact that the weight parameters in H_C for the u3r graphs are integers, but those in the w3r graphs are not.

The basic figure of merit for the ab-QAOA is $S(n)$, the speedup as a function of the number of vertices. This is plotted in Fig. 5. We see first of all that the improvement offered by the ab-QAOA is certainly not limited to very small graphs. If $S(n) = O(n^a)$, where $a > 0$, then the ab-QAOA gives a polynomial speedup over QAOA. If $a = 0$, then we can only hope for a constant speedup (which might still be of practical importance, of course). The curve in Fig. 5 shows no sign of saturating up to $n = 18$. The numerical results give evidence that the ab-QAOA gives a significant speed-up for the solution of the MaxCut problem compared to other classical-

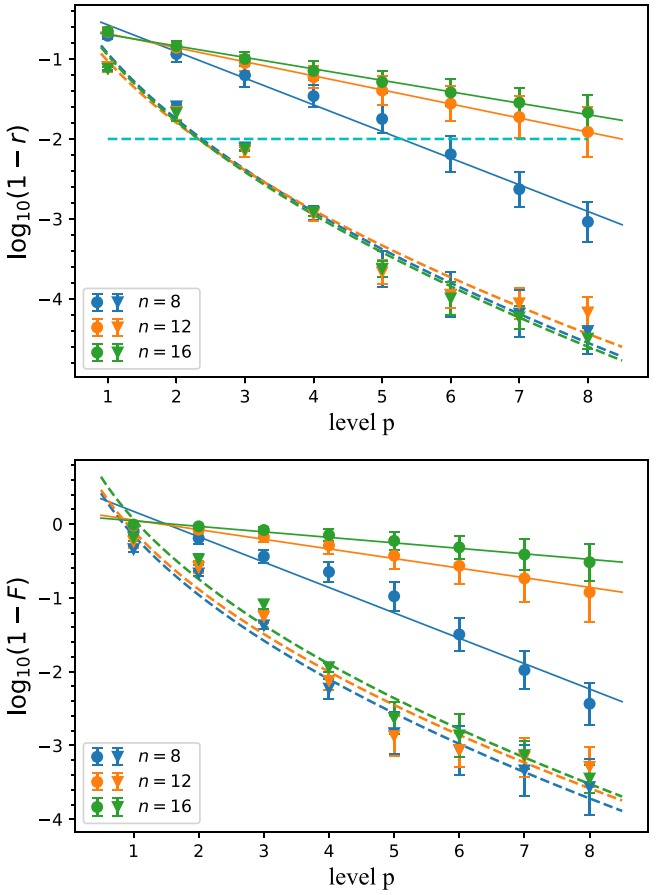

FIG. 4. Comparison of the accuracy (top panel) and the infidelity (bottom panel) for QAOA (solid lines) and ab-QAOA (dashed lines) for $n = 8, 12, 16$ for u3r graphs. The number of realizations is 40 for graphs with 10 or more vertices. For $n = 8$, there are only five different nonisomorphic u3r graphs, and we average over them. The error bars are standard deviations. In (a) the accuracies for QAOA and ab-QAOA are plotted vs p. The horizontal dashed line again represents $r^* = 0.99$. By comparing with Fig. 2, we find that the accuracy for both algorithms is slightly better when applied to unweighted graphs. The improvement of ab-QAOA over QAOA is very marked at even smaller p. In (b) the infidelities for QAOA and ab-QAOA are plotted as a function of p, now for unweighted graphs. Again, the improvement is clear at quite small p, and it continues to improve for all p. The fits are described in the text.

quantum hybrid optimization algorithms. We expect that the ab-QAOA can also be applied to other classical combinatorial optimization algorithms that can be mapped into classical Ising models, such as partitioning problems, covering and packing problems, and so on [53,54], since then the local field idea can be suitably modified. For problems that lack such a mapping, such as finding the ground states of molecules [11], the applicability of the ab-QAOA is much less clear.

B. Effect of bias fields for level $p = 1$

In this section, we illustrate the effect of bias fields for the smallest nontrivial graphs (the first graph in Fig. 6) and only at level $p = 1$. This isolates the effect of having these fields in the mixing Hamiltonian. We simply repeat the evolution,

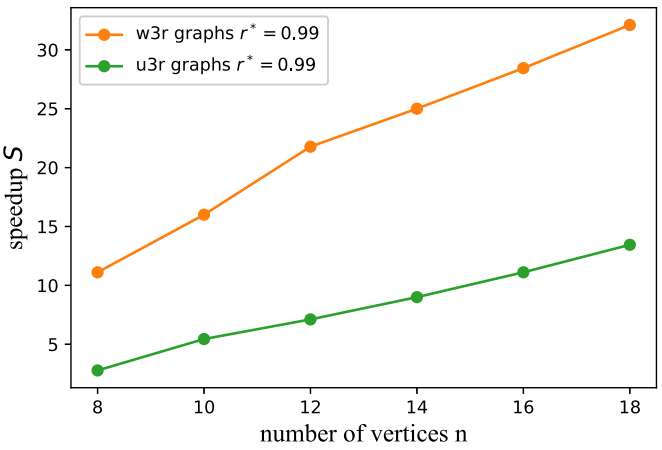

FIG. 5. Speedup $S(n)$ of ab-QAOA over standard QAOA, using an accuracy $r^* = 0.99$ as a criterion. The corresponding p^* values are given in Appendix E.

measuring $\langle H_C \rangle$ at the end of each step, and then we update the bias fields in the mixing Hamiltonian and the starting wave function using the prescription given above. Of course this leads to lower fidelities than for the full algorithm presented in Sec. III A.

In the ab-QAOA, the mixing Hamiltonian with bias field is

$$H_M^{\mathrm{ab}}(\{h_j\}) = \sum_{j=1}^{n} (X_j - h_j Z_j). \tag{10}$$

Once we know one product ground state $|\psi_{\max}^{\alpha}\rangle$ of the Max-Cut problem Hamiltonian in Eq. (2) (whose ground states are always degenerate and α is used to eliminate this degeneracy), then we have the expectation value of each Z_j. If h_j is fixed to $\langle \psi_{\max}^{\alpha} | Z_j | \psi_{\max}^{\alpha} \rangle$ in our ab-QAOA, then $|\psi_0^{\mathrm{ab}}\rangle$ is closer to $|\psi_{\max}^{\alpha}\rangle$ than $|-\rangle^{\otimes n}$ (the starting state of the standard QAOA), leading to a higher accuracy for the ab-QAOA.

The bias field parameter h_j is updated according to

$$h_j \to h_j - \ell(h_j - \langle Z_j \rangle). \tag{11}$$

This update strategy will bring h_j closer to $\langle \psi_{\max}^{\alpha} | Z_j | \psi_{\max}^{\alpha} \rangle$ and the starting state $|\psi_0^{\mathrm{ab}}\rangle$ closer to $|\psi_{\max}^{\alpha}\rangle$. In realistic calculations, prior knowledge of $|\psi_{\max}^{\alpha}\rangle$ may not be available. It turns out that we can still find $|\psi_{\max}^{\alpha}\rangle$ faster than the QAOA even without prior knowledge of $|\psi_{\max}^{\alpha}\rangle$, as we now show.

To illustrate this, we calculate the fidelity $\sum_{\alpha} |\langle \psi_0^{\mathrm{ab}} | \psi_{\max}^{\alpha} \rangle|^2$, where $|\psi_0^{\mathrm{ab}}\rangle$ is the ground state of $H_M^{\mathrm{ab}}(\{h_j\})$ in the level-1 ab-QAOA for the first u3r graph with eight vertices as shown in Fig. 6. The sum is over the ground states to which ab-QAOA steers the starting

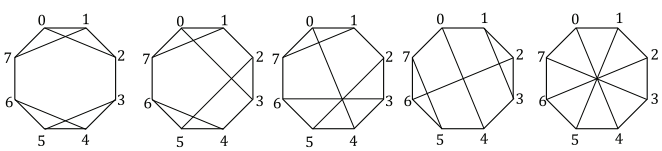

FIG. 6. All five different u3r graphs with eight vertices. These graphs are labeled as 1, 2, ... in sequence.

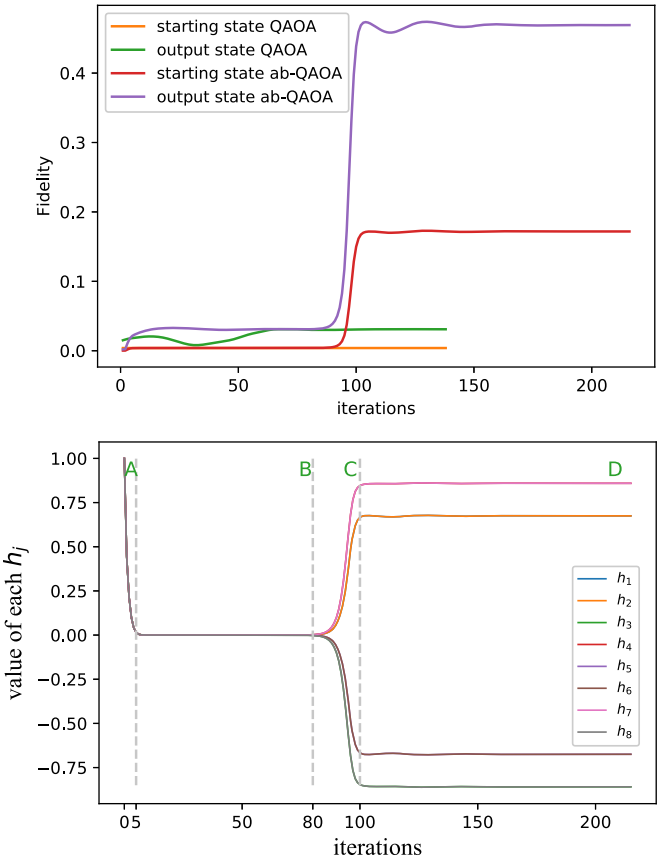

FIG. 7. The fidelity (top panel) between the target ground state and the starting or output states for graph 1 in Fig. 6 using level-1 QAOA and level-1 ab-QAOA, and the corresponding ab-QAOA bias fields (bottom panel), all plotted against the number of iterations. The target state is $|\psi_{\max}^{\alpha}\rangle = |00101101\rangle$. The fidelities show sudden jumps that correspond to rapid changes in the bias fields of the ab-QAOA. The fields steer the starting state to the desired output state. h_1, h_2, h_4, and h_7 are greater than 0 at the end of the optimization, corresponding to the form of $|\psi_{\max}^{\alpha}\rangle$. The ranges A, B, C, and D are discussed in the text.

state. For comparison, we also calculate $\sum_{\alpha} |\langle \psi_0^{\mathrm{s}} | \psi_{\max}^{\alpha} \rangle|^2$, $\sum_{\alpha} |\langle \psi_f^{\mathrm{s}} | \psi_{\max}^{\alpha} \rangle|^2$, and $\sum_{\alpha} |\langle \psi_f^{\mathrm{ab}} | \psi_{\max}^{\alpha} \rangle|^2$ in Fig. 7, where $|\psi_0^{\mathrm{s}}\rangle$ is the starting state of the standard QAOA, $|\psi_f^{\mathrm{s}}\rangle$ is the output state produced by the QAOA, and $|\psi_f^{\mathrm{ab}}\rangle$ is the output state produced by the ab-QAOA.

As shown in Fig. 7(a), it is clear that the bias field will bring the starting state closer to the ground state of H_C. There are some iterations for which both the starting state and output state curves of ab-QAOA grow rapidly, and that is when the bias field brings the starting state $|\psi_0^{\mathrm{ab}}\rangle$ close to the ground states. The operations $\exp(-i\beta_k H_M^{\mathrm{ab}})$ and $\exp(-i\gamma_k H_c)$ bring $|\psi_f^{\mathrm{ab}}\rangle$ closer to the target states than $|\psi_0^{\mathrm{ab}}\rangle$. Note that the fidelity approaches 0.5 for the output state of the ab-QAOA. This is bounded above by the ab-QAOA driven by fixed bias fields $h_j = \langle \psi_{\max}^{\alpha} | Z_j | \psi_{\max}^{\alpha} \rangle$ with $\ell = 0$.

To better investigate how the bias fields work, we also plot $\{h_j\}$ of graph 1 from level 1 ab-QAOA in the optimization iterations, as shown in Fig. 7(b). There are four regions in Fig. 7(b). In region A, all h_j decrease to 0 quickly in the first

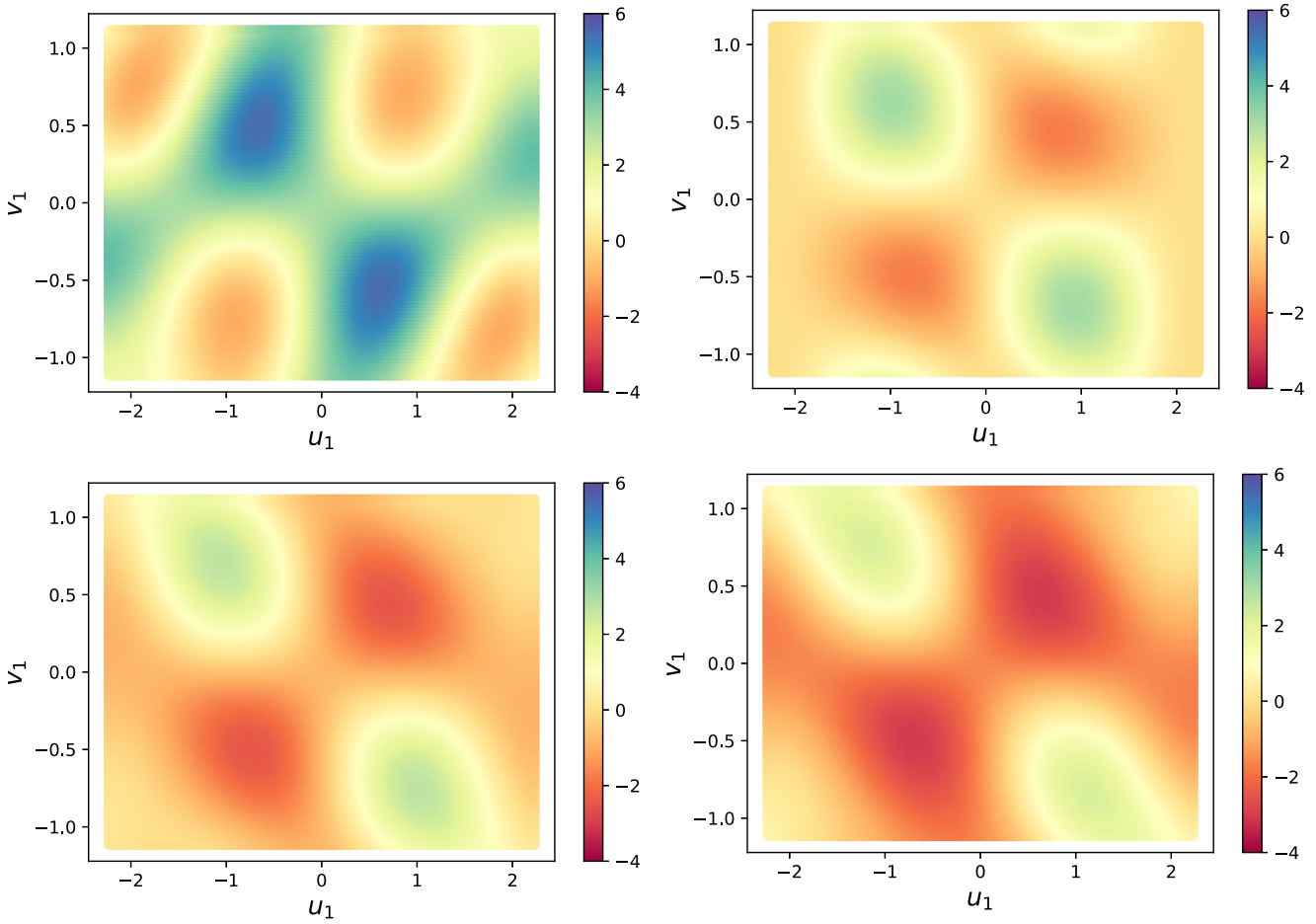

FIG. 8. The energy landscape of graph 1 as a function of the variational parameters u_1 and v_1 for increasing number of iterations. The 0th (top left), 50th (top right), 95th (bottom left), and final (bottom right) iterations that are shown belong to region A, B, C, and D, respectively. As analyzed in Ref. [35], γ_1 and β_1 can be restricted to $[-\pi/2, \pi/2]$ and $[-\pi/4, \pi/4]$, respectively, so u_1 and v_1 are restricted to $[-\sqrt{2}\pi/2, \sqrt{2}\pi/2]$ and $[-\sqrt{2}\pi/4, \sqrt{2}\pi/4]$ according to Eq. (6).

five iterations. In region B, from the 5th iteration to the 80th iteration, all h_j are near 0. In region C, from about the 80th iteration to the 100th iteration, the $\{h_j\}$ diverge and each h_j tries to find its true value, $\langle\psi^{\alpha}_{\max}|Z_j|\psi^{\alpha}_{\max}\rangle$. In region D, in the last half of the optimization, the value of each h_j does not change. The behavior of the fidelity in Fig. 7(a) is related to $\{h_j\}$ in Fig. 7(b). The divergence of $\{h_j\}$ from 0 implies a sharp rise in the fidelity.

For each of these four regions, we choose four specific points and plot the energy landscape using ab-QAOA as shown in Fig. 8. Note that for region B, the landscape is in close agreement with that of the QAOA since all h_j are small. In region A, due to the "wrong" bias fields, it is harder to find the target state using ab-QAOA than the QAOA, so the QAOA output state can be regarded as the optimal state of ab-QAOA's. As a result, each h_j moves towards 0 fast in

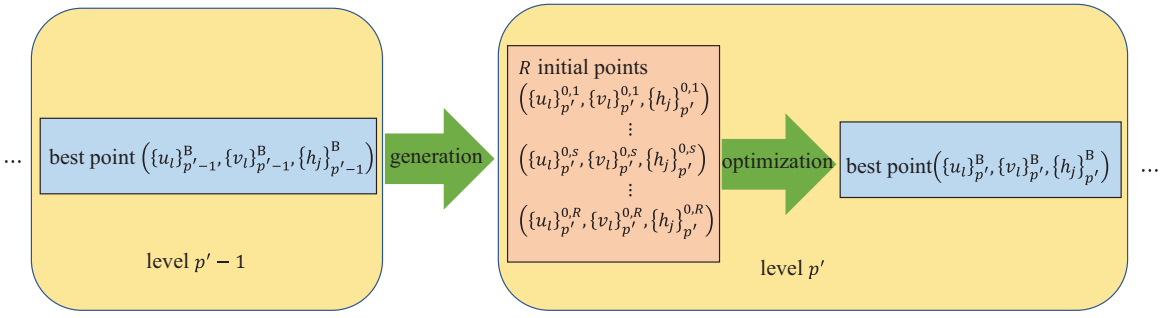

FIG. 9. Schematics of the outer loop of ab-QAOA. Using Eq. (A6), we generate R initial points in level p' from the best point in level $p'-1$. After the optimization of these R points, we get the point with the best energy. We do this procedure iteratively until the target level p.

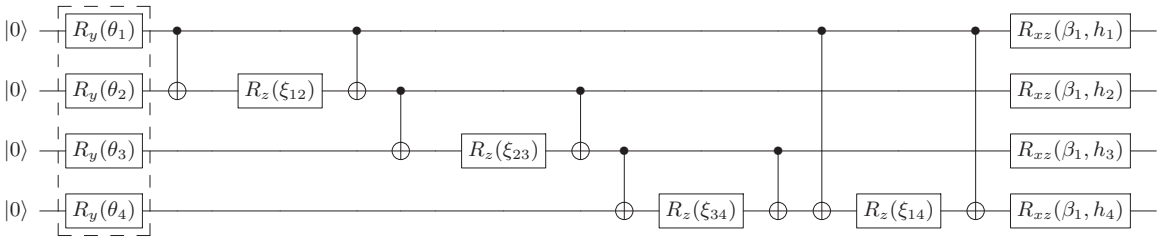

FIG. 10. Quantum circuit for 1-level ab-QAOA on 2-regular graphs with four vertices. $\xi_{v_1v_2}$ is the real coefficient of $Z_{v_1}Z_{v_2}$ appearing in $\exp(-i\gamma_1 H_C)$. R_y and R_z are the rotation operators around the $\hat{y}$ and $\hat{z}$ axis, respectively, while $R_{xz}(\beta_1, h_j) = \exp[-i\beta_1(X_j - h_j Z_j)]$. When $h_j = 0$, R_{xz} is the rotation operator around the $\hat{x}$ axis. The gates in the dashed box prepare the starting state for ab-QAOA. There are $4 + 3 \times 4 + 4 = 20$ gates in the circuit.

region A. In region B, although all h_j are small, their effects accumulate until the bias fields can have a significant effect on the cost function. In region C, all h_j change quickly because of the accumulation in region B. In the updating, all h_j are getting closer to their true values, so it is easier to find the target state in this region, which can be verified by the smaller lowest energy in the landscape. In region D, the output energy nearly meets the convergence criterion, and each h_j finds its true value, $\langle\psi^{\alpha}_{\max}|Z_j|\psi^{\alpha}_{\max}\rangle$, so the lowest energy is smaller than in the other regions.

IV. CONCLUSION

In this paper, we have shown how a generalization of the QAOA, the ab-QAOA, can greatly reduce the depth of the quantum circuit needed to solve optimization problems to a given accuracy. The understanding of this comes partly from

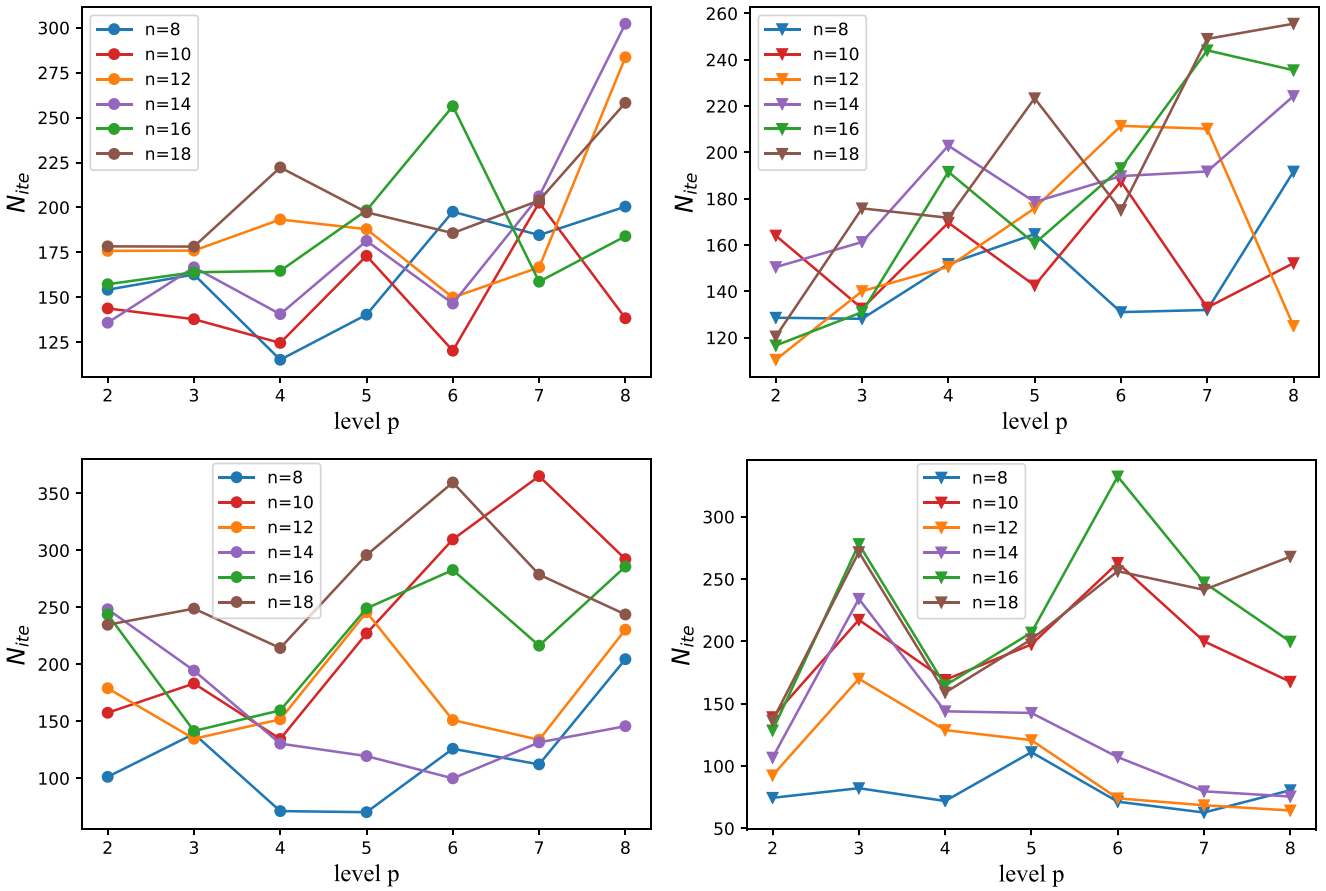

FIG. 11. Iterations needed for convergence N_{ite} in u3r graphs and w3r graphs for points generated by the above outer loop. The top left panel is QAOA for w3r graphs, the top right panel is ab-QAOA for w3r graphs, the bottom left panel is QAOA for u3r graphs, and the bottom right panel is ab-QAOA for w3r graphs. N_{ite} is the average over R samples. The classical optimizer is the Adam gradient-based stochastic optimization algorithm mentioned above. N_{ite} is very similar for different graphs and for the two different algorithms.

a study of the effects of the bias fields on small graphs. In the short and medium term (NISQ era), the results presented in Figs. 2 and 4 are the most important ones. They show that a quantum computer with of order 20 very high-quality logical qubits may produce impressive results at level $p = 5$, a machine that may be attainable quite soon [20]. In the longer term, we are more interested in how the performance of the ab-QAOA scales with n. Figure 5 shows that the speedup in fact increases as the size of the system increases, which suggests that the ab-QAOA may still be the algorithm of choice beyond the NISQ era.

The ability to carry out practical calculations in the NISQ era will depend on finding algorithms that can be implemented in circuits of relatively shallow depth, and converge quickly to an answer. The ab-QAOA contributes to these goals, converging to a desired accuracy with a computation time that is polynomially shorter in the system size than that of the standard QAOA.

ACKNOWLEDGMENTS

Y.Y. and X.-B.W. acknowledge National Natural Science Foundation of China Grants No. 11974204 and No. 12174215. N.S. acknowledges the support of the Theory of Quantum Matter Unit, Okinawa Institute of Science and Technology Graduate University (OIST). This research was performed using the computing resources and assistance of the UW-Madison Center For High Throughput Computing. We thank M. G. Vavilov and B. Özguler for useful discussions.

APPENDIX A: COMPUTATIONAL DETAILS OF THE AB-QAOA

In this Appendix, we give further details of the Fourier strategy proposed in [35] about how to choose the starting points in the optimization. For a p-level QAOA, as stated in the main text, the mixing Hamiltonian is $H_M^s = \sum_j X_j$. The quantum processor is initialized in $|\psi_0^s\rangle$, the ground state of H_M^s. Then we alternately apply problem Hamiltonian H_C and mixing Hamiltonian H_M^s to generate the final state,

$$|\psi_f^s\rangle = \prod_{k=1}^{p} e^{-i\beta_k H_M^s} e^{-i\gamma_k H_C} |\psi_0^s\rangle, \tag{A1}$$

where the level p is the number of times the unitary operators corresponding to H_M^s and H_C are applied to the initial state to move it to the final state. The scheduling parameters $\{\gamma_k\}, \{\beta_k\}$ in the operators are determined by optimizing

$$\langle H_C\rangle(\{\gamma_k\}, \{\beta_k\}) = \langle\psi_f^s|H_C|\psi_f^s\rangle. \tag{A2}$$

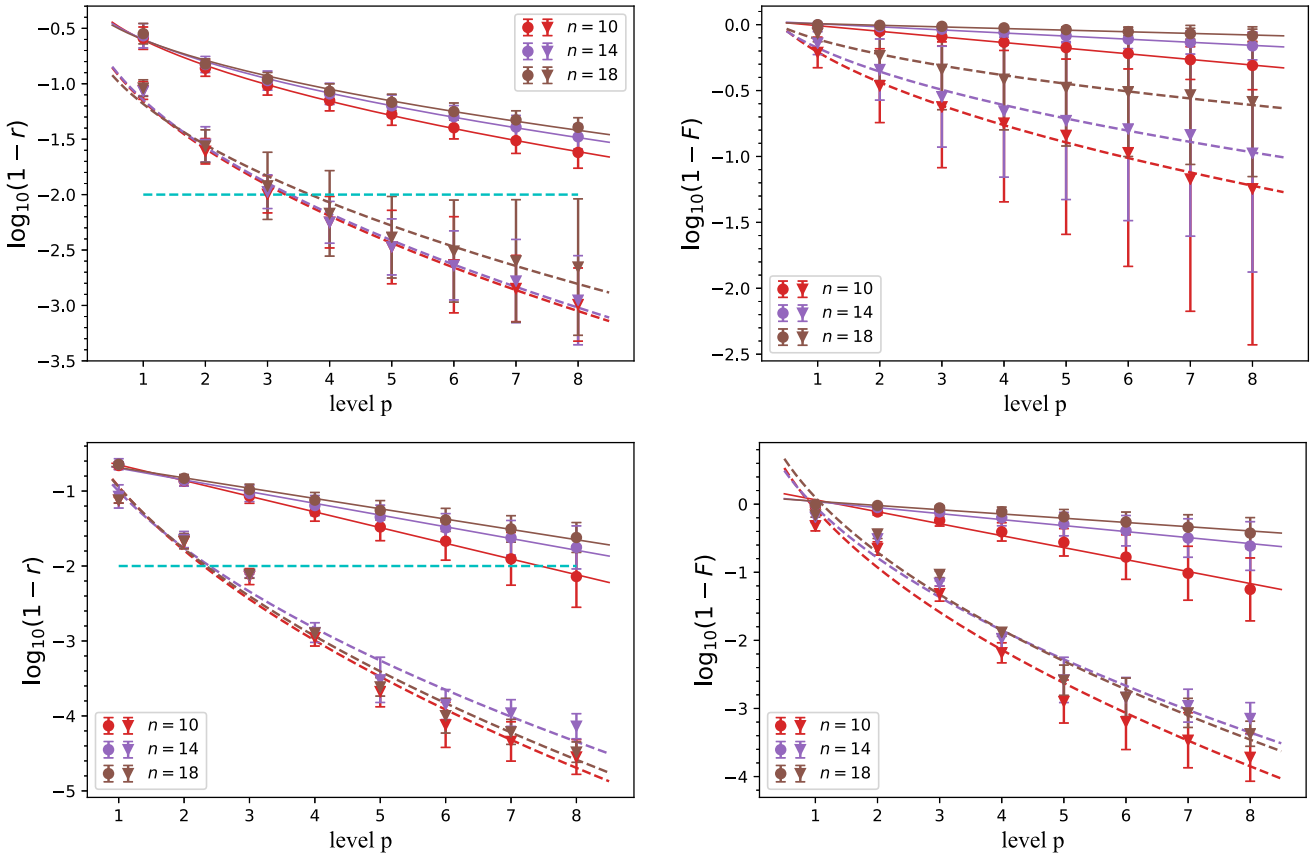

FIG. 12. Comparison of the accuracy $[\log_{10}(1-r)]$ and infidelity $[\log_{10}(1-F)]$ of results for QAOA (solid lines) and ab-QAOA (dashed lines) for $n = 10, 14, 18$ for w3r graphs and u3r graphs. The top row is w3r graphs, and the bottom row is u3r graphs. The horizontal dashed line represents $r^* = 0.99$. Each point is an average over 40 randomly chosen graphs. The fits are described in the main text. The error bars are standard deviations. As in Figs. 2 and 4 in the main text, the speedup of ab-QAOA over QAOA is clearly evident. The performances of both algorithms are slightly better in the unweighted case.

Note that for the original QAOA [18,35], $|\psi_0^{\rm s}\rangle$ is $|+\rangle^{\otimes n}$, but in our description, $|\psi_0^{\rm s}\rangle$ is $|-\rangle^{\otimes n}$, the ground state of $H_M^{\rm s}$. If we denote the QAOA final state from $|+\rangle^{\otimes n}$ by $|\psi_f^{\rm s+}\rangle$, the final state from $|-\rangle^{\otimes n}$ by $|\psi_f^{\rm s-}\rangle$, and define $\tilde{Z} = \prod_j Z_j$, it is easy to prove that

$$\begin{aligned}|\psi_f^{\rm s-}(\{\gamma_k\},\{\beta_k\})\rangle &= \prod_{k=1}^{p} e^{-i\beta_k H_M^{\rm s}} e^{-i\gamma_k H_C}\tilde{Z}|+\rangle^{\otimes n}\\ &= \tilde{Z}\prod_{k=1}^{p} e^{i\beta_k H_M^{\rm s}} e^{-i\gamma_k H_C}|+\rangle^{\otimes n}\\ &= \tilde{Z}|\psi_f^{\rm s+}(\{\gamma_k\},\{-\beta_k\})\rangle. \end{aligned} \tag{A3}$$

There is no difference in the classical optimization for both $|\psi_f^{\rm s-}\rangle$ and $|\psi_f^{\rm s+}\rangle$ as long as H_C is a classical Ising Hamiltonian since

$$\langle\psi_f^{\rm s-}|H_C|\psi_f^{\rm s-}\rangle(\{\gamma_k\},\{\beta_k\}) = \langle\psi_f^{\rm s+}|H_C|\psi_f^{\rm s+}\rangle(\{\gamma_k\},\{-\beta_k\}). \tag{A4}$$

There are two differences in the quantum part of the two algorithms.

(i) In the p-level ab-QAOA, $H_M^{\rm ab}$ contains local longitudinal fields as well as the usual global transverse field, $H_M^{\rm ab}(\{h_j\}) = \sum_{j=1}^{n}(X_j - h_j Z_j)$.

(ii) The wave function at the initial stage of each learning step is chosen to be the ground state of the updated $H_M^{\rm ab}$.

Hence both the longitudinal fields in the mixing Hamiltonian and the "reinitialized" wave function change during the course of the ab-QAOA algorithm.

Thus the final state of ab-QAOA is

$$\left|\psi_f^{\rm ab}\right\rangle = \prod_{k=1}^{p} e^{-i\beta_k H_M^{\rm ab}(\{h_j\})} e^{-i\gamma_k H_C}\left|\psi_0^{\rm ab}(\{h_j\})\right\rangle. \tag{A5}$$

In the Fourier strategy [35], what we optimize over are $\{u_l\}$ and $\{v_l\}$, the Fourier components of $\{\gamma_k\}$ and $\{\beta_k\}$. To avoid being trapped in the local optimum as far as possible, we start the optimization from R initial points and find the point with the best energy, as was done in Ref. [35]. To reach level p, we start from level 1 and find the point with the best energy from R initial points after the optimization. In level 2, R initial points are generated by adding some random numbers to the best point in level 1. Then repeat the optimization and initial point generation procedure with increasing level p' until $p' = p$.

Here, we use p to represent the target level and p' to represent the inner levels. The whole process from level 1 to level p including the initial points generation is the outer loop of the algorithm, while the update of $\{u_l\}$, $\{v_l\}$, $\{h_j\}$ until convergence for a fixed level p' (the algorithm presented in Sec. II B) is the inner loop of the algorithm. The same loops are used for the QAOA except that all the h_j are set to 0.

This more elaborate classical optimization is not strictly necessary to demonstrate the advantages of the ab-QAOA over the QAOA, but it does mean that the results can be compared more directly with those of Ref. [35]. The sampling parameter R was set to 10 in our calculations. We note once more that formally the QAOA can be considered as the limit of the ab-QAOA when $h_j \to 0$. This means that from a formal standpoint, the ab-QAOA is guaranteed to be at least as good as the QAOA. The detailed outer loop of the algorithm follows and is also illustrated in Fig. 9.

Algorithm: Outer Loop from level 1 to *p*

1. In level 1, we generate R initial "0" points $(\{u_l\}_1^{0,s}, \{v_l\}_1^{0,s}, \{h_j\}_1^{0,s})$, where the elements of $\{u_l\}_1^{0,s}$ and $\{v_l\}_1^{0,s}$ are random numbers drawn from a uniform distribution and all elements of $\{h_j\}_1^{0,s}$ are initialized to be 1. The subscripts refer to the ab-QAOA level in the outer loop, and the s superscript ranges from 1 to R representing the different initial points. Using the algorithm in Sec. II B we get the optimal "B" point $(\{u_l\}_1^{\rm B}, \{v_l\}_1^{\rm B}, \{h_j\}_1^{\rm B})$ with the best optimal energy $E_1^{\rm B}$ from R points for this level.
2. In level p' greater than 1, we use the best point $(\{u_l\}_{p'-1}^{\rm B}, \{v_l\}_{p'-1}^{\rm B}, \{h_j\}_{p'-1}^{\rm B})$ in level $p'-1$ to construct R initial points $(\{u_l\}_{p'}^{0,s}, \{v_l\}_{p'}^{0,s}, \{h_j\}_{p'}^{0,s})$. The s superscript refers to the elements of the following random selection procedure, representing the different points:

$$\begin{aligned}\{u_l\}_{p'}^{0,s} &= \begin{cases}\{u_l\}_{p'-1}^{\rm B}\cup\{0\}, & s=1,\\ \{u_l+\alpha{\rm Ran}^s[u_l]\}_{p'-1}^{\rm B}\cup\{0\}, & 2\leqslant s\leqslant R,\end{cases}\\ \{v_l\}_{p'}^{0,s} &= \begin{cases}\{v_l\}_{p'-1}^{\rm B}\cup\{0\}, & s=1,\\ \{v_l+\alpha{\rm Ran}^s[v_l]\}_{p'-1}^{\rm B}\cup\{0\}, & 2\leqslant s\leqslant R,\end{cases}\\ \{h_j\}_{p'}^{0,s} &= \begin{cases}\{h_j\}_{p'-1}^{\rm B}, & s=1,\\ \{h_j+\alpha{\rm Ran}^s[h_j]\}_{p'-1}^{\rm B}, & 2\leqslant s\leqslant R.\end{cases}\end{aligned} \tag{A6}$$

   $\{u_l\}_{p'}^{0,s}$ or $\{v_l\}_{p'}^{0,s}$ is a p'-element set whose p'th element is zero. The random number ${\rm Ran}^s[a]$ is the sth selection from a normal distribution with average 0 and variance a^2, i.e., ${\rm Ran}^s[a] = {\rm Norm}(0,a^2)$. We optimize these R initial points to find the best point $(\{u_l\}_{p'}^{\rm B}, \{v_l\}_{p'}^{\rm B}, \{h_j\}_{p'}^{\rm B})$ with the best energy $E_{p'}^{\rm B}$. The update parameter α was set to $\alpha = 0.6$.
3. Repeat step 2 until p' reaches the target level p.
4. Output all energies $E_{p'}^{\rm B}$ from level 1 to p.

APPENDIX B: COMPUTATION TIME

Here we give the analysis that leads to the conclusion in the main text that the total computation time is $O(p^2)$. We assume that the quantum part of the algorithm dominates the time. This will be true for the foreseeable future. The MaxCut cost Hamiltonian H_C is defined on an n-vertex $\mathcal{R}$-regular graph, and a p-level QAOA and ab-QAOA are implemented with optimization to find a target state. In our calculations in the main text, $\mathcal{R} = 3$. We denote the iterations needed for convergence by $N_{\rm ite}$. In each iteration of the optimization in our calculation, we need to calculate the expectation of the problem Hamiltonian $\langle H_C\rangle$ $2p+1$ times to get gradients of the input parameters. In both of these two QAOA, the gradient of E_p (the energy in one iteration for the p level QAOA or ab-QAOA) with respect to the $u_{l'}$ is

$$\begin{aligned}\frac{\partial E_p}{\partial u_{l'}} &= \frac{E_p(\{u_l\}',\{v_l\},\{h_j\}) - E_p(\{u_l\},\{v_l\},\{h_j\})}{\epsilon_g},\\ \{u_l\}' &= \{u_1,u_2,\ldots,u_{l'}+\epsilon_g,\ldots\},\end{aligned} \tag{B1}$$

where ϵ_g is a small quantity. There are p u_l, so p $E_p(\{u_l\}',\{v_l\},\{h_j\})$ and one $E_p(\{u_l\},\{v_l\},\{h_j\})$ are needed. As a result, $2p+1$ calculations of $\langle H_C\rangle$ are needed.

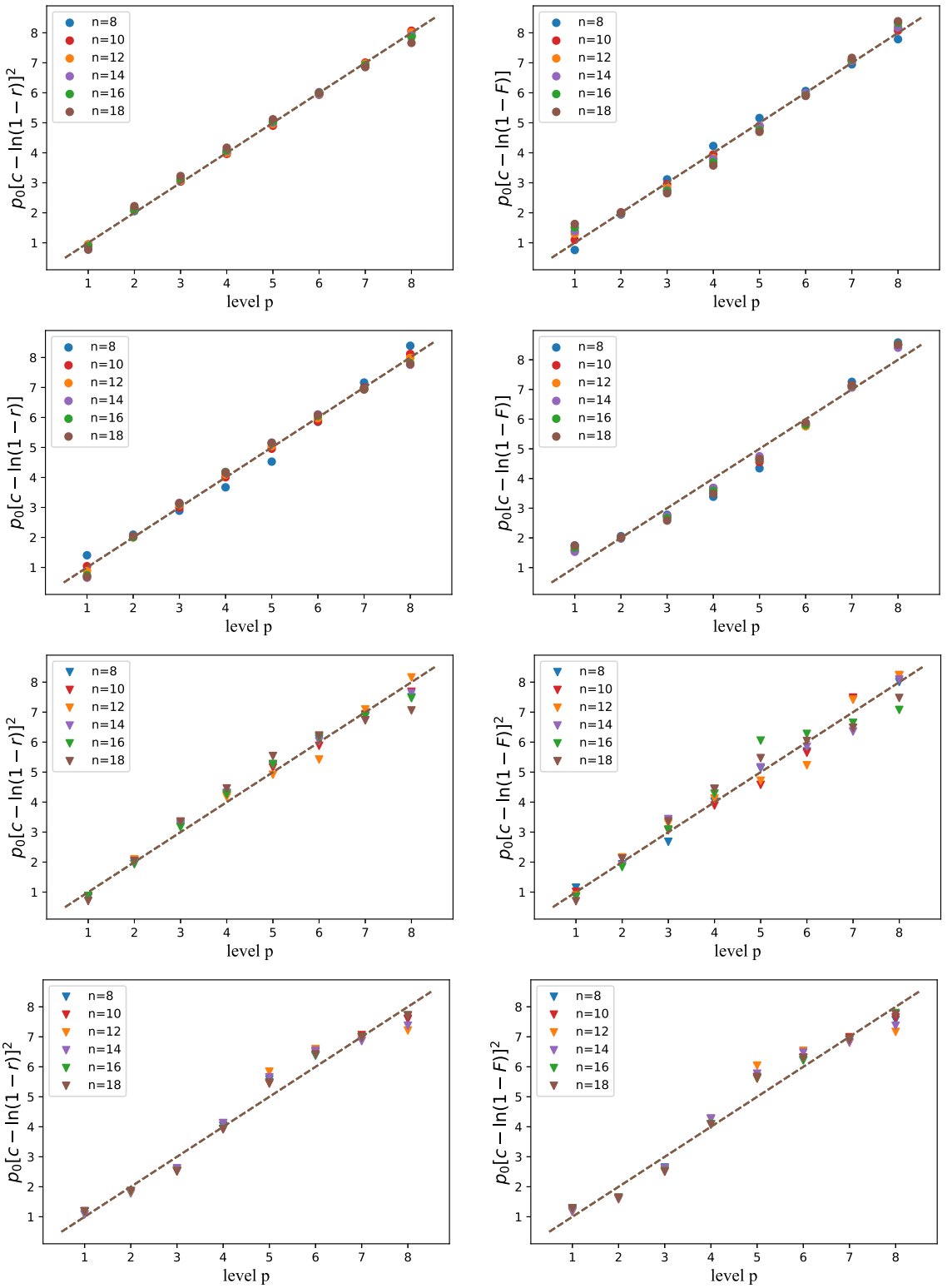

FIG. 13. Fits to the accuracy and infidelity curves for QAOA (top two rows) and ab-QAOA (bottom two rows). Rows 1 and 3 are for w3r graphs, and rows 2 and 4 are for w3r graphs. The dashed lines in the four subplots represent $p = p_0[c - \ln(1-r)]^2$ or $p = p_0[c - \ln(1-F)]^2$, which is equivalent to the fitting functions in the main text.

In a single calculation of $\langle H_C \rangle$, one needs to measure $n\mathcal{R}/2$ different ZZ terms of H_C. $|\psi_f\rangle$ (either $|\psi_f^{\rm s}\rangle$ or $|\psi_f^{\rm ab}\rangle$ in the main text) is prepared M_{ZZ} times to get an accurate expectation value for the ZZ term.

In the ab-QAOA, unlike the QAOA, knowledge of the Z term is also needed to guide $\{h_j\}$ in the flowing iteration. However, this does not require an additional measurement, since if we have the value of $\langle ZZ \rangle$ measured in the compu-

tational basis, we automatically also know $\langle Z\rangle$, as we now show. Consider a single ZZ term, $Z_{v_1}Z_{v_2}$. It has a spectral decomposition

$$\begin{aligned} Z_{v_1}Z_{v_2} = {} & |0_{v_1}\rangle\langle 0_{v_1}| \otimes |0_{v_2}\rangle\langle 0_{v_2}| - |0_{v_1}\rangle\langle 0_{v_1}| \otimes |1_{v_2}\rangle\langle 1_{v_2}| \\ & - |1_{v_1}\rangle\langle 1_{v_1}| \otimes |0_{v_2}\rangle\langle 0_{v_2}| + |1_{v_1}\rangle\langle 1_{v_1}| \otimes |1_{v_2}\rangle\langle 1_{v_2}|, \end{aligned} \tag{B2}$$

where $|1_{v_1}\rangle\langle 1_{v_1}| \otimes |1_{v_2}\rangle\langle 1_{v_2}|$ is short for $\mathbb{I} \otimes \cdots \otimes \underbrace{|1\rangle\langle 1|}_{v_1} \otimes \cdots \otimes \underbrace{|1\rangle\langle 1|}_{v_2} \otimes \cdots \otimes \mathbb{I}$, which is denoted as $T_{11}^{v_1v_2}$, so as for $T_{10}^{v_1v_2}$, $T_{01}^{v_1v_2}$, and $T_{00}^{v_1v_2}$. Once these four T operators are measured, then $\langle Z\rangle$ can be obtained:

$$\begin{aligned} \langle Z_{v_1}\rangle &= \left\langle T_{00}^{v_1v_2}\right\rangle + \left\langle T_{01}^{v_1v_2}\right\rangle - \left\langle T_{10}^{v_1v_2}\right\rangle - \left\langle T_{11}^{v_1v_2}\right\rangle, \\ \langle Z_{v_2}\rangle &= \left\langle T_{00}^{v_1v_2}\right\rangle + \left\langle T_{10}^{v_1v_2}\right\rangle - \left\langle T_{01}^{v_1v_2}\right\rangle - \left\langle T_{11}^{v_1v_2}\right\rangle. \end{aligned} \tag{B3}$$

As a result, there are no additional measurements needed in the ab-QAOA compared to the QAOA.

In one preparation of $|\psi_f\rangle$, the operator $\exp(-i\gamma_k H_C)$ is applied p times and $\exp(-i\beta_k H_M^{\mathrm{s}})$ or $\exp(-i\beta_k H_M^{\mathrm{ab}})$ is applied p times. The operator $\exp(-i\gamma_k H_C)$ can be decomposed into three quantum gates while $\exp(-i\beta_k H_M^{\mathrm{ab}})$ can be represented by one, as shown in Fig. 10, so $p(3n\mathcal{R}/2 + n)$ quantum gates are needed. In the meanwhile, for ab-QAOA, n R_y rotation gates around the $\hat{y}$ axis are needed for the starting state preparation from $|0\rangle^{\otimes n}$, and for the QAOA, n Hadamard gates are needed.

In conclusion, there are

$$N_{\mathrm{gate}} = N_{\mathrm{ite}}(2p+1)M_{ZZ}\frac{n\mathcal{R}}{2}\left[p\left(\frac{3n\mathcal{R}}{2}+n\right)+n\right] \tag{B4}$$

quantum gates for a p-level QAOA or ab-QAOA with full optimization, $N_{\mathrm{gate}} \sim O(N_{\mathrm{ite}}p^2n^2\mathcal{R}^2)$.

In our simulation, there are two kinds of initial points. One kind is the randomly generated points in level 1, and the other one is the points generated with the above outer loop in the other levels. Since p^* is always larger than 1 for $r^* = 0.99$, we focus on N_{ite} when the level $p \geqslant 2$. In this case, the iterations are similar among different graphs and between the QAOA and ab-QAOA, as shown in Fig. 11. So we conclude that N_{ite} is the same constant for different levels and for both algorithms, so $N_{\mathrm{gate}} \sim O(p^2n^2\mathcal{R}^2)$. The additional classical cost for the ab-QAOA is only a small constant. This is because essentially the entire classical cost is in the optimization routine, which does not depend on whether bias fields are included, since these fields are not optimized over.

Of course this analysis assumes that there is no error correction. It also assumes that two-qubit gates can be applied to any pair of qubits, thus avoiding the necessity of SWAP gates. These considerations apply equally to QAOA and ab-QAOA, so they should not affect the speedup that is defined in the main text since it is a *relative* speedup. Similarly, $\mathcal{R}$ and n are the same for the two algorithms, and the same reasoning may be applied. For a given accuracy and problem size, only p is different.

TABLE I. Fitting parameters p_0 and c of QAOA and ab-QAOA for w3r graphs and u3r graphs. e_{p_0} and e_c represent the standard deviation errors. Top left entry in each table specifies accuracy (r) or infidelity (F), algorithm, and graph type.

r for QAOA (w3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	0.4280	0.6223	0.7332	0.8023	0.9069	0.9443
e_{p_0}	0.0479	0.0514	0.0340	0.0592	0.0443	0.0925
c	0.1074	−0.1276	−0.2325	−0.2635	−0.3967	−0.3586
e_c	0.0333	0.0430	0.0308	0.0562	0.0448	0.0953
r for ab-QAOA (w3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	0.1733	0.1758	0.1770	0.1796	0.1842	0.2392
e_{p_0}	0.4616	0.4285	0.4949	0.4619	0.4258	0.6381
c	−0.2451	−0.2796	−0.2611	−0.2788	−0.2483	−0.6791
e_c	0.2038	0.1906	0.2209	0.2076	0.1939	0.3310
F for QAOA (w3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	8.4398	10.1726	12.8145	18.5465	29.5196	35.3103
e_{p_0}	0.0034	0.0012	0.0023	0.0019	0.0017	0.0018
c	0.0340	0.0805	0.0881	0.0695	0.0485	0.0453
e_c	0.0170	0.0060	0.0116	0.0095	0.0085	0.0089
F for ab-QAOA (w3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	0.5021	0.6129	0.7549	1.0054	1.5762	2.5342
e_{p_0}	0.1585	0.1335	0.1490	0.1112	0.0992	0.0551
c	1.0155	0.7984	0.6930	0.5899	0.5140	0.3716
e_c	0.1192	0.1109	0.1373	0.1188	0.1321	0.0930
r for QAOA (u3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	1.3023	2.0753	2.4677	2.7788	3.0333	3.1562
e_{p_0}	0.0694	0.0246	0.0304	0.0917	0.0800	0.0937
c	−0.5508	−1.0162	−1.1651	−1.2412	−1.2671	−1.2632
e_c	0.2069	0.0288	0.0252	0.0600	0.0439	0.0475
r for ab-QAOA (u3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	0.0481	0.0448	0.0541	0.0565	0.0474	0.0478
e_{p_0}	0.0058	0.0053	0.0080	0.0070	0.0052	0.0054
c	2.4151	2.5535	1.9384	1.8957	2.3950	2.3987
e_c	0.5815	0.5896	0.6777	0.5551	0.5354	0.5466
F for QAOA (u3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	1.2631	2.4719	3.3318	4.9315	5.8239	6.9639
e_{p_0}	0.1080	0.1576	0.2050	0.2623	0.3985	0.5160
c	1.1888	0.5537	0.4275	0.2897	0.2734	0.2436
e_c	0.3418	0.1302	0.0932	0.0545	0.0593	0.0537
F for ab-QAOA (u3r)	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$
p_0	0.0496	0.0442	0.0519	0.0574	0.0490	0.0499
e_{p_0}	0.0067	0.0060	0.0093	0.0088	0.0067	0.0072
c	4.1400	4.5860	4.1717	4.0884	4.6724	4.7042
e_c	0.6389	0.6792	0.8383	0.6831	0.6599	0.6857

APPENDIX C: NUMERICAL RESULTS FOR $N = 10, 14, 18$

This Appendix contains numerical results for both the w3r and u3r graphs with $n = 10, 14, 18$. The same fitting functions are used as those in the main text, and the fitting parameters are tabulated in Appendix D. The scaling analysis of these fitting functions will also be presented in Appendix D.

The results for $n = 18$ (brown triangles) in Figs. 12(c) and 12(d) are below those for $n = 14$ (purple triangles), as is also

observed in Figs. 2 and 4. From our results in Figs. 4(a) and 12(c) and inspection of the local fields, it appears that there are some special u3r graphs for which the ab-QAOA can find the solutions to the MaxCut problems using only a very shallow circuit depth. This creates the inversion with respect to n. In any case, one must keep in mind that the cost function is the energy, not the fidelity. A very low-lying excited state may have little overlap with the true ground state. We confirmed that the effect is mitigated by averaging over more graphs or by increasing R, the number of starting points, so we do not believe that there is anything very fundamental about it.

APPENDIX D: FITTING PARAMETERS

In Table I we list the fitting parameters defined in the main text, which are computed using the *scipy.optimize.curve_fit* function in python. The corresponding standard deviation errors e_{p_0} and e_c are also listed. On average, the fitting functions work better for the QAOA, where the fitting errors of p_0 for the u3r infidelity are a little bigger than those for the other cases. Overall, the fitting errors in ab-QAOA are bigger. However, what we care about most are the fits for the accuracy, which give the estimated p^* in the speedup, and the errors there are small.

We choose to fit all the points even though it might have been preferable to leave out the results in level 1 when fitting the results of Fourier strategy, since the R initial points are randomly generated instead of using the information from the last level. If we did leave the $p = 1$ points out, e_{p_0} and e_c in the QAOA would decrease slightly. For the ab-QAOA, e_{p_0} would decreased a small amount but e_c would increase significantly, and there would be a noticeable deviation between the points and the fitting curves. However, the main point about the fitting is the extrapolation of the QAOA data, so this does not affect any of our conclusions.

Using these fitting parameters and redefining the vertical axes of Figs. 2 and 4 of the main text, we can collapse the graphs for the accuracy and infidelity onto straight lines, as shown in Fig. 13.

TABLE II. p^* in speedup for w3r and u3r graphs.

w3r	$n = 8$	$n = 10$	$n = 12$	$n = 14$	$n = 16$	$n = 18$
standard QAOA	10	12	14	15	16	17
ab-QAOA	3	3	3	3	3	3
u3r	$n = 8$	$n = 10$	$n = 12$	$n = 14$	$n = 16$	$n = 18$
standard QAOA	5	7	8	9	10	11
ab-QAOA	3	3	3	3	3	3

APPENDIX E: SPEEDUP PARAMETER P^*

In Table II, we list p^* for the calculation of the speedup $S(n)$ shown in Fig. 5 in the main text. For the QAOA, p^* is obtained from the fitting function. For the ab-QAOA, p^* is obtained directly from the numerical simulation. For clarity of the speedup, all p^* are rounded to integers.

[1] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

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

[3] 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).

[4] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke *et al.*, Noisy intermediate-scale quantum algorithms, Rev. Mod. Phys. **94**, 015004 (2022).

[5] 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).

[6] D. Wang, O. Higgott, and S. Brierley, Accelerated Variational Quantum Eigensolver, Phys. Rev. Lett. **122**, 140504 (2019).

[7] K. M. Nakanishi, K. Mitarai, and K. Fujii, Subspace-search variational quantum eigensolver for excited states, Phys. Rev. Research **1**, 033062 (2019).

[8] J. Zeng, C. Cao, C. Zhang, P. Xu, and B. Zeng, A variational quantum algorithm for Hamiltonian diagonalization, Quantum Sci. Technol. **6**, 045009 (2021).

[9] C. Cao, Y. Yu, Z. Wu, N. Shannon, B. Zeng, and R. Joynt, Mitigating algorithmic errors in quantum optimization through energy extrapolation, arXiv:2109.08132.

[10] R. LaRose, A. Tikku, É. O'Neel-Judy, L. Cincio, and P. J. Coles, Variational quantum state diagonalization, npj Quantum Inf. **5**, 57 (2019).

[11] M. Cerezo, K. Sharma, A. Arrasmith, and P. J. Coles, Variational quantum state eigensolver, arXiv:2004.01372.

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

[13] D. Bondarenko and P. Feldmann, Quantum Autoencoders to Denoise Quantum Data, Phys. Rev. Lett. **124**, 130502 (2020).

[14] C. Cao and X. Wang, Noise-Assisted Quantum Autoencoder, Phys. Rev. Applied **15**, 054012 (2021).

[15] M. Cerezo, A. Poremba, L. Cincio, and P. J. Coles, Variational quantum fidelity estimation, Quantum **4**, 248 (2020).

[16] R. Chen, Z. Song, X. Zhao, and X. Wang, Variational quantum algorithms for trace distance and fidelity estimation, Quantum Sci. Technol. **7**, 015019 (2021).

[17] K. Sharma, S. Khatri, M. Cerezo, and P. J. Coles, Noise resilience of variational quantum compiling, New J. Phys. **22**, 043006 (2020).

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

[19] E. Farhi and A. W. Harrow, Quantum supremacy through the quantum approximate optimization algorithm, arXiv:1602.07674.

[20] 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).

[21] J. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. S. Fried, S. Hong *et al.*, Unsupervised machine learning on a hybrid quantum computer, arXiv:1712.05771.

[22] X. Qiang, X. Zhou, J. Wang, C. M. Wilkes, T. Loke, S. O'Gara, L. Kling, G. D. Marshall, R. Santagati, T. C. Ralph *et al.*, Large-scale silicon quantum photonics implementing arbitrary two-qubit processing, Nat. Photon. **12**, 534 (2018).

[23] G. Pagano, A. Bapat, P. Becker, K. S. Collins, A. De, P. W. Hess, H. B. Kaplan, A. Kyprianidis, W. L. Tan, C. Baldwin *et al.*, Quantum approximate optimization of the long-range Ising model with a trapped-ion quantum simulator, Proc. Natl. Acad. Sci. (USA) **117**, 25396 (2020).

[24] M. Willsch, D. Willsch, F. Jin, H. De Raedt, and K. Michielsen, Benchmarking the quantum approximate optimization algorithm, Quant. Inf. Proc. **19**, 197 (2020).

[25] D. M. Abrams, N. Didier, B. R. Johnson, M. P. da Silva, and C. A. Ryan, Implementation of XY entangling gates with a single calibrated pulse, Nat. Electron. **3**, 744 (2020).

[26] A. Bengtsson, P. Vikstål, C. Warren, M. Svensson, X. Gu, A. F. Kockum, P. Krantz, C. Križan, D. Shiri, I.-M. Svensson, G. Tancredi, G. Johansson, P. Delsing, G. Ferrini, and J. Bylander, Improved Success Probability with Greater Circuit Depth for the Quantum Approximate Optimization Algorithm, Phys. Rev. Applied **14**, 034010 (2020).

[27] G. Guerreschi and A. Matsuura, QAOA for Max-Cut requires hundreds of qubits for quantum speed-up, Sci. Rep. **9**, 6903 (2019).

[28] N. Moll *et al.*, Quantum optimization using variational algorithms on near-term quantum devices, Quantum Sci. Technol. **3**, 030503 (2018).

[29] E. Farhi, J. Goldstone, S. Gutmann, and H. Neven, Quantum algorithms for fixed qubit architectures, arXiv:1703.06199.

[30] Z.-C. Yang, A. Rahmani, A. Shabani, H. Neven, and C. Chamon, Optimizing Variational Quantum Algorithms Using Pontryagin's Minimum Principle, Phys. Rev. X **7**, 021027 (2017).

[31] L. T. Brady, C. L. Baldwin, A. Bapat, Y. Kharkov, and A. V. Gorshkov, Optimal Protocols in Quantum Annealing and Quantum Approximate Optimization Algorithm Problems, Phys. Rev. Lett. **126**, 070505 (2021).

[32] D. Wecker, M. B. Hastings, and M. Troyer, Training a quantum optimizer, Phys. Rev. A **94**, 022309 (2016).

[33] L. Li, M. Fan, M. Coram, P. Riley, and S. Leichenauer, Quantum optimization with a novel Gibbs objective function and ansatz architecture search, Phys. Rev. Research **2**, 023074 (2020).

[34] S. Bravyi, A. Kliesch, R. Koenig, and E. Tang, Obstacles to Variational Quantum Optimization from Symmetry Protection, Phys. Rev. Lett. **125**, 260505 (2020).

[35] 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).

[36] L. Zhu, H. L. Tang, G. S. Barron, F. Calderon-Vargas, N. J. Mayhall, E. Barnes, and S. E. Economou, An adaptive quantum approximate optimization algorithm for solving combinatorial problems on a quantum computer, arXiv:2005.10258.

[37] G. Crooks, Performance of the quantum approximate optimization algorithm on the maximum cut problem, arXiv:1811.08419 [quant-ph].

[38] S. Muthukrishnan, T. Albash, and D. Lidar, When diabatic trumps adiabatic in quantum optimization, arXiv:1505.01249.

[39] A. B. Ozguler, R. Joynt, and M. G. Vavilov, Steering random spin systems to speed up the quantum adiabatic algorithm, Phys. Rev. A **98**, 062311 (2018).

[40] H. R. Grimsley, S. E. Economou, E. Barnes, and N. J. Mayhall, An adaptive variational algorithm for exact molecular simulations on a quantum computer, Nat. Commun. **10**, 3007 (2019).

[41] H. L. Tang, V. O. Shkolnikov, G. S. Barron, H. R. Grimsley, N. J. Mayhall, E. Barnes, and S. E. Economou, Qubit-ADAPT-VQE: An adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor, PRX Quantum **2**, 020310 (2021).

[42] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019).

[43] T. Graß, Quantum Annealing with Longitudinal Bias Fields, Phys. Rev. Lett. **123**, 120501 (2019).

[44] A. Magann, K. Rudinger, M. Grace, and M. Sarovar, Feedback-based quantum optimization, arXiv:2103.08619.

[45] A. Magann, K. Rudinger, M. Grace, and M. Sarovar, Lyapunov control-inspired strategies for quantum combinatorial optimization, arXiv:2108.05945.

[46] K. Binder and A. P. Young, Spin glasses: Experimental facts, theoretical concepts, and open questions, Rev. Mod. Phys. **58**, 801 (1986).

[47] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).

[48] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, Proof verification and the hardness of approximation problems, J. ACM **45**, 501 (1998).

[49] J. Håstad, Some optimal inapproximability results, J. ACM **48**, 798 (2001).

[50] M. X. Goemans and D. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, J. ACM **42**, 1115 (1995).

[51] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[52] J. Wurtz and P. Love, MaxCut quantum approximate optimization algorithm performance guarantees for $p > 1$, Phys. Rev. A **103**, 042612 (2021).

[53] A. Lucas, Ising formulations of many NP problems, Front. Phys. **2**, 5 (2014).

[54] F. Glover, G. Kochenberger, and Y. Du, Quantum Bridge Analytics I: A tutorial on formulating and using QUBO models, 4OR **17**, 335 (2019).