Quantum optimization with a novel Gibbs objective function and ansatz architecture search

Li Li (李力),[1,*] Minjie Fan (范敏杰),[1] Marc Coram,[1] Patrick Riley,[1] and Stefan Leichenauer[2]
[1]*Google Research, Mountain View, California 94043, USA*
[2]*X, The Moonshot Factory, Mountain View, California 94043, USA*

(Received 18 October 2019; accepted 26 March 2020; published 24 April 2020)

The quantum approximate optimization algorithm (QAOA) is a standard method for combinatorial optimization with a gate-based quantum computer. The QAOA consists of a particular ansatz for the quantum circuit architecture, together with a prescription for choosing the variational parameters of the circuit. We propose modifications to both. First, we define the Gibbs objective function and show that it is superior to the energy expectation value for use as an objective function in tuning the variational parameters. Second, we describe an ansatz architecture search (AAS) algorithm for searching the discrete space of quantum circuit architectures near the QAOA to find a better ansatz. Applying these modifications for a complete graph Ising model results in a 244.7% median relative improvement in the probability of finding a low-energy state while using 33.3% fewer two-qubit gates. For Ising models on a 2d grid we similarly find 44.4% median improvement in the probability with a 20.8% reduction in the number of two-qubit gates. This opens a new research field of quantum circuit architecture design for quantum optimization algorithms.

DOI: 10.1103/PhysRevResearch.2.023074

I. INTRODUCTION

The quantum approximate optimization algorithm (QAOA) [1,2] is a general-purpose algorithm for finding a low-energy state of a given computational-basis Hamiltonian. This is a classical problem which can be combinatorially difficult, but using a quantum computer to find the solution might be more efficient than a classical method. The QAOA has performance guarantees in certain combinatorial problems [2] and quantum state transfer [3], and it has been shown that in general the output of the QAOA is not classically simulable [4]. The QAOA and related algorithms offer a promising avenue for near-term applications of quantum computers [5].

As emphasized in the original QAOA paper, the correct way to frame the goal of quantum optimization is in the probably approximately correct framework [6]. That is, the goal is to obtain a high likelihood of finding a nearly optimal solution. However, the standard objective function for QAOA does not reflect this goal. We introduce a new Gibbs objective function and show its superiority in the probably approximately correct sense. In numerical experiments for grid and complete graph Ising models, using the Gibbs objective function results in 10.8% and 8.6% median relative improvement of the probability of finding a low-energy state, respectively. We then proceed to try and find a superior circuit ansatz for the Gibbs objective function that is closely related to the general QAOA circuit through ansatz architecture search (AAS). Using AAS, the median relative improvement increased to 44.4% and 244.7% for the grid and complete graph models, together with a median reduction in the number of two-qubit gates by 20.8% and 33.3%, respectively. Figure 1 shows two exemplary instances and the improvement of probability of low energy with ansatzes found by AAS with the Gibbs objective function. The existence of these superior circuits opens a new field of research to design a search procedure for optimal problem-specific circuits.

II. ISING MODELS

A model $\mathcal{I}$ is defined on a graph $\mathcal{G}^{\mathcal{I}}$ with n vertices $\boldsymbol{v} \in \{1, 2, \ldots, n\}$ and a set of undirected edges $\mathcal{E} = \{\boldsymbol{e}_{ij}\}$. We select a 4×4 grid and complete graph with 10 vertices to cover the extreme cases of sparse and dense graphs.

Grid. A 4×4 square lattice. Edges only exist between nearest-neighbor vertices. This graph contains 16 vertices and $|\mathcal{E}| = 24$ edges. The average degree of the vertex in this graph is 3.

Complete graph. A complete graph with 10 vertices. Edges exist between any pair of vertices. This graph contains $|\mathcal{E}| = 45$ edges. The degree of each vertex in this graph is 9.

Each instance consists of a set of couplings $\boldsymbol{J}$ sampled independently from a uniform distribution $J_{ij} \sim U(-1, 1)$. A coupling J_{ij} is assigned to each undirected edge $\boldsymbol{e}_{ij}$ between vertices i and j. The Hamiltonian is written as a sum over edges, $E = \sum_{\boldsymbol{e}_{ij}} J_{ij} Z_i Z_j$. We denote a problem instance as $\mathcal{I} = \mathcal{I}(\mathcal{G}^{\mathcal{I}}, \boldsymbol{J})$. In Fig. 2 we plot histograms of the exact ground state energies per vertex for these instances.

The QAOA specifies a particular quantum circuit architecture which depends on the Hamiltonian. The prescription is very similar to a discretized adiabatic algorithm. The quantum state produced by the QAOA at level p is

$$|\psi\rangle = e^{i\beta_p X} e^{i\gamma_p E} \cdots e^{i\beta_1 X} e^{i\gamma_1 E} H^{\otimes n} |0^n\rangle. \quad (1)$$

*leeley@google.com

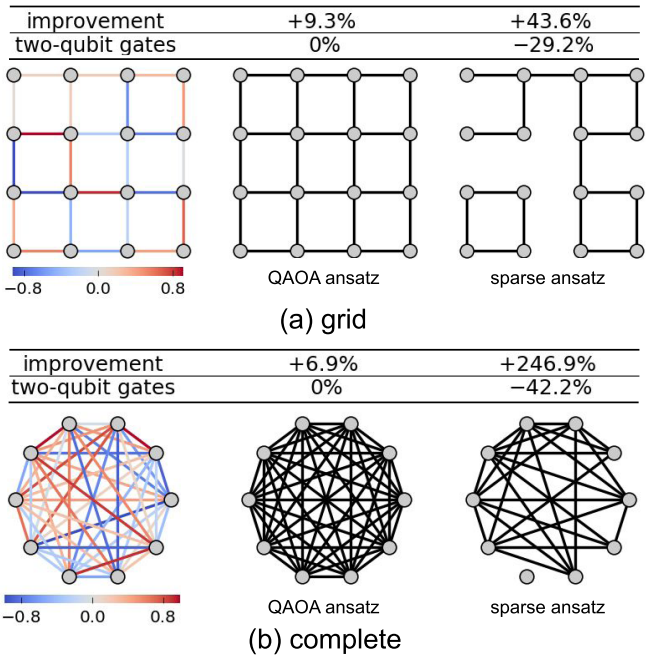

FIG. 1. Particular instances of random couplings (left) and the structures of the associated QAOA ansatzes (middle) and best sparse ansatzes (right) for (a) grid and (b) complete graph problems with the Gibbs objective function. On the left, each edge in the instance graph is colored by its coupling from blue (−1) to red (1). At the middle and on the right, each edge denotes the existence of a two-qubit gate on the corresponding edge in the ansatz graph. We show the relative improvement of the probability of low energy and reduction of the number of two-qubit gates compared to the usual prescription of the QAOA.

The $2p$ parameters $\vec{\beta}$ and $\vec{\gamma}$ are variational parameters of the model. For the Ising models in this paper, only two-qubit gates are required to construct the circuit for $e^{i\gamma E}$. We will also focus on $p = 1$ for simplicity.

Success in approximate optimization is measured according to the *probability of finding a low-energy state*. With that in mind, we evaluate the performance of our quantum circuits according to $P(E < E_0)$, where P is the Born probability distribution of the output quantum state $|\psi\rangle$ and E_0 is the cutoff for what we consider low energy. For definiteness, in this paper we use $E_0 = 0.95E_{\rm gs}(\mathcal{I})$ as our definition, where $E_{\rm gs}(\mathcal{I})$ is the exact ground state energy of the given instance $\mathcal{I}$ (which is always negative for the models we consider).

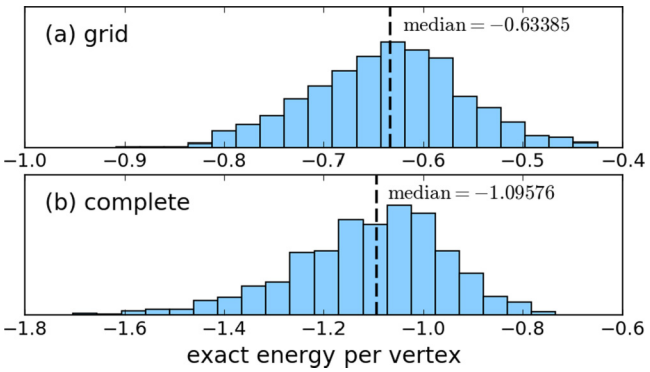

FIG. 2. Exact ground state energies per vertex of (a) grid and (b) complete graph instances. Black dashed lines indicate the medians of the exact energies per vertex.

III. GIBBS OBJECTIVE FUNCTION

A. Theory

We first address the problem of choosing the optimal values of the variational parameters. The standard prescription of minimizing the expectation value of the energy, $\langle E\rangle$, is just a proxy for maximizing $P(E < E_0)$. Recent work [7] has explored using conditional value-at-risk (CVaR) as the objective function. As an alternative, we propose minimizing the *Gibbs objective function*, defined as follows:

$$f = -\ln\langle e^{-\eta E}\rangle. \tag{2}$$

Here $\eta > 0$ is a hyperparameter based on the general properties of the class of problems. The function f is very similar to the Gibbs free energy from statistical mechanics, which is the origin of the name.

The reason why $\langle e^{-\eta E}\rangle$ might be preferred over $\langle E\rangle$ is easily understood intuitively. The exponential profile rewards us for increasing the probability of low energy, and deemphasizes the shape of the probability distribution at higher energies. Note that the Gibbs objective function is just as easy to measure as the energy expectation value itself when the energy is diagonal in the computational basis: we just perform a different computation with our measurement samples.

The Gibbs objective function is essentially the cumulant generating function of the energy [8]. The Taylor expansion reads $f(\eta) = \mu_E\eta - \sigma_E^2\eta^2/2 + \kappa_3\eta^3/6 + \cdots$. For small η, then, minimizing the Gibbs objective function is equivalent to minimizing $\mu_E = \langle E\rangle$. As η increases, the higher-order cumulants become more important.

To better understand the Gibbs objective function, we can try to estimate the best value of the hyperparameter η. For any $\eta > 0$ the probability of low energy is bounded from above as follows:

$$\begin{aligned} P(E < E_0) &= \langle 1_{E<E_0}\rangle \\ &\leqslant \left\langle 1_{E<E_0} e^{-\eta(E-E_0)}\right\rangle \\ &\leqslant \left\langle e^{-\eta(E-E_0)}\right\rangle. \end{aligned} \tag{3}$$

Choosing η to minimize the right-hand side gives the strongest inequality out of this one-parameter family. That value of η is the one which satisfies the equation

$$E_0 = \frac{\langle E e^{-\eta E}\rangle}{\langle e^{-\eta E}\rangle}. \tag{4}$$

Now, η is meant to be a fixed hyperparameter that is maintained throughout parameter optimization, whereas the η satisfying this equation depends functionally on the probability distribution itself. Our prescription for estimating η is to find an approximate solution to this equation, valid for a large class of probability distributions that we may encounter during parameter optimization.

If E_0 is meant to be close to $E_{\rm gs}$, then it is clear that the interesting limit of Eq. (4) is the large-η limit.[1] The first correction at large η to the right-hand side is equal to η^{-1}:

$$\frac{\langle E e^{-\eta E}\rangle}{\langle e^{-\eta E}\rangle} \approx E_{\rm gs} + \eta^{-1}.$$

Combined with Eq. (4), this suggests that we should set $\eta = (E_0 - E_{\rm gs})^{-1}$. We may only be able to estimate values for E_0 and $E_{\rm gs}$ based on the specification of our problem, but in practice these estimates are good enough. For the problems we consider, $E_0 = 0.95E_{\rm gs}$ and $E_{\rm gs} \approx -1$ gives $\eta \approx 20$ as an estimate, which we use for the majority of our numerical experiments below.

Note that in the large-η/small-$(E_0 - E_{\rm gs})$ regime one can make much stronger statements about the relationship between $\langle e^{-\eta E}\rangle$ and $P(E < E_0)$. We will sketch some of them here. In taking the large-η limit above, we effectively approximated the probability density function for the energy, $p(E)$, by its constant term $p(E_{\rm gs})$ near the ground state energy. If $p(E)$ is treated as a constant, then $P(E < E_0)$ and $\langle e^{-\eta(E-E_{\rm gs})}\rangle$ are actually *equal* when $\eta = (E_0 - E_{\rm gs})^{-1}$. More generally, if $p(E)$ is well approximated by a finite-degree polynomial in $E - E_{\rm gs}$ with bounded coefficients, then we have the slightly weaker condition $P(E < E_0) \sim \langle e^{-\eta(E-E_{\rm gs})}\rangle$, meaning that either quantity is bounded from above and below by constant multiples of the other. This further motivates the use of the Gibbs objective function.

B. Numerical experiments

To evaluate the performance of the Gibbs objective function, as well as the ansatz search described later, we analyze 1000 instances each of the grid and complete graph Ising models. For each instance we optimize the variational parameters β and γ using the Nelder-Mead algorithm [9] to minimize either the expectation value of the energy or the Gibbs objective function. The underlying circuit ansatz is either the QAOA or an optimized sparse ansatz as described in the next section. In all cases we evaluate the algorithm performance according to the probability of finding a low-energy state, $P(E < 0.95E_{\rm gs}(\mathcal{I}))$. As discussed above, the quantum optimization algorithm is not designed to find the exact ground state, and we chose the metric as 5% around the ground state energy so the task of finding low energy is hard but not impossible. The exact value chosen here is not important.

In Fig. 3 we show the effect of changing the hyperparameter η in the Gibbs objective function using the QAOA circuit ansatz. When η is small, the Gibbs objective function is equivalent to the energy expectation value as an objective function. Therefore the Gibbs objective function cannot perform worse than the energy expectation value when η is properly tuned. Numerically, we find that the probability of low energy increases monotonically with η before plateauing at large values of η. Our estimated value $\eta = 20$ falls within the convergence range for the problems we consider, so we use that value throughout. Finally, we also observe that in the extreme-η regime, e.g., 10^5, the parameter optimization does not converge due to the fact that the objective function is approximately zero except when the exact ground state is sampled. This is an obstacle to efficient optimization at those extreme values.

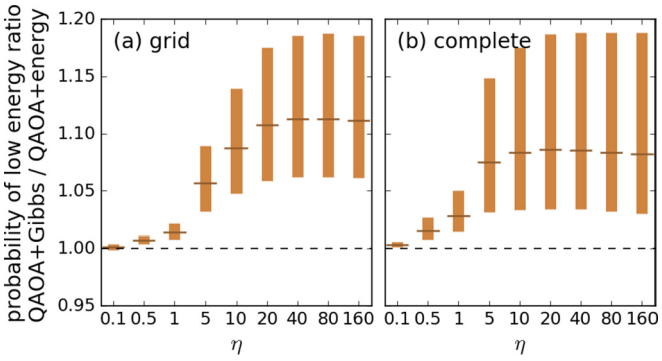

FIG. 3. Comparison of Gibbs objective function with different η to the energy expectation objective function on QAOA ansatz. For every instance given η, we measure the probability of low energy of QAOA + Gibbs divided by QAOA + energy. The bars show the range from 5% to 95% and the horizontal segments are median. For small values of η, the Gibbs objective function is equivalent to the energy expectation value for purposes of optimization, while for large values of η it is equivalent to maximizing the probability of finding the ground state.

Figure 4 displays the probability of finding a low-energy state for each quantum optimization algorithm, denoted as {ansatz type} + {objective}. QAOA + energy is the original QAOA prescription and provides the baseline for comparison. The sparse ansatz is the subject of the next section. As shown in the scatter plots of QAOA + Gibbs vs QAOA + energy, using the Gibbs objective function improves the solution. More significant improvement can be achieved using a sparse ansatz in addition to the Gibbs objective function, especially for complete graph instances.

IV. OPTIMIZING THE ANSATZ

Next we discuss alternatives to the QAOA circuit ansatz. For the Ising Hamiltonians, the operator $e^{i\gamma E}$ of Eq. (1) involves a two-qubit operator for each edge in the instance graph $\mathcal{G}^{\mathcal{I}}$. We denote by $\mathcal{G}^{\mathcal{A}}$ the *ansatz graph*, which is obtained from $\mathcal{G}^{\mathcal{I}}$ by removing some edges. The associated circuit ansatz $\mathcal{A}$ is obtained by removing from $e^{i\gamma E}$ those two-qubit operators corresponding to the edges which were removed from $\mathcal{G}^{\mathcal{I}}$. The rest of the quantum circuit remains the same as in the QAOA. This is clearly not the most general possible prescription for $\mathcal{G}^{\mathcal{A}}$, but makes use of the intuition that the QAOA ansatz $\mathcal{G}^{\mathcal{A}} = \mathcal{G}^{\mathcal{I}}$ is a good starting point for the architecture search.

In total, an ansatz $\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \beta, \gamma)$ is determined by its graph architecture $\mathcal{G}^{\mathcal{A}}$ and continuous parameters β, γ. The optimal ansatz graph and variational parameters for a given instance are denoted by $\hat{\mathcal{G}}^{\mathcal{A}}, \hat{\beta}, \hat{\gamma}$, and they are the ones that minimize the objective function:

$$\hat{\mathcal{G}}^{\mathcal{A}}, \hat{\beta}, \hat{\gamma} = \underset{\mathcal{G}^{\mathcal{A}},\beta,\gamma}{\arg\min}\, f(\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \beta, \gamma), \mathcal{I}). \tag{5}$$

[1]We are assuming that η is large compared to the inverse of the energy scale of the Hamiltonian, but not large compared to the gap. In other words, even when η is large there should still be many states between $E_{\rm gs}$ and $E_{\rm gs} + \eta^{-1}$.

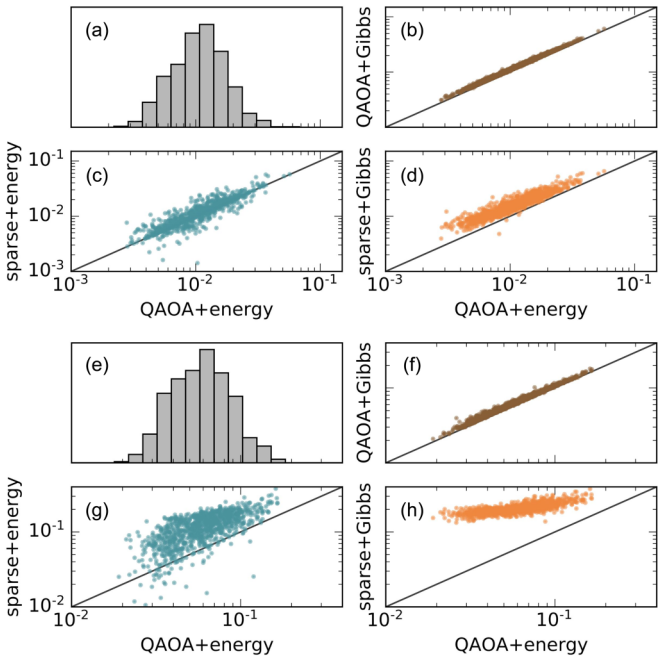

FIG. 4. Comparison of the objective functions and ansatzes on 1000 grid [(a)–(d)] and complete [(e)–(h)] graph instances. The histograms show the distributions of probability of low energy for QAOA + energy. The scatter plots compare the probability of low energy for {ansatz} + {objective} pairs against the QAOA + energy baseline.

For each $\mathcal{G}^{\mathcal{A}}$, $\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \beta, \gamma)$ represents a family of ansatzes differing by β, γ. We can optimize Eq. (5) in a nested manner,

$$\hat{\mathcal{G}}^{\mathcal{A}} = \arg\min_{\mathcal{G}^{\mathcal{A}}} f(\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \hat{\beta}, \hat{\gamma}), \mathcal{I}) \tag{6}$$

$$\text{with } \hat{\beta}, \hat{\gamma} = \arg\min_{\beta,\gamma} f(\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \beta, \gamma), \mathcal{I}). \tag{7}$$

The outer step [Eq. (6)] searches the space of the architectures $\{\mathcal{G}^{\mathcal{A}}\}$. For a fixed architecture $\mathcal{G}^{\mathcal{A}}$, the inner step [Eq. (7)] returns the optimal ansatz $\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \hat{\beta}, \hat{\gamma})$ in the family of $\mathcal{A}(\mathcal{G}^{\mathcal{A}}, \beta, \gamma)$. We should understand that $\hat{\beta}$ and $\hat{\gamma}$ are implicit functions of the ansatz graph $\mathcal{G}^{\mathcal{A}}$ through Eq. (7). We denote the outer step as AAS and the inner step as parameter optimization.

A. Ansatz architecture search

A good design of the search space is essential in discrete structure optimization problems, e.g., neural architecture search [10–12], molecule optimization [13], composite design [14], and symbolic regression [15,16]. Since the QAOA is a well-recognized ansatz for combinatorial problems, we have designed the search space for $\mathcal{G}^{\mathcal{A}}$ based on gradual modifications of the QAOA ansatz. The QAOA prescription is to take $\mathcal{G}^{\mathcal{A}} = \mathcal{G}^{\mathcal{I}}$, and our search through architectures is a search through graphs obtained by removing edges from $\mathcal{G}^{\mathcal{I}}$.

Denote by $\mathcal{G}_k$ a graph containing k edges. If m is the number of edges in $\mathcal{G}^{\mathcal{I}}$, then there is only one $\mathcal{G}_m$ in our search space, namely $\mathcal{G}^{\mathcal{I}}$ itself. Thus we say $|\{\mathcal{G}_m\}| = 1$. If we remove up to n edges from the graph, then the total search space is

$$\bigcup_{l=0}^{n} \{\mathcal{G}_{m-l}\} = \{\mathcal{G}_m\} \cup \{\mathcal{G}_{m-1}\} \cup \ldots \cup \{\mathcal{G}_{m-n}\}. \tag{8}$$

The size of this space is $\sum_{l=0}^{n} \binom{m}{l}$. A brute-force enumerative search is impractical since the size grows quickly as n increases. For example, considering a complete graph with 10 vertices, $\sum_{l=0}^{5} \binom{45}{l} \sim 1 \times 10^6$ and $\sum_{l=0}^{15} \binom{45}{l} \sim 6 \times 10^{11}$. We propose greedy search as an affordable strategy for AAS. Given an instance $\mathcal{I}$, the search starts with $\mathcal{G}^{\mathcal{A}} = \mathcal{G}_m$ at level 0. Then level by level, ansatzes are expanded by removing one two-qubit gate from the best ansatz of previous level, scored, and the best of them is selected as the output of this level. The output architectures at level l have l two-qubit gates (i.e., edges of the graph) removed. The total number of architectures visited in the greedy search is $\mathcal{N} \leqslant \sum_{l=0}^{n}(m - l) = (n + 1)(m - n/2)$.

B. Scoring an ansatz

1. Nelder-Mead

The score for each ansatz $\widetilde{\mathcal{G}}_{m-l}$ in AAS is obtained by specifying parameters β^* and γ^* and computing $f(\mathcal{A}(\widetilde{\mathcal{G}}_{m-l}, \beta^*, \gamma^*), \mathcal{I})$. One prescription is to let β^* and γ^* simply be the optimal values of β and γ minimizing the objective function, $\beta^*, \gamma^* = \arg\min_{\beta,\gamma} f(\mathcal{A}(\widetilde{\mathcal{G}}_{m-l}, \beta, \gamma), \mathcal{I})$. We use the Nelder-Mead algorithm [9] to perform this minimization. In other words, we take β^* and γ^* to be close approximations to the optimal values $\hat{\beta}$ and $\hat{\gamma}$ for the given ansatz graph. Nelder-Mead is a black-box optimization algorithm popular in the quantum variational circuit literature [17,18]. Using this algorithm requires running quantum circuit simulations at each iteration and reporting the objective function value to the optimizer until convergence. Thus it is extremely expensive in terms of calls to the (simulated) quantum computer. Since we want to limit the number of such calls, we are motivated to consider other strategies for finding β^*, γ^*.

2. Estimated β, γ

Rather than using an optimization algorithm such as Nelder-Mead to minimize the objective function and thereby obtain β^* and γ^*, we can use analytical estimates to approximate the parameters instead. This saves the computation time required to evaluate the quantum circuits for parameter optimization during scoring. The β^* and γ^* we find will not necessarily be close to the optimal values $\hat{\beta}$ and $\hat{\gamma}$, but the idea is that this may not be important for the purposes of scoring. We may still wish to use Nelder-Mead for evaluation of the final ansatz at the conclusion of the AAS.

The estimates for β^* and γ^* we use in this section come from making several simplifying assumptions that are not necessarily valid. The first assumption is that it is reasonable to use β^* and γ^* values obtained by minimizing the energy expectation value instead of the Gibbs objective function. Focusing on the grid Ising model, we can write down an exact formula for $\langle E \rangle$ in a $p = 1$ QAOA as follows:

$$\langle E \rangle = \sin 4\beta \sum_{\mathbf{e}_{ij}} \left[\prod_k \cos(2\gamma J_{kj}) \right] J_{ij} \tan(2\gamma J_{ij}). \tag{9}$$

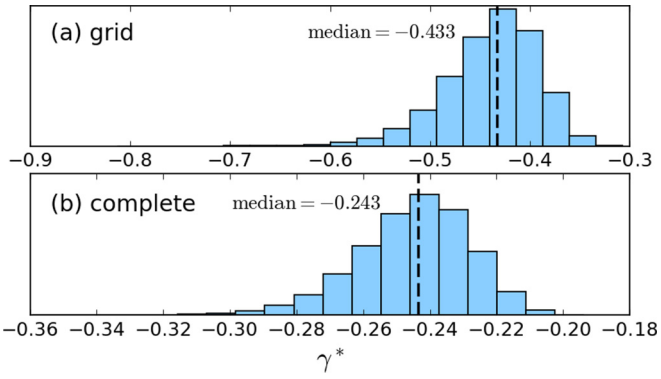

FIG. 5. Histogram of γ^* as determined by Eq. (10) for 10^5 independently drawn sets of couplings $\mathbf{J}$. Black dashed lines are the medians of the distributions.

To use this formula for an ansatz graph $\mathcal{G}^{\mathcal{A}}$ other than $\mathcal{G}^{\mathcal{I}}$ one just sets to 0 the J_{ij} associated to the missing edges.

Equation (9) determines $\beta^* = \pi/8$. To find a formula for γ^* we make another simplifying assumption, namely that $\gamma^* \ll 1$.[2] Expanding Eq. (9) to third order in γ and minimizing the resulting cubic polynomial gives

$$\gamma^* = -\sqrt{\frac{\sum_{ij} J_{ij}^2}{6\left(\sum_{ijk,j\neq k} J_{ki}^2 J_{ij}^2 + \frac{1}{3}\sum_{ij} J_{ij}^4\right)}}. \qquad (10)$$

All of this was in the context of grid instances, and in particular in deriving Eq. (9) we made use of the fact that two neighboring vertices in the graph do not have any neighbors in common. Nevertheless, as our final simplifying assumption we will insist on using Eq. (10) for the complete graph as well. We will see from the numerical experiments that these simplifying assumptions are good enough for scoring.

3. Fixed β, γ

The estimated β^* above is already instance-independent, and the estimated γ^* performed well despite the approximations involved not being fully justified. This suggests that the precise value of γ^* used in scoring is not crucially important. This is similar to the observation in Ref. [19] that the behavior of the QAOA tends to concentrate across instances. This motivates the third scoring prescription, the "fixed-parameter" prescription.

We generated the distribution of estimated γ^* values according to the formula Eq. (10) for both the grid and complete graph models by looking at 10^5 choices of couplings drawn independently from the uniform distribution, $\boldsymbol{J} \sim U(-1, 1)$, for each model with $\mathcal{G}^{\mathcal{A}} = \mathcal{G}^{\mathcal{I}}$. The associated histograms are shown in Fig. 5. The "fixed-parameter" prescription for γ^* is defined by using the medians of these distributions for all instances of the associated model. These values are listed in Fig. 5.

[2]Rather than finding an explicit formula, one could also choose to minimize Eq. (9) numerically to find γ^*. This does not affect the results.

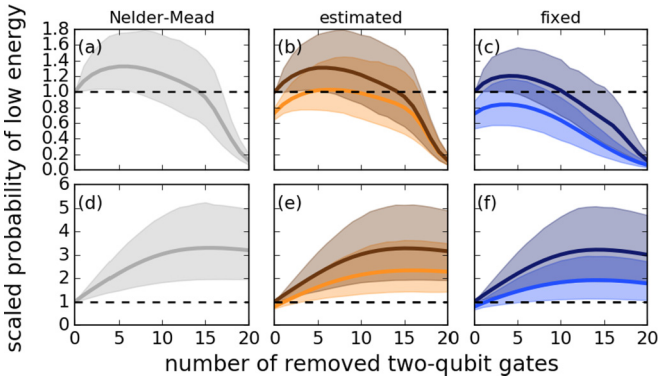

FIG. 6. Comparison of different prescriptions for the scoring function of AAS. The solid curves are the scaled probability of low energy of the best ansatz found through greedy search at each of the first 20 levels. The shadows show the range from 5% to 95%. Panels (a)–(c) are results from 1000 grid instances and panels (d)–(f) are results from 1000 complete graph instances. The scoring prescriptions are (a), (d) Nelder-Mead; (b), (e) estimated parameters; and (c), (f) fixed parameters. The dark orange and blue curves are the scaled probability of the best ansatz graph after using Nelder-Mead to conduct a final optimization of the parameters after AAS, while the light orange [(b), (e)] and light blue [(c), (f)] curves represent the scaled probability obtained for the best ansatz graph without the final parameter optimization step.

C. Numerical experiments

We apply the AAS procedure with the three prescriptions to choose the scoring parameters β^* and γ^* on grid and complete graph Ising models. The performance of an ansatz $\mathcal{A}$ produced by AAS is measured by the scaled probability of low energy,

$$\widetilde{P}(\mathcal{A}, \mathcal{I}) = P_{\mathcal{A}}(E < 0.95E_{\rm gs})/P_{\mathcal{I}}(E < 0.95E_{\rm gs}). \qquad (11)$$

The probability in the numerator is the one associated with the ansatz graph $\mathcal{G}^{\mathcal{A}}$ with parameters equal to their optimal values $\hat{\beta}$, $\hat{\gamma}$ obtained by minimizing the Gibbs objective function for that ansatz. The probability in the denominator is similar, except using the instance graph $\mathcal{G}^{\mathcal{I}}$ as the ansatz. In other words, the prescription for computing the denominator probability is similar to the standard QAOA, except that the parameters are optimized using the Gibbs objective function rather than the energy expectation value. We chose to use the Gibbs objective function for both numerator and denominator in order to isolate the effects of the AAS. The optimization is done using Nelder-Mead for both.

Figure 6 shows the scaled probabilities of low energy of the optimal ansatz at each level for both grid and complete graph instances. Each column corresponds to a different prescription for the scoring function of AAS.

We first discuss the results of grid instances. In (a), the scoring is done using parameters that are optimized by Nelder-Mead. The scaled probabilities of low energy increase as more two-qubit gates are removed. But they start to decrease when more than 5 two-qubit gates are removed. In (b) and (c) the scoring was performed according to the estimated parameter prescription and fixed parameter prescription, respectively. The dark curves in each case represent the performance of the

final circuit found by AAS using the optimal $\hat{\beta}$, $\hat{\gamma}$ obtained by minimizing the Gibbs objective function. We see that there is not a strong dependence on the scoring prescription, though the "fixed" procedure is slightly worse. However, the light curves in (b) and (c) represent the performance of those same output circuits if, rather than using $\hat{\beta}$ and $\hat{\gamma}$, we use the β^* and γ^* values used in the scoring step of AAS. Then we see that there is a significant decrease in performance, especially for the fixed method in (c). The lesson here is that for scoring in AAS, which only cares about relative performance for ranking, the circuit parameter values are less important. In fact, good relative performance from these two prescriptions suggests that it is possible to construct inexpensive heuristic functions for scoring without calls to the quantum computer. We explore this further in Appendix D. On the other hand, it is crucial to get the parameters right when considering absolute performance.

The trend for complete graphs in (d)–(f) is very similar. The main qualitative change is that the performance does not drop off as steeply as a function of the number of removed gates. This is easily understood from the fact that the complete graphs have far more edges than the grid (45 vs 24). We also see that the "fixed" procedure is closer in performance to the others for complete graphs.

V. CONCLUSION

We have proposed using the Gibbs objective function and AAS as two improvements to the QAOA. There are several potential follow-ups and opportunities for further developments:

The Gibbs objective function may be useful more broadly for quantum optimization problems, such as variational approaches to molecular ground states [17,20]. In those cases, where the energy is not diagonal in the computational basis, it will be more challenging to evaluate $\langle \exp(-\eta E)\rangle$ by sampling, but may still be worthwhile.

Even within combinatorial optimization, AAS is costly because each quantum circuit must be simulated (or run on a real quantum computer) during the scoring step. Performance improvements could be offset by this extra cost. That is the motivation for the alternative heuristic methods we explore in Appendix D, and it remains an open problem to find an effective heuristic. Our estimated parameter and fixed parameter prescriptions for scoring show that it is possible to capture relative performance without reproducing the absolute performance. This leaves open the possibility that a good heuristic scoring function exists.

In this paper, we computed probabilities and expectation values directly from the wave function. On a real quantum computer this is impossible. Instead, one estimates expectation values based on a finite number of samples. The number of samples is another hyperparameter, and it directly affects the cost of running the algorithm on a quantum computer. An open question is whether the scoring in AAS can work with a very small number of samples, mitigating the cost. Finally, one may want to include other effects in the scoring, e.g., the fidelity of the two-qubit gates in the circuit, and search for the Pareto optimal [21] for multiobjective optimization.[3]

ACKNOWLEDGMENTS

The authors thank Steven Kearnes for code review and discussions, Zan Armstrong and Nathan Neibauer for suggestions and help in data visualization, and Dave Bacon, Edward Farhi, Murphy Yuezhen Niu, Thomas Fösel, and John Platt for their review and comments. X, formerly known as Google[x], is part of the Alphabet family of companies, which includes Google, Verily, Waymo, and others [22]. Quantum simulation and AAS in this paper were implemented using Cirq [23] and Apache Beam [24].

APPENDIX A: GREEDY SEARCH

The following three steps are performed at level l in the search:

(1) *Expansion*. Generate all the unique $\{\widetilde{\mathcal{G}}_{m-l}\}$ by removing one two-qubit gate from the output of the previous level.

(ii) *Scoring*. Evaluate a scoring function $\mathcal{S}$ on each of the architectures $\{\widetilde{\mathcal{G}}_{m-l}\}$ generated by the previous step. Ideally, the scoring function would exactly match the final target function. However, that can be expensive to compute so we will examine alternative scoring functions below. In particular, we will consider different methods for specifying variational parameters β^*, γ^* for each circuit and then evaluating the Gibbs objective function by simulation using those parameters:

$$\{\mathcal{S}(\widetilde{\mathcal{G}}_{m-l}, \mathcal{I})\} = \{f(\mathcal{A}(\widetilde{\mathcal{G}}_{m-l}, \beta^*, \gamma^*), \mathcal{I})\}. \tag{A1}$$

(iii) *Selection*. Select the architecture with the best score as the output of this level.

APPENDIX B: INITIAL VALUES IN NELDER-MEAD

The initial values of β and γ are sampled independently from the uniform distribution $U(0, 0.1)$.

APPENDIX C: BEAM SEARCH

We introduce beam search, a generalized search algorithm of greedy search. It has been used in combinatorial optimization [25], program synthesis [26], and machine translation [27]. Beam search differs from greedy search in the selection step:

Selection. Select w architectures with the best scores, where the integer w is called the *beamwidth*. These best-performing architectures are the output of this level. At early stages in the beam search we may have fewer than w candidates available, in which case all candidates are returned.

At the lth level, $|\{\widetilde{\mathcal{G}}_{m-l}\}| \leqslant w \times (m-l)$. The total number of architectures visited in the beam search is $\mathcal{N} \leqslant w \sum_{l=0}^{n}(m-l) = w(n+1)(m-n/2)$. As special cases, we recover enumerative search as $w \to \infty$ and greedy search as $w = 1$.

[3] We thank Edward Farhi for discussion of this point.

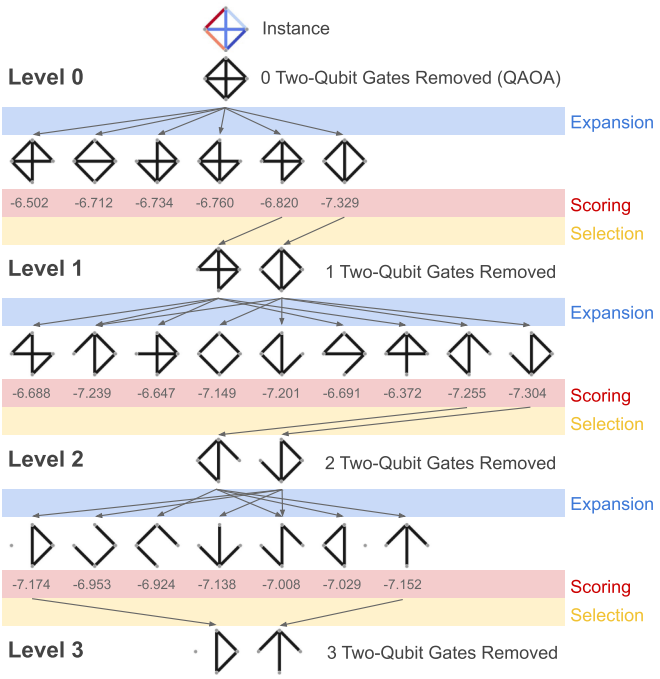

FIG. 7. Illustration of ansatz architecture search (AAS) by removing up to 3 two-qubit gates with beamwidth $w = 2$.

Figure 7 illustrates the procedure of AAS for a complete graph with 4 vertices.

From numerical experiments, we found that there was not much improvement in performance from increasing the beamwidth $w \geqslant 1$.

APPENDIX D: CHALLENGE: SEARCH WITHOUT QUANTUM SIMULATION

We demonstrated that with AAS and parameter optimization, a circuit ansatz that significantly improves the probability of low energy can be found. However, all of the methods in Fig. 6 made use of quantum circuit simulation at each stage in the search. While we were able to show that the method with the most quantum circuit simulation (Nelder-Mead) does not improve significantly on cheaper scoring methods, all the methods require some quantum circuit simulation at each level. In this section, we investigate some heuristic functions for replacement of quantum simulation for the purpose of the scoring step of AAS. Our results are mixed, and fully solving this problem remains an open challenge for the community.

Random. For each ansatz in the scoring step, we assign a random number to replace $f(\mathcal{A}, \mathcal{I})$ in Eq. (A1). This baseline does not use any information from the ansatz architecture and problem instance, and amounts to removing edges from the graph randomly during AAS.

Energy approximation. Our second heuristic uses the estimated energy expectation value as the scoring function. That is, we plug $\beta^* = \pi/8$ and γ^* as given by Eq. (10) into Eq. (9) and use that as the score.[4]

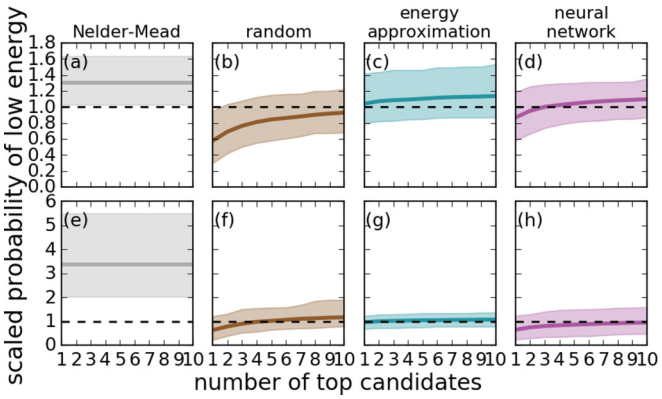

FIG. 8. Comparison of different heuristics for the scoring function of AAS. Panels (a)–(d) search sparse ansatzes by removing exactly 5 two-qubit gates on 200 grid instances, and panels (e)–(h) search sparse ansatzes by removing exactly 15 two-qubit gates on 200 complete graph instances. Panels (a) and (e) use the Nelder-Mead scoring function at each level in greedy search and serve as the baseline for measuring performance. The remaining prescriptions, explained in detail in Appendix D, are (b), (f) random; (c), (g) energy approximation; and (d), (h) neural networks. For each of these three prescriptions we use beamwidth $w = 100$. In all cases, Nelder-Mead is used at the end to optimize the parameters of the top candidates from the AAS (the number of which varies along the horizontal axis), and the candidate with the lowest Gibbs objective function value is selected for the plot. The solid lines are the mean performance across instances, the dashed lines are at 1, and the shadows show the range from 5% to 95%.

Neural network. We use a neural network to approximate $f(\mathcal{A}, \mathcal{I})$ in Eq. (A1). It contains two dense layers with 128 hidden units and ReLU activation functions. The instance is represented by 2nd and 4th powers of couplings on edges. The ansatz graph $\mathcal{G}^{\mathcal{A}}$ is represented by Booleans indicating whether a two-qubit gate is placed on an edge of the instance. We concatenate these features as input of the network. We take all the ansatzes generated by AAS and Nelder-Mead and split them randomly by their instances into a training set and test set. For grid instances, the training set contains 800 instances with 232 800 ansatzes. For complete graph instances, the training set contains 800 instances with 568 800 ansatzes. Both test sets contain 200 instances not seen in the training set. To fix normalization we use the scaled objective function value $f(\mathcal{A}, \mathcal{I})/f(\text{QAOA}, \mathcal{I})$ as the label.

Although a training set is not required for the random and energy approximation heuristics, for a fair comparison we restrict each heuristic to the same 200 test instances for each instance type in Fig. 8. We find optimal sparse ansatzes by removing 5 two-qubit gates for grid instances and 15 two-qubit gates for complete graph instances. Panels (a) and (e) show the results of AAS and Nelder-Mead. They are the best ansatzes we can find for the test instances. Since random, energy approximation, and neural network score functions are inexpensive to compute compared to quantum simulation, we

[4]Really, we first expanded Eq. (9) to third order in γ before plugging in γ^*. This is for consistency, but does not make a large difference. We also experimented with numerically minimizing Eq. (9) rather than using any estimates, and this, too, does not make much difference.

have given these an advantage in the search by setting the beamwidth w to 100. At the end of AAS, we sort the ansatzes from the last level by their objective function values. Then we run quantum simulations for the top candidates and report the best scaled probability of low energy. As more candidates are taken into consideration, the performance of all three heuristic functions improves but the cost of quantum simulation for evaluation also increases. The number of top candidates chosen in each case is listed along the horizontal axis in the plots. Note that a reasonable fraction of cases produce a scaled probability less than 1, indicating that one would be better off just using the original circuit. For grid instances, random (b) is the worst. Energy approximation (c) performs better than neural network (d) and is comparable (though still inferior) to simulation (a). However, for complete graph, none of the heuristic functions is comparable to simulation (e).

APPENDIX E: THE EFFECT OF NOISE

The purpose of this section is to analyze the effects of a simple noise model on the Gibbs objective function of Eq. (2). In the absence of noise, the ideal Gibbs objective function is given by

$$f_{\text{ideal}} = -\ln\langle e^{-\eta E}\rangle_\psi,$$

where E is the Hamiltonian we are optimizing and the angled brackets represent the quantum expectation value in the output state ψ of the quantum circuit ansatz.

The noise model we consider is a simple depolarizing channel. With probability $1-p$ the quantum circuit is executed perfectly and the output state is the one we expect. With probability p, there is some error in the execution and the output state is the maximally mixed one. In other words, with probability p we sample from the uniform distribution on bit strings instead of the desired Born distribution. In the language of density operators, we can say that the effective density operator describing the quantum state is

$$(1-p)|\psi\rangle\langle\psi| + \frac{p}{2^n}I,$$

where n is the number of qubits.

Using this error model, we can ask what the noise does to the Gibbs objective function f. We simply replace the expectation value in the ideal state ψ with an expectation value in the noisy state $(1-p)|\psi\rangle\langle\psi| + pI/2^n$. Equivalently, we can take a weighted average of the $\langle\cdot\rangle_\psi$ expectation value with an expectation value according the uniform distribution over bit strings. We find

$$\begin{aligned} f_{\text{noisy}} &= -\ln\left[(1-p)\langle e^{-\eta E}\rangle_\psi + \frac{p}{2^n}\operatorname{Tr} e^{-\eta E}\right] \\ &= f_{\text{ideal}} - \ln\left[1 - p\frac{\langle e^{-\eta E}\rangle_\psi - \operatorname{Tr} e^{-\eta E}/2^n}{\langle e^{-\eta E}\rangle_\psi}\right]. \end{aligned}$$

Note that one expects $\langle e^{-\eta E}\rangle_\psi \geqslant \operatorname{Tr} e^{-\eta E}/2^n$ if the circuit is properly trained, and so the correction makes the objective function larger (worse), as it should. We also have the following bound on the change in the objective function, coming from the positivity of $e^{-\eta E}$:

$$f_{\text{noisy}} - f_{\text{ideal}} \leqslant -\ln(1-p). \tag{E1}$$

For small p the right-hand side is just $\approx p$. It is reasonable to expect that $\langle e^{-\eta E}\rangle_\psi \gg \operatorname{Tr} e^{-\eta E}/2^n$—in other words, the trained ansatz should have a much better Gibbs objective function value than the uniform distribution over bit strings—which means that the bound in Eq. (E1) will be approximately saturated.

This means that we can directly translate improvements to the objective function into resilience against depolarizing noise. An improvement of size Δf in the objective function can counteract the effect of depolarizing noise with size $p \approx \Delta f$ (assuming $p \ll 1$).

It is also worth noting the effects of noise on the probability of finding a low-energy bit string, $P(E < E_0)$. Using the same depolarizing noise model,

$$P_{\text{noisy}} = P_{\text{ideal}} - p(P_{\text{ideal}} - P_{\text{uni}}).$$

Here P_{uni} is just the probability for success by random guessing using the uniform distribution on bit strings. Then, taking logarithms, we find

$$\ln P_{\text{noisy}} = \ln P_{\text{ideal}} + \ln\left(1 - p\frac{P_{\text{ideal}} - P_{\text{uni}}}{P_{\text{ideal}}}\right).$$

This is very similar to what we saw for the behavior of the Gibbs objective function. This is not a coincidence: part of the reason why the Gibbs objective function was chosen is that the operator $e^{-\eta E}$ for appropriate values of η behaves very similarly to the projection operator one would use to define P. For large η and E_0 close to E_{gs}, P and $\langle e^{-\eta E}\rangle_\psi$ become equal up to a state-independent multiplicative factor.

TABLE I. Relative improvement of the probability of low energy [Eq. (F1)] and reduction of the number of two-qubit gates [Eq. (F2)] compared to the usual prescription of the QAOA for 1000 grid instances and 1000 complete graph instances. The values of 5th percentile, median, and 95th percentile are reported for two instance types.

		Relative Probability of Low Energy [Eq. (F1)]			Relative Number of Two-Qubit Gates [Eq. (F2)]		
Instance Type	Prescription	5th percentile (%)	Median (%)	95th percentile (%)	5th percentile (%)	Median (%)	95th percentile (%)
Grid	QAOA + Gibbs	+5.9	+10.8	+17.5	0	0	0
	Sparse + Gibbs	+15.7	+44.4	+102.7	−54.2	−20.8	−8.3
Complete	QAOA + Gibbs	+3.4	+8.6	+18.7	0	0	0
	Sparse + Gibbs	+114.4	+244.7	+485.6	−44.4	−33.3	−24.4

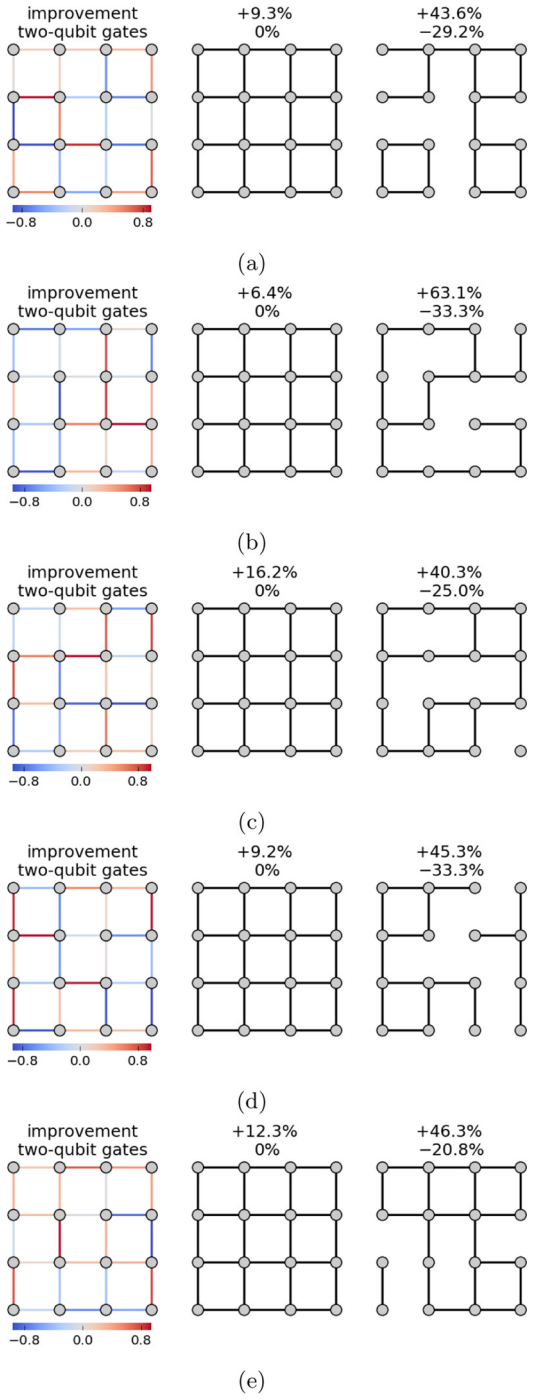

FIG. 9. Five instances of random couplings and the structures of the associated QAOA and best sparse ansatzes for grid problems with the Gibbs objective function. On the left, each edge in the instance graph is colored by its coupling from blue (−1) to red (1). We show the relative improvement of the probability of low energy and reduction of the number of two-qubit gates compared to the usual prescription of the QAOA.

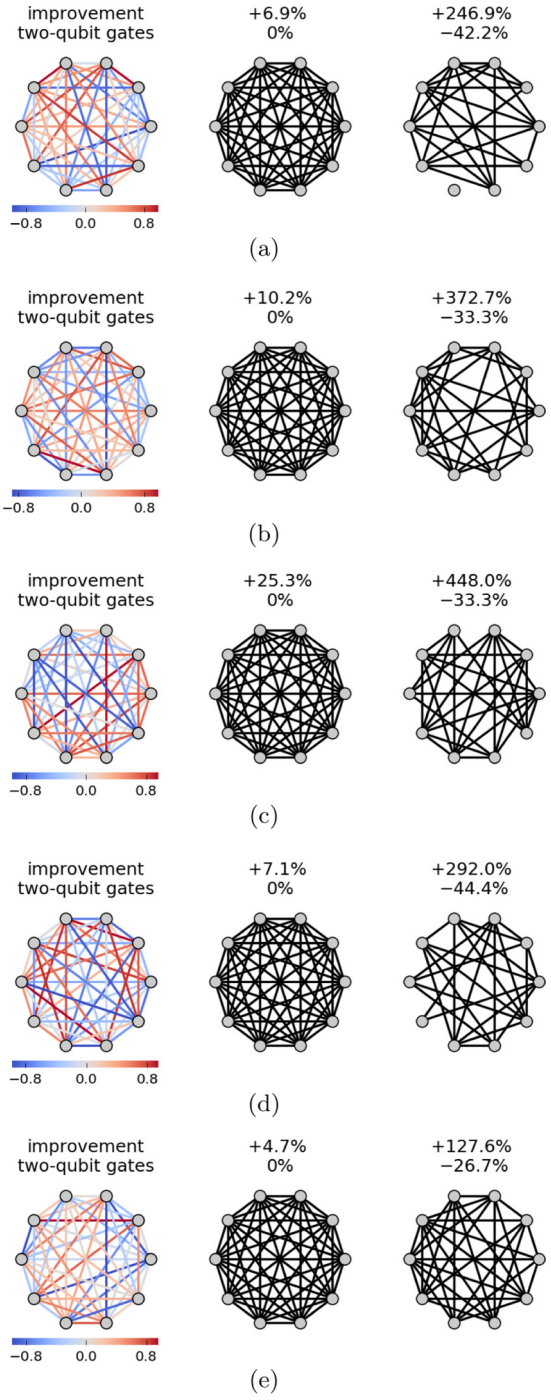

FIG. 10. Five instances of random couplings and the structures of the associated QAOA and best sparse ansatzes for complete graph problems with the Gibbs objective function. On the left, each edge in the instance graph is colored by its coupling from blue (−1) to red (1). We show the relative improvement of the probability of low energy and reduction of the number of two-qubit gates compared to the usual prescription of the QAOA.

APPENDIX F: RELATIVE IMPROVEMENT OF THE PROBABILITY OF LOW ENERGY AND REDUCTION OF THE NUMBER OF TWO-QUBIT GATES

In Table I, we report the relative improvement of the probability of low energy and reduction of the number of two-qubit gates compared to the usual prescription of the QAOA (QAOA + energy) for 1000 grid instances and 1000 complete graph instances. The relative improvement of the

probability of low energy is

$$\left(\frac{P_{\{\text{ansatz}\}+\text{Gibbs}}(E < 0.95E_{\text{gs}})}{P_{\text{QAOA}+\text{energy}}(E < 0.95E_{\text{gs}})} - 1\right) \times 100\%. \quad \text{(F1)}$$

For sparse + Gibbs, the ansatz for each instance is the ansatz with the lowest Gibbs objective function value in AAS by removing up to 20 two-qubit gates. The relative reduction of the number of two-qubit gates is

$$\left(\frac{N(\mathcal{A}_{\{\text{ansatz}\}+\text{Gibbs}})}{N(\mathcal{I})} - 1\right) \times 100\%, \quad \text{(F2)}$$

where $N(\cdot)$ counts the number of edges in the ansatz graph. For the usual prescription of the QAOA, $\mathcal{A}_{\{\text{ansatz}\}+\text{energy}} = \mathcal{I}$, so the relative reduction is always 0%. QAOA + Gibbs brings 10.8% and 8.6% median relative improvement of the probability of low energy for grid and complete graph instances, respectively. By using a sparse ansatz together with the Gibbs objective function, the median relative improvement of the probability of low energy is 44.4% and 244.7%, with reduction of the number of two-qubit gates by 20.8% and 33.3%, for grid and complete graph instances, respectively.

In Figs. 9 and 10, we randomly sample five instances out of 1000 for each instance type and show the structure of their associated QAOA and best sparse ansatzes with the Gibbs objective function.

[1] E. Farhi, J. Goldstone, and S. Gutmann, arXiv:1411.4028.
[2] E. Farhi, J. Goldstone, and S. Gutmann, arXiv:1412.6062.
[3] M. Y. Niu, S. Lu, and I. L. Chuang, arXiv:1905.12134.
[4] E. Farhi and A. W. Harrow, arXiv:1602.07674.
[5] J. Preskill, Quantum **2**, 79 (2018).
[6] L. G. Valiant, Commun. ACM **27**, 1134 (1984).
[7] P. K. Barkoutsos, G. Nannicini, A. Robert, I. Tavernelli, and S. Woerner, Quantum **4**, 256 (2020).
[8] P. McCullagh and J. Kolassa, Scholarpedia **4**, 4699 (2009).
[9] J. A. Nelder and R. Mead, Comput. J. **7**, 308 (1965).
[10] B. Zoph and Q. V. Le, in *Proceedings of 2017 International Conference on Learning Representations* (OpenReview, 2017).
[11] S. Xie, A. Kirillov, R. Girshick, and K. He, in *Proceedings of the 2019 IEEE International Conference on Computer Vision* (Computer Vision Foundation, New York, NY, 2019), pp. 1284–1293.
[12] L. Li and A. Talwalkar, in *Proceedings of the 2019 Conference on Uncertainty in Artificial Intelligence* (2019), http://auai.org/uai2019/accepted.php.
[13] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, Sci. Rep. **9**, 10752 (2019).
[14] G. X. Gu, C.-T. Chen, D. J. Richmond, and M. J. Buehler, Mater. Horiz. **5**, 939 (2018).
[15] M. Schmidt and H. Lipson, Science **324**, 81 (2009).
[16] S.-M. Udrescu and M. Tegmark, arXiv:1905.11481.
[17] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, Nat. Commun. **5**, 4213 (2014).
[18] G. Verdon, J. Pye, and M. Broughton, arXiv:1806.09729.
[19] F. G. Brandao, M. Broughton, E. Farhi, S. Gutmann, and H. Neven, arXiv:1812.04170.
[20] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New J. Phys. **18**, 023023 (2016).
[21] K. Deb, in *Search Methodologies* (Springer, Boston, MA, 2014), pp. 403–449.
[22] See http://www.x.company.
[23] Cirq: A Python Framework for Creating, Editing, and Invoking Noisy Intermediate Scale Quantum (NISQ) Circuits, https://github.com/quantumlib/Cirq.
[24] Apache Beam, https://beam.apache.org.
[25] Z. Li, Q. Chen, and V. Koltun, in *Advances in Neural Information Processing Systems 31 (NIPS 2018)* (Curran Associates, Red Hook, NY, 2018), pp. 539–548.
[26] R. Bunel, M. Hausknecht, J. Devlin, R. Singh, and P. Kohli, in *Proceedings of 2018 International Conference on Learning Representations* (OpenReview, 2018).
[27] I. Sutskever, O. Vinyals, and Q. V. Le, in *Advances in Neural Information Processing Systems 27 (NIPS 2014)* (Curran Associates, Red Hook, NY, 2014), pp. 3104–3112.