Progress toward favorable landscapes in quantum combinatorial optimization

Juneseo Lee, Alicia B. Magann, Herschel A. Rabitz, and Christian Arenz

Phys. Rev. A **104**, 032401 — Published 2 September 2021

DOI: 10.1103/PhysRevA.104.032401

Towards favorable landscapes in quantum combinatorial optimization

Juneseo Lee,[1,2] Alicia B. Magann,[3] Herschel A. Rabitz,[2] and Christian Arenz[2,4]

[1]*Department of Mathematics, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Department of Chemical & Biological Engineering, Princeton University, Princeton, New Jersey 08544, USA*
[4]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*

(Dated: August 6, 2021)

The performance of variational quantum algorithms relies on the success of using quantum and classical computing resources in tandem. Here, we study how these quantum and classical components interrelate. In particular, we focus on algorithms for solving the combinatorial optimization problem MaxCut, and study how the structure of the classical optimization landscape relates to the quantum circuit used to evaluate the MaxCut objective function. In order to analytically characterize the impact of quantum features on the critical points of the landscape, we consider a family of quantum circuit ansätze composed of mutually commuting elements. We identify multiqubit operations as a key resource, and show that overparameterization allows for obtaining favorable landscapes. Namely, we prove that an ansatz from this family containing exponentially many variational parameters yields a landscape free of local optima for generic graphs. However, we further prove that these ansätze do not offer superpolynomial advantages over purely classical MaxCut algorithms. We then present a series of numerical experiments illustrating that non-commutativity and entanglement are important features for improving algorithm performance.

I. INTRODUCTION

Quantum computers promise to offer computational advantages over classical computers for certain high-value tasks [1–3]. However, the availability of fault-tolerant quantum computers that can achieve these speedups at meaningful scales is likely years away. In the meantime, the advent of noisy, intermediate-scale quantum (NISQ) [4] devices has inspired tremendous interest in variational quantum algorithms (VQAs), which aim to leverage the computing power of NISQ devices to solve a broad range of scientific problems, with applications spanning quantum chemistry [5], combinatorial optimization [6], machine learning [7], and linear systems [8]. VQAs function by using NISQ hardware in tandem with a classical processor [9]. At the outset, an objective function J is defined that encodes the solution to a problem of interest. Then, a classical computer is used to iteratively optimize J. The optimization is conventionally performed over a set of parameters associated with a quantum circuit, or *ansatz*, which is used to evaluate J on the NISQ device. In order to achieve strong performance, the classical optimization procedure and the quantum objective function evaluation must function successfully together. As such, an understanding of the associated costs and challenges of the quantum and classical aspects of VQAs, and how they interrelate, is highly desirable.

To date, significant attention has been paid to the quantum component of VQAs, in an effort to develop effective strategies for evaluating the objective functions associated with different problems of interest, and a plethora of ansätze have been developed to target different applications [5–7]. These application-oriented ansätze are often motivated by physical intuition. However, hardware-efficient ansätze have also been developed that are motivated by a convenient implementation on typical NISQ platforms [5, 10]. Furthermore, a variety of error mitigation schemes have been developed in order

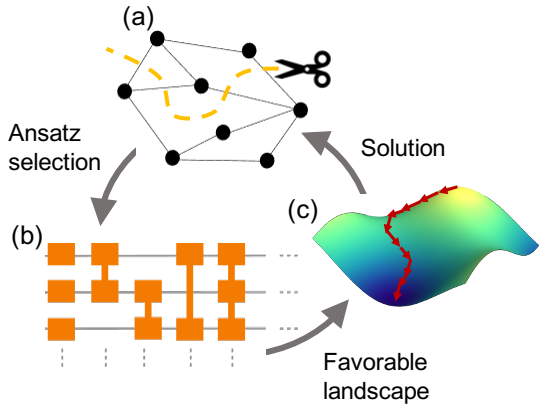

Figure 1. Pictorial representation of how ansatz selection can be informed by the interplay between the problem instance and the underlying optimization landscape structure for solving the MaxCut problem. In particular, the graph structure (a) can be utilized to guide the development of an ansatz (b) for a VQA that yields favorable landscape properties (c), resulting in enhanced convergence. Here, we introduce a family of ansätze as a toy model that allows for analytically studying the associated landscape critical point structure, and prove that an ansatz from this family containing exponentially many variational parameters yields a landscape that is free of local optima. We go on to use this ansatz family as a starting point for exploring relations between inclusion of quantum features in ansätze, scalability, and VQA performance.

to bolster the utility of ansätze in different settings [11–14]. In comparison to these efforts studying the quantum component of VQAs, fewer analyses have been done with respect to the classical component [15–18]. The latter aspect of VQA performance can carry a significant computational cost. This circumstance arises because the classical optimization problem is non-convex in general [19], due to the fact that the quantum circuit parameters typically enter in a nonlinear manner into J. This can cause local optima to appear, which can render the classical search for a global optimum of J prohibitively difficult [20].

Here, we seek to obtain a deeper understanding of these issues by analyzing the optimization landscapes, defined by the objective J as a function of the quantum circuit parameters. The precise manner in which J depends on the parameters is dictated by the interplay between the ansatz and the problem at hand, and consequently, the quantum and classical components of VQAs are intimately related. Thus far, numerical and theoretical observations have confirmed the presence of local optima in VQA landscapes for commonly employed ansätze [20–22]. However, numerical observations have suggested that overparameterization of an ansatz can yield a more favorable landscape structure [18, 23]. We note that similar findings on overparameterization have appeared in the analysis of quantum control landscapes [24–27], where in the latter setting, overparameterization takes the form of "sufficient" pulse-level control resources. In fact, VQAs can themselves be considered a form of quantum learning control experiment [28–32], where the control is performed at the quantum circuit level, rather than at the conventional pulse level [27]. Similar findings on the effects of overparameterization have also appeared in the study of classical neural network landscapes [33–35]. Despite these observations, analytical analyses and rigorous results regarding the critical point structure of VQA landscapes remain scarce, and a comprehensive understanding of VQA landscapes has not yet been attained.

In this article, we make a step in this direction. Namely, we explore quantum-classical tradeoffs in VQAs by studying how the structure of the classical optimization landscape relates to the problem instance and to the quantum ansatz used to evaluate J. In particular, we investigate how ansätze can be designed to yield favorable landscapes that are free of local optima. To do so, we examine how quantum resources, such as entangling gates, can be harnessed to improve the optimization landscape structure to contain only global optima and saddle points, the latter of which often do not hinder local searches from finding global optima efficiently [36–39]. However, as these aims are difficult to achieve in general, we focus here on applications of VQAs for solving the combinatorial optimization problem MaxCut, and consider ansätze that allow for analytically characterizing the critical point structure of the optimization landscape in this setting. Our general approach for this is depicted in Fig. 1. That is, in section III we consider a family of ansätze with elements generated by mutually commuting k-body operators [40, 41], and use these ansätze as a toy model for our studies. The considered ansätze can be tailored to the structure of the graph under consideration, and allow for analyzing the impact of adding variational parameters in a systematic fashion. To this end, we first prove that for generic graphs on n vertices, an ansatz from this family containing $2^{n-1}-1$ variational parameters yields a landscape free of local optima. We go on to explore the prospect of achieving favorable landscapes through ansätze with polynomially many parameters, and discuss how these findings relate to classical algorithmic capabilities. For instance, despite the inclusion of entangling operations, we show that the considered ansatz family does not offer a superpolynomial advantage over purely classical MaxCut schemes. We then numerically explore in section IV how the algorithm performance is affected by incorporating non-commutativity into the ansatz, and compare our findings against the quantum approximate optimization algorithm (QAOA) [6].

II. PRELIMINARIES

Consider a quantum circuit

$$U(\boldsymbol{\theta}) = \prod_{j=1}^{M} e^{-i\theta_j H_j}, \tag{1}$$

parameterized by a set of M (variational) parameters collected in the vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)$, where H_j are Hermitian operators. The parameterized circuit (1) defines an ansatz for optimizing an objective function $J(\boldsymbol{\theta})$. The objective function considered here takes the form

$$J(\boldsymbol{\theta}) = \langle \varphi(\boldsymbol{\theta}) | H_p | \varphi(\boldsymbol{\theta}) \rangle , \tag{2}$$

where H_p is the so-called problem Hamiltonian, and the state $|\varphi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta}) |\phi\rangle$ is created through the parameterized circuit starting from a fixed initial state $|\phi\rangle$. The goal of VQAs is then to solve the optimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^M} J(\boldsymbol{\theta}). \tag{3}$$

Solving (3) is typically accomplished by iteratively searching for the parameters $\boldsymbol{\theta}$ that minimize J in a hybrid quantum-classical fashion. In each iteration a quantum device is used to create the state $|\varphi(\boldsymbol{\theta})\rangle$, followed by expectation value measurements to infer the value of J. Then, a classical search routine is employed to determine how to update the values of $\boldsymbol{\theta}$ for subsequent iteration. This procedure is repeated until convergence is achieved. In order for this procedure to be scalable, the depth of the parameterized circuit, and the number of variational parameters M, should each scale at most polynomially in the number of qubits.

A. Optimization landscapes of VQAs

The ease of finding the parameter configuration that minimizes J depends on the structure of the optimization landscape, given by J as a function of $\boldsymbol{\theta}$. This landscape structure depends on how the components of $\boldsymbol{\theta}$ enter in the associated quantum circuit $U(\boldsymbol{\theta})$ in conjunction with the form of H_p. In order to obtain a favorable landscape, one obvious choice is to consider ansätze [42, 43] that allow for directly varying over the $\mathcal{O}(2^n)$ coefficients of $|\varphi(\boldsymbol{\theta})\rangle$ in the eigenbasis of H_p, assuming the eigenbasis is known, e.g., as for the MaxCut problem studied below. In this case the optimization problem is convex and constrained due to the normalization of $|\varphi(\boldsymbol{\theta})\rangle$. However, such ansätze lack the flexibility to systematically reduce the number of variational parameters, while still maintaining guarantees on the reachability of the ground state. Furthermore, there is no obvious mechanism for tailoring their structure to the problem instance at hand.

In order to address these challenges, here we consider the more common case where the variational parameters $\boldsymbol{\theta}$ enter in a non-convex manner, and consequently, the associated optimization landscapes may contain local optima. In this setting, an assessment of the associated optimization landscape relies on an analysis of the set of critical points $\{\boldsymbol{\theta}^*\}$ at which the gradient $\nabla J(\boldsymbol{\theta})$ vanishes. For objective functions of the form (2), the j-th component of the gradient takes the form

$$\frac{\partial}{\partial\theta_j}J(\boldsymbol{\theta}) = -i\,\langle\varphi(\boldsymbol{\theta})|\,[H_p, W_jH_jW_j^\dagger]\,|\varphi(\boldsymbol{\theta})\rangle\,, \tag{4}$$

where $W_j = \prod_{k=j}^{M} e^{-i\theta_k H_k}$. Due to the form of the gradient (4), it is evident for each of the eigenstates of H_p, which are reachable through $U(\boldsymbol{\theta})$, that the corresponding parameter configurations constitute critical points. In addition, there could also be situations where parameter configurations that yield non-eigenstates constitute critical points.

In order to characterize the type of critical point (e.g. saddle point, minimum, maximum), the Hessian matrix, denoted by $\nabla^2 J(\boldsymbol{\theta})$, can be used. Using the short-hand notation $A_j \equiv W_jH_jW_j^\dagger$, the components of the Hessian are given by

$$\begin{aligned}\frac{\partial^2}{\partial\theta_j\partial\theta_k}J(\boldsymbol{\theta}) = &-\langle\varphi(\boldsymbol{\theta})|\,[[H_p, A_j], A_k]\,|\varphi(\boldsymbol{\theta})\rangle \\ &- i\,\langle\varphi(\boldsymbol{\theta})|\,\frac{\partial}{\partial\theta_j}A_k\,|\varphi(\boldsymbol{\theta})\rangle\,.\end{aligned} \tag{5}$$

To distinguish a saddle point from a local optimum, we define a local optimum as follows:

Definition 1. *A local optimum is a critical point that does not correspond to a global optimum or a saddle, but at which the Hessian is positive or negative semidefinite.*

As recent studies have suggested that saddle points often do not hinder gradient-based algorithms from efficiently finding global optima [36–39], here we focus on the question of whether it is possible to remove local optima by appropriately choosing the ansatz and the initial state $|\phi\rangle$. In particular, we explore which parameterized gates allow for such favorable landscape properties. These considerations emphasize the interplay between quantum resources and classical capabilities. Addressing them in the most general form is challenging, as it would require the ability to analytically track the dependence of H_p and $U(\boldsymbol{\theta})$ on ∇J and $\nabla^2 J$. As a consequence, here we focus on a particular Hamiltonian H_p that has high practical relevance. In particular, we restrict ourselves to Ising Hamiltonians whose ground states encode solutions to the graph-partitioning problem MaxCut.

B. Formulations of the MaxCut problem

Consider a weighted, undirected graph $G = (V, E)$, where V denotes the set of n vertices, E denotes the set of edges, and $w_{a,b} \geq 0$, with $(a, b) \in E$, denote the corresponding non-negative edge weights. The MaxCut problem then corresponds to partitioning V into two subsets such that the sum of the weights belonging to the edges connecting the two subsets is maximized. More formally, for some subset of vertices $S \subset V$, whose complement is denoted by S^c, we define the cut set $\mathsf{Cut}(S)$ with respect to the partition $\{S, S^c\}$ by $\mathsf{Cut}(S) = \{(a, b) \in E, |\, a \in S, b \in S^c\}$. If we denote by $\mathsf{CutVal}(S) = \sum_{(a,b)\in\mathsf{Cut}(S)} w_{a,b}$ the corresponding cut value, the MaxCut problem for G reduces to solving

$$\mathsf{MaxCut}(G) = \max_{S\subset V} \mathsf{CutVal}(S). \tag{6}$$

We note that the MaxCut problem is equivalent to the binary quadratic program

$$\begin{aligned}&\text{minimize} \sum_{(a,b)\in E} w_{a,b}(x_a x_b - 1)/2, \\ &\text{subject to } x_a \in \{\pm 1\} \text{ for every } a \in V,\end{aligned} \tag{7}$$

which is known to be both NP-hard [44] and APX-hard [45] for generic graphs G. As such, significant effort has been dedicated to the development of heuristics and approximation algorithms that efficiently yield high cut values. For example, the Goemans-Williamson (GW) algorithm involves the relaxation of the discrete optimization problem (7) into a semidefinite program, whose initial solution is then rounded to obtain a final solution [46].

It is well-known [47, 48] that solving the MaxCut problem is also equivalent to finding the ground state of an n-qubit Ising Hamiltonian

$$H_p = \sum_{(a,b)\in E} w_{a,b} Z_a Z_b, \tag{8}$$

where $Z_a = \mathbb{1} \otimes \cdots \otimes Z \otimes \cdots \otimes \mathbb{1}$ denotes the Pauli operator $Z = \mathrm{diag}(1, -1)$ acting non-trivially on the ath

qubit. In this setting, MaxCut can be formulated as the optimization problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^M} \sum_{(a,b)\in E} w_{a,b} \left(\langle\varphi(\boldsymbol{\theta})| Z_a Z_b |\varphi(\boldsymbol{\theta})\rangle - 1\right)/2. \quad (9)$$

If we denote the eigenstates of H_p by $|z\rangle$ with $z \in \{0,1\}^n$, we see that each $|z\rangle$ corresponds to a cut of the graph as $\frac{1-\langle z|Z_a Z_b|z\rangle}{2} \in \{0,1\}$, so it is useful to denote by $S_z \subset V$ the set of vertices that are assigned a "1" in the bitstring z associated with $|z\rangle$. The equivalence $\mathsf{CutVal}(S_z) = \mathsf{CutVal}(S_z^c)$ is reflected in fact that the eigenstates of H_p come in pairs with the same eigenvalue, or equivalently cut value, due to the $\mathbb{Z}_2$ symmetry of the Ising Hamiltonian (8). We refer to a set of $2^{n-1}-1$ vertex subsets without any of their complements as being non-symmetric. The minimization problem (9) can be considered a relaxation of the discrete MaxCut optimization problem (7) to quantum states, rather than real vectors on a sphere as in the GW relaxation. It is widely hoped that varying over quantum states $|\varphi(\boldsymbol{\theta})\rangle$ will offer advantages. However, it is largely unknown which quantum features could provide such advantages.

Investigating the optimization landscape of $J(\boldsymbol{\theta})$ for the Ising Hamiltonians (8) offers a potential path forward for assessing what quantum features have to offer in the context of solving the MaxCut problem. To this end, in a recent preprint [20] it has been shown that using an ansatz of the form

$$U(\boldsymbol{\theta}) = \prod_{j=1}^{n} e^{-i\theta_j X_j}, \quad (10)$$

and taking the initial condition $|\phi\rangle = |\mathbf{0}\rangle = |0\rangle \otimes \cdots \otimes |0\rangle$ to be the highest excited state of (8), yields local optima in general, from which it is concluded that the classical optimization problem is NP-hard. In Eq. (10), X_j denotes the Pauli operator $X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ that acts non-trivially on the jth qubit. Since the eigenstates of the Ising Hamiltonian take the form $|z\rangle$ with $z \in \{0,1\}^n$, the classical ansatz (10) can be interpreted as continuously flipping qubits. We note that even though $\exp(-i\theta_j X_j)$ allows for coherent superpositions of qubit states of the form $\cos(\theta_j)|0\rangle - i\sin(\theta_j)|1\rangle$, as H_p is diagonal in the computational basis, such coherent superpositions do not give any advantage over convex combinations. That is, in both cases the objective function takes the form,

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} w_{a,b} \cdot \cos(2\theta_a)\cos(2\theta_b), \quad (11)$$

whose optimization landscape provably contains local optima depending on the graph structure. As such, we refer to (10) as a "classical" ansatz. Furthermore, given that the relative phase of each qubit state does not change J, even an ansatz consisting of generic local SU(2) operations applied to each qubit will produce an objective function of the form (11). Consequently, starting from $|\mathbf{0}\rangle$, the use of generic local operations alone does not yield favorable optimization landscapes. With this in mind, below we explore the effect of incorporating entangling gates.

III. A FAMILY OF ANSÄTZE FOR REMOVING LOCAL OPTIMA

Looking beyond the classical ansatz (11), we proceed by including ansatz elements that are created by k-body operators $\prod_{i\in S} X_i$ acting non-trivially on a subset $S \subset V$ of $|S| = k$ qubits. This leads to a family of ansätze, which we refer to as $\mathbb{X}$-ansätze.

Definition 2. *An $\mathbb{X}$-ansatz is of the form*

$$U(\boldsymbol{\theta}) = \prod_{j=1}^{M} e^{-i\theta_j H_j}, \qquad H_j = \prod_{i\in S_j} X_i, \quad (12)$$

where $\mathcal{A} = \{S_j\}$ with $S_j \subset V$ being a collection of vertex subsets that the ansatz elements non-trivially act on.

We remark that due to commutativity of its elements the set $\mathcal{A}$ uniquely defines the ansatz (12), and that (10) is contained in the family of $\mathbb{X}$-ansätze through $\mathcal{A} = \Big\{\{1\},\{2\},\cdots,\{n\}\Big\}$. As the application of each H_j in (12) on $|\mathbf{0}\rangle$ has the effect of creating a cut, varying over $\boldsymbol{\theta}$ in a given $\mathbb{X}$-ansatz can be interpreted as varying continuously over cuts. Furthermore, these $\mathbb{X}$-ansätze have strong ties to instantaneous quantum polynomial-time (IQP) circuits [40, 41]. In this regard, it is interesting to note that the ability to calculate J for a given $\mathbb{X}$-ansatz efficiently on a classical computer is related to the ability to determine whether barren plateaus are present [49–51]. For further details, we refer to Appendix B.

Since the ansatz elements in (12) mutually commute, the components of the gradient (4) take a particularly simple form

$$\frac{\partial}{\partial\theta_j} J(\boldsymbol{\theta}) = -i \langle\varphi(\boldsymbol{\theta})| [H_p, H_j] |\varphi(\boldsymbol{\theta})\rangle, \quad (13)$$

while the elements of the Hessian (5) are given by

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k} J(\boldsymbol{\theta}) = -\langle\varphi(\boldsymbol{\theta})| [[H_p, H_j], H_k] |\varphi(\boldsymbol{\theta})\rangle,$$

which enables the optimization landscape to be analyzed analytically. We proceed by fixing $|\phi\rangle = |\mathbf{0}\rangle$. Writing out the objective function explicitly and utilizing techniques from [52] for degenerate critical points, at which the Hessian is not invertible [19], allows for establishing the following lemma, whose proof is given in Appendix A.

Lemma 1. *Given an $\mathbb{X}$-ansatz, any critical point of $J(\boldsymbol{\theta})$ not corresponding to an eigenstate of H_p is a saddle.*

Given the goal of understanding the optimization landscape critical point structure, and importantly, the presence of local optima in this landscape, the importance of this lemma is that it allows us to focus completely on critical points with parameter configurations $\boldsymbol{\theta}_E^*$ that correspond to eigenstates of H_p, as all other critical points are saddle points. Since $\langle z| [[H_p, H_j], H_k] |z\rangle = 0$ for all $j \neq k$ and all $z \in \{0,1\}^n$ we immediately have that at these parameter configurations $\boldsymbol{\theta}_E^*$, the Hessian is diagonal, with elements given by

$$\frac{\partial^2}{\partial \theta_j^2} J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_E^*} = -2(J(\boldsymbol{\theta}_E^*) - \langle \varphi(\boldsymbol{\theta}_E^*)| H_j H_p H_j |\varphi(\boldsymbol{\theta}_E^*)\rangle). \tag{14}$$

Instead of considering J as a function of $\boldsymbol{\theta}_E^*$ we can also think of J as dependent on the set S_z that corresponds to the eigenstate created. In this case we write $J\{S_z\}$, where we denote by S_0 and S_g the vertex sets corresponding to highest excited and ground states, respectively. The second term in (14) describes the expectation value of H_p with respect to an eigenstate $H_j |z\rangle$. The corresponding vertex set can be described by the symmetric difference of the set S_j corresponding to H_j, describing which vertices are flipped, and the set S_z, describing the assignment of ones in $|z\rangle$. More formally, if we introduce the symmetric difference of two sets A and B as

$$A \oplus B = A \cup B - A \cap B, \tag{15}$$

we can express the diagonal elements of the Hessian as

$$\frac{\partial^2}{\partial \theta_j^2} J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_E^*} = 2(J\{S_z \oplus S_j\} - J\{S_z\}). \tag{16}$$

We observe that for S_z to be a local minimum, the condition

$$J\{S_z\} \leq J\{S_z \oplus S_j\}, \quad \forall S_j \in \mathcal{A}, \tag{17}$$

has to be satisfied, while for S_z to be a local maximum we need

$$J\{S_z\} \geq J\{S_z \oplus S_j\}, \quad \forall S_j \in \mathcal{A}, \tag{18}$$

to hold. Together with Lemma 1 this allows for establishing the following theorem.

Theorem 1. *The optimization landscape associated with an $\mathbb{X}$-ansatz for which $\mathcal{A}$ contains all non-symmetric $2^{n-1}-1$ vertex subsets exhibits no local optima.*

Proof. We prove Theorem 1 by contradiction. By Lemma (1) we need only consider local optima corresponding to eigenstates. Thus, assume there exists a local minimum with vertex set S_z. By the assumption that all non-symmetric vertex subsets are contained in $\mathcal{A}$, we can pick $S_j = S_z \oplus S_g$. Using properties of the symmetric difference we then have

$$\begin{aligned} J\{S_z\} &\leq J\{S_z \oplus (S_z \oplus S_g)\} \\ &= J\{S_g\}, \end{aligned} \tag{19}$$

which contradicts that by definition $J\{S_g\} < J\{S_z\}$. Analogously, a local maximum S_z satisfying (18) would contradict $J\{S_0\} > J\{S_z\}$, which completes the proof. □

Theorem 1 shows that an $\mathbb{X}$-ansatz with $M = 2^{n-1}-1$ variational parameters yields an optimization landscape whose critical points consists of global optima and saddle points only. It also shows that local optima occurring in the classical ansatz (10) vanish when ansatz elements that contain k-body entangling operators are added.

While Theorem 1 holds for any graph, it requires exponentially many classical parameters θ_j and is therefore not scalable. We now consider whether there exist graphs for which an $\mathbb{X}$-ansatz with polynomially many parameters can be sufficient to obtain an optimization landscape that is free from local optima.

We begin by considering the example of an Ising chain with nearest-neighbor interactions, described by the Hamiltonian

$$H_p = \sum_{j=1}^{n-1} w_{j,j+1} Z_j Z_{j+1}. \tag{20}$$

Depending on the weights $\omega_{j,j+1}$, the classical ansatz (8) yields local optima. However, an $\mathbb{X}$-ansatz described by a path given by $\mathcal{A} = \Big\{\{1\}, \{1,2\}, \cdots, \{1,2,\cdots,n-1\}\Big\}$ does allow for turning all local optima into saddle points. To see this, note that from (14) we have that the diagonal elements of the Hessian at the $|z\rangle$ critical points are given by

$$\frac{\partial^2}{\partial \theta_j^2} J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_E^*} = -4 w_{j,j+1} \langle z| Z_j Z_{j+1} |z\rangle . \tag{21}$$

Together with Lemma 1, we can then conclude that the only critical points at which the Hessian is positive (negative) semidefinite are global minima (maxima). Consequently, for the Ising chain, an optimization landscape free from local optima can be obtained with an $\mathbb{X}$-ansatz consisting of $n-1$ variational parameters and with a circuit depth polynomial in n. To see the latter, we note that in addition to the circuit containing only linearly many ansatz elements, each ansatz element can itself be implemented efficiently with standard universal quantum gate sets [53]. Furthermore, it is straightforward to generalize this conclusion to chains with periodic boundary conditions (i.e., to all connected 2-regular graphs). In this latter case, an $\mathbb{X}$-ansatz described by n paths, each of length $n-1$ but starting at a different vertex, gives an optimization landscape exhibiting global optima and saddle points only, using $\mathcal{O}(n^2)$ variational parameters.

These examples illustrate that including quantum features in the ansatz, here in the form of k-body entangling operators, can allow for obtaining a favorable landscape while maintaining scalability. We now consider whether scalable $\mathbb{X}$-ansätze can yield a landscape free from local optima for other classes of graphs, where MaxCut is

nontrivial. Mathematically, this translates into the question of whether for a given graph G, an $\mathbb{X}$-ansatz with $|\mathcal{A}| = \text{poly}(n)$ exists so that conditions (17) and (18) can only be satisfied at the global optima.

Theorem 2. *For any graph G and an $\mathbb{X}$-ansatz $\mathcal{A}$ with size $|\mathcal{A}|$, there exists a purely classical algorithm that has the same solution set as the set of local optima satisfying conditions (17) and (18).*

Proof. Consider the purely classical algorithm for solving MaxCut, shown in Fig. 2 and outlined in the associated figure caption. The condition that an output of the classical algorithm is a cut in which the CutVal cannot be increased any further by a flip of a single set of vertices S_k is precisely condition (17). Changing "increase" to "decrease" in the algorithm immediately yields those that satisfy (18), although for the purposes of MaxCut this set is irrelevant. This completes the proof. □

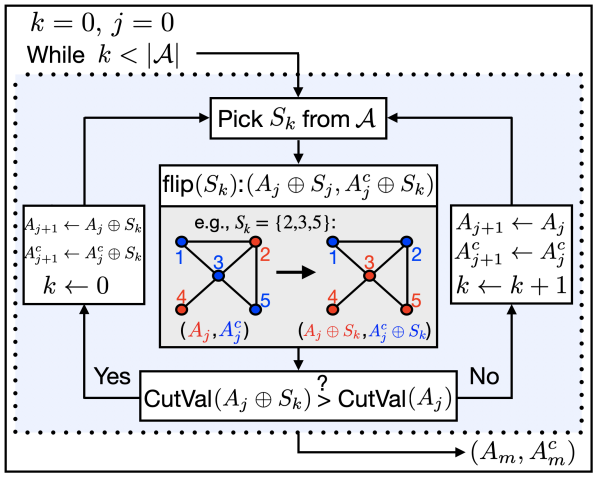

Figure 2. Purely classical algorithm for solving MaxCut. Start with a random bipartition (A_0, A_0^c) of the vertices V. Then, iteratively construct new bipartitions $(A_1, A_1^c), (A_2, A_2^c), \ldots, (A_m, A_m^c)$. At each iteration $j = 1, \cdots, m$, pick an ansatz element $S_k \in \mathcal{A}$. Beginning with an ansatz element labeled by $k = 0$, implement the flip(S_k) operation by flipping the assignments of each vertex in S_k to obtain $A_j \oplus S_k$ and $A_j^c \oplus S_k$. If flip(S_k) increases the CutVal such that CutVal$(A_j \oplus S_k) >$ CutVal(A_j) (or, equivalently, decreases the objective function J), then increment j and set $A_{j+1} = A_j \oplus S_k$ and $A_{j+1}^c = A_j^c \oplus S_k$ and return to $k = 0$. Otherwise, increment k and repeat the flip operation until CutVal$(A_j \oplus S_k) >$ CutVal(A_j) is satisfied. Continue this procedure until a bipartition (A_m, A_m^c) is reached such that CutVal(A_m) cannot be increased any further by a single flip.

We remark that the manner in which a particular S_k is picked at each step in the algorithm shown in Fig. 2 is irrelevant. Choosing S_k uniformly at random from $\mathcal{A}$ yields a randomized algorithm, whereas iteratively testing each $S_k \in \mathcal{A}$ and choosing the largest-increasing S_k at each step yields a greedy algorithm akin to the classical 0.5-approximation scheme for MaxCut [54].

This immediately yields the following corollary:

Corollary 2.1. *Unless P=NP, there does not exist a class of $\mathbb{X}$-ansätze with $|\mathcal{A}| = poly(n)$ that yields a landscape free from local optima for generic graphs.*

Proof. If there exists a landscape free from local optima, then the only cut that satisfies either (17) or (18) is the one corresponding to the global minimum (namely, the maximum cut) or global maximum. By Theorem (2), if there exists an $\mathbb{X}$-ansatz $\mathcal{A}$ with size $|\mathcal{A}| = poly(n)$, there also exists a purely classical greedy algorithm that would always converge to the global minimum as well. Notice that for an unweighted graph the maximum cut value is at most $|E|$, so the purely classical algorithm converges in at most $|\mathcal{A}| \cdot |E|$ steps, thus solving unweighted MaxCut with polynomial cost. Since unweighted MaxCut is NP-hard, and so is MaxCut for arbitrary graphs (see [54] for the extension of greedy algorithms to weighted graphs), this shows that unless P=NP, there does not exist a class of $\mathbb{X}$-ansätze with $|\mathcal{A}| = poly(n)$ that generically yields landscapes consisting of global optima and saddle points only. □

This relation can be extended even further, by considering approximation schemes for MaxCut, aimed at achieving an approximation ratio $\alpha = \text{CutVal}(S)/\text{MaxCut}(G)$ for some $S \subset V$. Then, assuming the existence of an algorithm that can escape saddle points:

Corollary 2.2. *Given a fixed approximation ratio α, and any $\mathbb{X}$-ansatz with $|\mathcal{A}| = poly(n)$, even an algorithm that can escape saddle points cannot provide a superpolynomial advantage over a purely-classical α-approximation scheme for MaxCut.*

Proof. Analogously as to the proofs of Theorem 2 and Corollary 2.1, the key idea is that the solution set among the local optima that satisfy conditions (17) or (18) is indifferent to each such potential solution in a gradient algorithm, even one that can escape saddle points. Similarly, the classical approximation scheme presented in the proof of Theorem (2) is also indifferent to each of the potential solutions, and for a fixed α it converges in at most $\frac{2 \cdot |\mathcal{A}|}{1-\alpha}$ steps, which is polynomial in n if $|\mathcal{A}| = poly(n)$. As such, any provable approximation ratios α given by each of the algorithms are identical. Consequently, no superpolynomial advantage exists. □

This result begs the question: what is required for a quantum advantage in this setting? In the following section, we assess the role of non-commutativity through a series of numerical experiments.

IV. NUMERICAL EXPERIMENTS

To systematically assess how including k-body elements in an ansatz affects the structure of the underlying optimization landscape, we define the k-body depth of an $\mathbb{X}$-ansatz as $D = \max_{S_j \in \mathcal{A}} |S_j|$. Here, we remark

that while "width" or "span" may be more apt descriptors of this quantity with respect to ansatz elements acting on a graph, we choose to refer to it as a "depth" in order to serve as a proxy for the more common circuit depth complexity, as depicted in Figure (3)(a). Given this, we first aim to explore how the structure of the optimization landscape changes when the k-body depth is successively increased, towards an ansatz containing all k-body operators, which according to Theorem (1) yields a landscape free from local optima. We then proceed by introducing non-commutative elements in the $\mathbb{X}$-ansatz, and investigate whether such extensions exhibit faster convergence to better approximation ratios. Finally, we compare the $\mathbb{X}$-ansatz and its non-commutative variants against QAOA.

We focus our numerical analyses on complete graphs K_n with random positive edge weights, for which MaxCut is known to be NP-hard [44]. In each numerical experiment we solve (9) for K_n with $w_{a,b}$ chosen uniformly randomly from $[0, 5]$, and utilize the first-order gradient BroydenFletcherGoldfarbShanno (BFGS) algorithm with a randomly chosen initial parameter configuration $\boldsymbol{\theta}$. Details regarding the hyperparameters used can be found in Appendix C. In each run we calculate the approximation ratio α, given as the ratio between the actual MaxCut value $\mathsf{MaxCut}(K_n)$, obtained from exact diagonalization of H_p, and the cut value obtained from solving (9) using BFGS. The curves in the figures below show the average taken over 100 realizations and the shaded areas show the corresponding standard deviation.

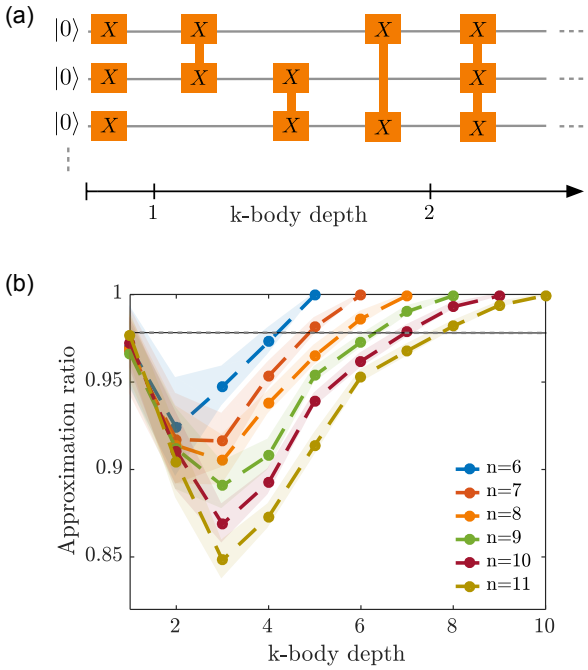

Figure 3. The performance of the $\mathbb{X}$-ansatz, whose circuit diagram is shown in (a), for solving MaxCut on complete graphs is shown in (b). In (a) boxes represent variational ansatz elements. Connected boxes represent entangling ansatz elements that are generated by k-body X operators acting non-trivially on k "boxed" qubits. In (b) the approximation ratio is shown as a function of the k-body depth for different problem sizes n. The circles correspond to the average taken over 100 randomly chosen graph instances and BFGS initial conditions. The shaded areas show the corresponding standard deviations. The grey dashed line indicates the threshold for when better approximation ratios than the classical ansatz (10) are achieved.

A. Dependence on the k-body depth

We begin by investigating how the approximation ratio that is obtained changes when the k-body depth in the $\mathbb{X}$-ansatz is increased. In particular, we consider the $\mathbb{X}$-ansatz schematically represented in Fig. 3(a) where increasing the k-body depth by one is achieved by adding $\binom{n}{k}$ k-body operators. That is, for a fixed k-body depth D the ansatz consists of $M = \sum_{k=1}^{D} \binom{n}{k}$ ansatz elements, noting that $D = 1$ corresponds to the classical ansatz (10) with $M = n$ local rotations. We further note that for $D = n$ we have $M = 2^n - 1$, so that for $D = n - 1$ the number of variational parameters $M = 2^n - 2$ scales exponentially in the number of qubits n. The approximation ratio as a function of the k-body depth is shown in Fig. 3(b) for different n. We first observe that the classical ansatz performs very well as approximation ratios ≥ 0.95 are achieved. Numerical simulations shown in Appendix C indeed confirm that solving (9) using BFGS for graphs with a large vertex degree yield approximation ratios that are slightly better than the ones obtained from the GW algorithm. However, from Fig. 3(b) we also see that adding quantumness in the form of 2-body entangling terms does not increase the approximation ratio. Instead, performance drops until a sufficiently large k-body depth is reached (here > 3). This behavior suggests that simply adding additional 2-body terms to the classical ansatz does not automatically make the optimization landscape structure more favorable, as a drop in the approximation ratio indicates that randomly initialized first-order gradient algorithms are in this case more prone to get stuck in local optima or saddle points. However, increasing the k-body depth even further allows for obtaining better approximation ratios than the classical ansatz, which is indicated by a grey dashed line. We further observe in Fig. 3(b) that when $D = n-1$, i.e, exponentially many variational parameters are used, average approximation ratios of ≥ 0.99 with standard deviations $\leq 10^{-6}$ are achieved. This behavior is in line with Theorem 1, as an $\mathbb{X}$-circuit with exponentially many parameters yields a landscape free from local optima, while the appearance of saddle points does not seem to affect the performance of BFGS. However, we remark here that according to Theorem 1, a landscape not exhibiting local optima is already obtained at a lower k-body depth $D > n/2$, as all non-symmetric vertex subsets are then contained in $\mathcal{A}$. It is interesting to note that in this case,

smaller approximation ratios correspond to runs converging to degenerate saddle points.

One way to justify this behavior disappearing when we increase the k-body depth from $n/2$ to $n-1$ is to consider that at depth $n-1$, each parameter is effectively included twice (for any ansatz element S_k with parameter θ_k, there exists its complementary element $S_j = S_k^c$ for some j, with parameter θ_j). Since the critical point conditions are equivalent for θ_k and θ_j, the probability of all pairs satisfying the conditions (and thus yielding a critical point) at depth $n-1$ is squared relative to the probability that each of the $2^{n-1}-1$ values at depth $n/2$ satisfies them. Thus, the probability of observing this phenomenon at depth $n-1$ is significantly lower than at depth $n/2$.

The results shown in Fig. 3(b) suggest that the $\mathbb{X}$-ansatz with sufficiently many k-body terms performs better than the classical ansatz (10). However, according to Corollary (2.2), there is also a purely classical strategy to achieve the same approximation ratios. A natural next question to consider is whether introducing non-commutativity in the $\mathbb{X}$-ansatz will improve performance.

B. Assessing the role of non-commutativity

We proceed by introducing non-commutative ansatz elements into the $\mathbb{X}$-ansatz. We consider ansätze of the form

$$U(\boldsymbol{\theta}) = \prod_j e^{-i\tilde{\theta}_j \tilde{H}_j} e^{-i\theta_j H_j}, \tag{22}$$

where $H_j \in \{\prod_{i\in S} X_i \,|\, S \in \mathcal{A}\}$ are the generators of the $\mathbb{X}$-ansatz elements determined by $\mathcal{A}$ and non-commutativity is introduced through the Hermitian operators $\tilde{H}_j$. As schematically represented in Fig. 4(a) and (b), we consider the case where the k-body depth is increased by including ansatz elements generated by Pauli Z operators between the elements of the $\mathbb{X}$-ansatz in the last section, which we refer to as $\mathbb{XZ}$-ansätze. We treat two different cases. Namely, in (a) we consider $\tilde{H}_k \in \{\prod_{i\in S} Z_i \,|\, S \in \mathcal{A}\}$ while in (b) $\tilde{H}_k = \sum_{i=1}^n Z_i$ for all k. As such, now the number of variational parameters is increased by $2\binom{n}{k}$ when the k-body depth is increased by one. The results are shown for $n = 8$ in Fig. 4(c).

We first observe that both $\mathbb{XZ}$-ansätze yield faster convergence than the $\mathbb{X}$-ansatz, which is not surprising as the number of variational parameters has doubled. However, it is interesting to observe that the performance between (a) and (b) differs only slightly; at a k-body depth of 4, both $\mathbb{XZ}$-ansätze yield approximation ratios of ≈ 0.98 while the $\mathbb{X}$-ansatz achieves ≈ 0.94, which is even lower than for the classical ansatz (10).

Another way of introducing non-commutativity is repeating a given structure, which consists of ansatz elements that do not mutually commute. A standard example for such an ansatz is QAOA [6].

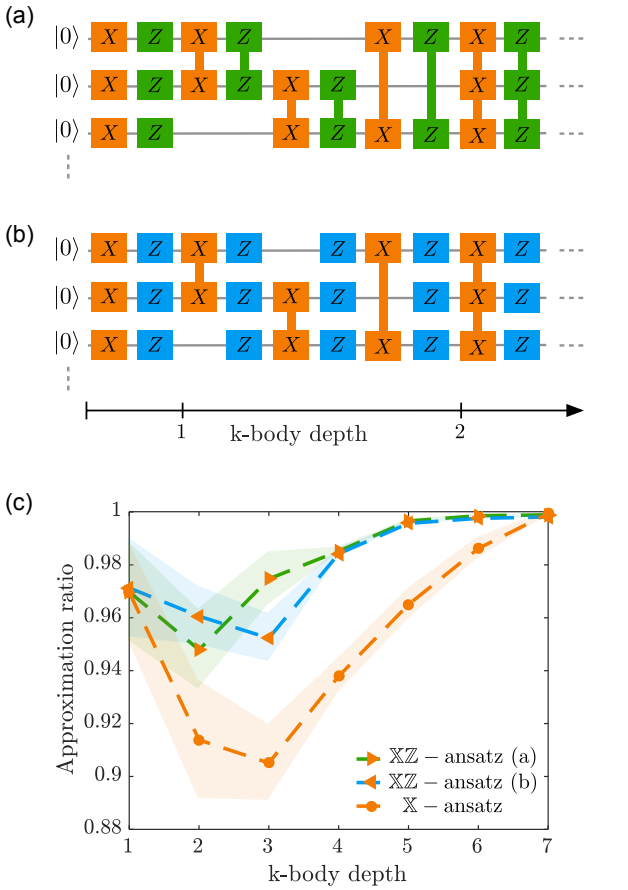

Figure 4. The role of non-commutativity is assessed by introducing variational ansatz elements generated by Pauli Z operators into the $\mathbb{X}$-ansatz. This is achieved by alternating between ansatz elements generated by k-body X (orange) and (a) k-body Z (green) operators and (b) single-qubit Z operators (blue). In (c), the approximation ratio is shown as a function of the k-body depth for $n = 8$ qubits. The circles/triangles correspond to the average taken over 100 randomly chosen complete graphs and BFGS initial conditions. The shaded areas show the corresponding standard deviations.

C. Comparison with QAOA and initial state dependence

QAOA is a VQA designed to solve combinatorial optimization problems that was developed in 2014 [6] and has since inspired many works [55–61]. QAOA aims to generate approximate solutions to combinatorial optimization problems such as MaxCut through an ansatz of the form

$$U(\boldsymbol{\theta}) = \prod_j e^{-i\tilde{\theta}_j H_P} e^{-i\theta_j H_X}, \tag{23}$$

where $H_X = \sum_{j=1}^n X_j$. In QAOA the initial state is given by $|\phi\rangle = |+\rangle$ where $|+\rangle = |+\rangle \otimes \cdots \otimes |+\rangle$ with $|+\rangle$ being an eigenstate of X, so that for sufficiently many alternations between H_p and H_X as in (23) a ground state of H_p is reachable. In contrast to the $\mathbb{X}$- and $\mathbb{XZ}$-

ansätze in the last section, for which the ground state is reachable at a k-body depth of $D = 1$, a sufficiently large circuit depth is needed in QAOA before the ground state can be created.

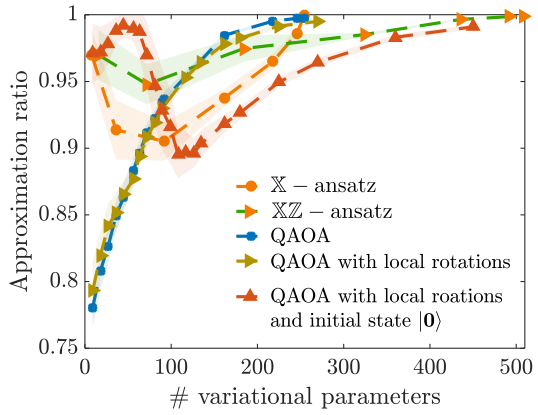

Figure 5. A comparison of different versions of QAOA with the $\mathbb{X}$- and $\mathbb{XZ}$-ansätze for $n = 8$ qubits is shown. The circles/triangles correspond to the average approximation ratio taken over 100 randomly chosen complete graphs and BFGS initial conditions. The shaded areas show the corresponding standard deviations.

In Fig. 5 we compare the approximation ratios obtained from QAOA (blue) with the $\mathbb{X}$- and $\mathbb{XZ}$-ansatz (orange and orange/green, respectively) represented in Fig. 3(a) and Fig. 4(a) as a function of the number of variational parameters for $n = 8$ qubits. We note that since Fig. 5 compares the performance obtained using the $\mathbb{X}$- and $\mathbb{XZ}$-ansätze against the performance of QAOA and its variants, we plot the results as a function of the number of variational parameters, rather than the k-body depth, as the latter is not a relevant quantity in QAOA. However, for the former $\mathbb{X}$- and $\mathbb{XZ}$-ansätze, the number of variational parameters serves as a proxy for the k-body depth, as per Figs. 3 and 4.

We first observe that for a small number of variational parameters, QAOA performs worse than the $\mathbb{X}$- and the $\mathbb{XZ}$-ansätze. This behavior can be traced back to the fact that for a small number of variational parameters (and accordingly, a shallow circuit), the ground state of H_p may not be reachable through the ansatz (23). In contrast, as the classical ansatz (10) is contained in the $\mathbb{X}$- and the $\mathbb{XZ}$-ansätze, for $M = n$ variational parameters the ground state is already reachable for a small number of classical parameters, which explains why the $\mathbb{X}$- and the $\mathbb{XZ}$-ansätze perform better in this regime than QAOA. However, we also see from Fig. 5 that with increasing numbers of variational parameters, QAOA outperforms the $\mathbb{X}$- and the $\mathbb{XZ}$-ansatz, as for QAOA a better approximation ratio can be obtained with fewer variational parameters. One may then wonder whether we can combine these favorable aspects of QAOA and the $\mathbb{X}$- and the $\mathbb{XZ}$-ansätze, e.g., by modifying QAOA such that for a few variational parameters high approximation ratios ≥ 0.95 are obtained, while increasing the number of variational parameters continuously improves the approximation ratios, or conversely, whether it is possible to avoid the drop in the approximation ratios observed in Fig. 3(b) and Fig. 4(c) when the k-body depth is increased.

A natural modification of QAOA is to incorporate n independent local X rotations at each layer [62], such that the classical ansatz (10) is now contained in the ansatz (23). This modification is well-motivated, due to the fact that even on non-interacting spin problems, conventional QAOA (i.e., using only global X rotations) can fail to converge [63]. However, we note that this does not automatically guarantee that the ground state is then reachable for a smaller number of variational parameters, as reachability also depends on the initial state $|\phi\rangle$. And indeed, from the olive-green curve in Fig. 5, we see that a modification of QAOA to include local rotations while having $|\phi\rangle = |+\rangle$ does not substantially change the convergence behavior, suggesting that a lack of reachability may affect the performance in this case. In comparison, if additionally the initial state is changed to $|\phi\rangle = |\mathbf{0}\rangle$ (red), then $M = n$ variational parameters do allow for retaining high approximation ratios ≥ 0.96. Furthermore, increasing M allows for increasing the approximation ratio to ≈ 0.99 at $M \approx 45$. These results suggest that in addition to the incorporation of non-commutativity into an ansatz, the ability to guarantee reachability of the set of solution states with shallow circuits can enhance the performance of variational quantum algorithms. However, we do note that continuing to increase M causes the approximation ratio to drop again, indicating that further work is needed to fully understand the tradeoffs in how these aspects of algorithm design impact performance. We also note that the numerical simulations in Fig. 5 suggest that the modified version of QAOA (red) and the $\mathbb{XZ}$-ansatz require $M \approx 2^{n+1}$ variational parameters in order to achieve approximation ratios asymptotically approaching 1.

V. CONCLUSIONS

The successful implementation of VQAs rests on the ability to solve the underlying optimization problem, which in turn is determined by the structure of the corresponding optimization landscape. The landscape structure depends on the problem under consideration, and on the parameterized quantum circuit ansatz used to evaluate the associated objective function. In this work, we have focused on VQAs for solving the combinatorial optimization problem MaxCut, and studied the interplay between the graph type, ansatz, and critical point structure of the optimization landscape in this setting. In particular, we introduced a family of ansätze consisting of mutually commuting elements generated by k-body Pauli X operators, which we termed $\mathbb{X}$-ansätze. We proved that for generic graphs, an ansatz from this family containing exponentially many variational parameters yields

an optimization landscape that consists of global optima and saddle points only. Next, we considered examples for which an ansatz from this family with polynomially many variational parameters yields a landscape free from local optima, but for which MaxCut is also efficiently solvable classically. We then showed that for a given graph, a polynomially sized ansatz from the considered family exists if and only if there exists a purely classical algorithm that allows for solving MaxCut efficiently. As a consequence, we concluded that this ansatz family cannot offer a superpolynomial advantage over purely classical schemes.

We went on to numerically study whether introducing non-commutative ansatz elements could improve performance. For the $\mathbb{X}$-ansatz and its variants, we found that the addition of k-body terms did not improve the landscape structure in a manner that yields monotonically improving approximation ratios. However, for sufficiently high k-body depth, non-commutative versions of the $\mathbb{X}$-ansatz did perform better than their commutative counterparts. In total, we believe that the $\mathbb{X}$-ansatz family can be useful as a baseline, upon which new ideas for further performance improvement can be studied. We note that for relatively shallow circuit depth, the best performance overall was obtained through a modified version of QAOA, motivated by our prior findings, that includes independent local rotations.

More research is needed to determine how quantum resources contribute to favorable landscape properties and if effective ansätze can be identified in a problem-dependent manner in the future. To this end, it is interesting to note that a "classical" ansatz, consisting of only local rotations, in combination with a first-order gradient algorithm, achieves remarkably high ≈ 0.95 approximation ratios, and performs slightly better on average than the purely classical Goemans and Williamson MaxCut approximation algorithm for graphs with sufficiently high vertex degree. Due to the provable existence of local optima when utilizing the former classical ansatz, this suggests that the optimization landscape could be favorable in the sense that gradient algorithms converge to critical points with high associated approximation ratios.

In order to improve beyond these results, it would be desirable to steadily remove local optima from the landscape, while retaining only local optima with high corresponding approximation ratios. Furthermore, an equally important goal would be to widening the basin of attraction of global optima and reducing those of the remaining local optima [63]. In addition, analyzing the curvature of the optimization landscape can reveal information about an algorithm's robustness to noise [25], suggesting that it could be desirable to design future ansätze to take this into account in order to enhance the quality of NISQ implementations.

Future studies may benefit from using the former "classical" ansatz as a starting point, from which additional "quantum" features are added, i.e., through the design of adaptive ansätze [15, 64–67], e.g. seeded from local rotations and then grown by incorporating gates that introduce non-commutativity and entanglement with the goal of continuously improving the landscape structure such that higher approximation ratios are attainable. Such procedures could additionally be guided by characterizing the geometry [68] and entangling power [18] of the parameterized circuit.

The ideal outcome would be the development of VQAs that are *scalable*, and have optimization landscapes with a favorable critical point structure. The most favorable landscapes would be free of local optima, and also free from certain types of saddle points. That is, while in this work we did not distinguish between classes of saddle points, the convergence of classical optimization routines can be significantly hindered by the presence of degenerate saddle points, as opposed to more favorable "strict" saddle points. Furthermore, the most favorable landscapes would also be free of barren plateaus [49–51] (see also Appendix B), where the gradient becomes exponentially small, thereby impeding the progress of optimization algorithms. In fact, taken together, ansätze with polynomially many variational parameters lacking in barren plateaus, degenerate saddles, and local optima would allow a global optimum to be found efficiently using first-order gradient methods [37–39]. However, for problems like MaxCut, we do not expect that such efficient solutions will be feasible in general, as we do not anticipate that VQAs will be able to solve NP-hard problems efficiently. Nevertheless, the achievement of any subset of these goals for scalable VQA ansätze would be highly desirable. Looking ahead, we hope that the strategies outlined above will allow for systematically assessing how to improve VQA performance across a range of applications.

ACKNOWLEDGEMENTS

We thank B. Hanin, J. McClean, M. McConnell, and M. Zhandry for valuable conversations and insights. C.A. acknowledges support from the ARO (Grant No. W911NF-19-1-0382). A.M. acknowledges support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-FG02-97ER25308. H. R. acknowledges support from the ARO (Grant No. W911NF-19-1-0382) for the theoretical aspects of the research and from the DOE STTR (Grant No. DE-SC0020618) for the algorithmic implications of the research.

The authors are pleased to acknowledge that the work reported on in this paper was substantially performed using the Princeton Research Computing resources at Princeton University which is consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing.

This report was prepared as an account of work sponsored by an agency of the United States Government.

[1] P. W. Shor, Algorithms for quantum computation: Discrete logarithms and factoring, in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, SFCS '94 (IEEE Computer Society, USA, 1994) pp. 124–134.
[2] L. K. Grover, A fast quantum mechanical algorithm for database search, in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* (1996) pp. 212–219.
[3] S. Lloyd, Universal Quantum Simulators, Science **273**, 1073 (1996).
[4] J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum **2**, 79 (2018).
[5] 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 (2013).
[6] E. Farhi, J. Goldstone, and S. Gutmann, A Quantum Approximate Optimization Algorithm, arXiv e-prints , arXiv:1411.4028 (2014), arXiv:1411.4028 [quant-ph].
[7] V. Dunjko and P. Wittek, A non-review of Quantum Machine Learning: trends and explorations, Quantum Views **4**, 32 (2020).
[8] C. Bravo-Prieto, R. LaRose, M. Cerezo, Y. Subasi, L. Cincio, and P. J. Coles, Variational Quantum Linear Solver: A Hybrid Algorithm for Linear Systems, arXiv e-prints , arXiv:1909.05820 (2019), arXiv:1909.05820 [quant-ph].
[9] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).
[10] 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 **549**, 242 (2017).
[11] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, Phys. Rev. X **7**, 021050 (2017).
[12] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Phys. Rev. X **8**, 031027 (2018).
[13] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, Phys. Rev. Lett. **119**, 180509 (2017).
[14] X. Bonet-Monroig, R. Sagastizabal, M. Singh, and T. E. O'Brien, Low-cost error mitigation by symmetry verification, Phys. Rev. A **98**, 062339 (2018).
[15] H. R. Grimsley, S. E. Economou, E. Barnes, and N. J. Mayhall, An adaptive variational algorithm for exact molecular simulations on a quantum computer, Nature Communications **10**, 3007 (2019).
[16] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum Natural Gradient, Quantum **4**, 269 (2020).
[17] B. Koczor and S. C. Benjamin, Quantum analytic descent (2020), arXiv:2008.13774 [quant-ph].
[18] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the hamiltonian variational ansatz, PRX Quantum **1**, 020319 (2020).
[19] P. Jain and P. Kar, Non-convex optimization for machine learning, arXiv:1712.07897 (2017), arXiv:1712.07897.
[20] L. Bittel and M. Kliesch, Training variational quantum algorithms is np-hard – even for logarithmically many qubits and free fermionic systems, arXiv:2101.07267 [quant-ph] (2021), arXiv: 2101.07267.
[21] D. Wierichs, C. Gogolin, and M. Kastoryano, Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer, Phys. Rev. Research **2**, 043246 (2020).
[22] N. Moll, P. Barkoutsos, L. S. Bishop, J. M. Chow, A. Cross, D. J. Egger, S. Filipp, A. Fuhrer, J. M. Gambetta, M. Ganzhorn, A. Kandala, A. Mezzacapo, P. Müller, W. Riess, G. Salis, J. Smolin, I. Tavernelli, and K. Temme, Quantum optimization using variational algorithms on near-term quantum devices, Quantum Science and Technology **3**, 030503 (2018).
[23] B. T. Kiani, S. Lloyd, and R. Maity, Learning unitaries by gradient descent, arXiv:2001.11897 [quant-ph] (2020), 2001.11897.
[24] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, Quantum optimally controlled transition landscapes, Science **303**, 1998 (2004).
[25] R. Chakrabarti and H. Rabitz, Quantum control landscapes, Int. Rev. Phys. Chem. **26**, 671 (2007).
[26] B. Russell, H. Rabitz, and R.-B. Wu, Control landscapes are almost always trap free: A geometric assessment, J. Phys. A **50**, 205302 (2017).
[27] A. B. Magann, C. Arenz, M. D. Grace, T.-S. Ho, R. L. Kosut, J. R. McClean, H. A. Rabitz, and M. Sarovar, From pulses to circuits and back again: A quantum optimal control perspective on variational quantum algorithms, PRX Quantum **2**, 010101 (2021).
[28] R. S. Judson and H. Rabitz, Teaching lasers to control molecules, Phys. Rev. Lett. **68**, 1500 (1992).
[29] J. Li, X. Yang, X. Peng, and C.-P. Sun, Hybrid quantum-classical approach to quantum optimal control, Phys. Rev. Lett. **118**, 150503 (2017).
[30] Q.-M. Chen, X. Yang, C. Arenz, R.-B. Wu, X. Peng, I. Pelczer, and H. Rabitz, Combining the synergistic control capabilities of modeling and experiments: Illustration of finding a minimum-time quantum objective, Phys. Rev. A **101**, 032313 (2020).
[31] X.-d. Yang, C. Arenz, I. Pelczer, Q.-M. Chen, R.-B.

Wu, X. Peng, and H. Rabitz, Assessing three closed-loop learning algorithms by searching for high-quality quantum control pulses, Phys. Rev. A **102**, 062605 (2020).
[32] D. Dong, M. A. Mabrok, I. R. Petersen, B. Qi, C. Chen, and H. Rabitz, Sampling-based learning control for quantum systems with uncertainties, IEEE Transactions on Control Systems Technology **23**, 2155 (2015).
[33] A. Shevchenko and M. Mondelli, Landscape connectivity and dropout stability of SGD solutions for over-parameterized neural networks, Proceedings of the 37th International Conference on Machine Learning **119**, 8773 (2020).
[34] Z. Allen-Zhu, Y. Li, and Y. Liang, Learning and generalization in overparameterized neural networks, going beyond two layers, arXiv:1811.04918 (2018), arXiv:1811.04918.
[35] Z. Chen, Y. Cao, D. Zou, and Q. Gu, How much over-parameterization is sufficient to learn deep relu networks?, arXiv:1911.12360 (2019), arXiv:1911.12360.
[36] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, First-order methods almost always avoid saddle points, arXiv:1710.07406 (2017), arXiv:1710.07406.
[37] R. Ge, F. Huang, C. Jin, and Y. Yuan, Escaping from saddle points — online stochastic gradient for tensor decomposition, in *Proceedings of The 28th Conference on Learning Theory*, Proceedings of Machine Learning Research, Vol. 40, edited by P. Grünwald, E. Hazan, and S. Kale (PMLR, Paris, France, 2015) pp. 797–842.
[38] K. Y. Levy, The power of normalization: Faster evasion of saddle points, arXiv:1611.04831 (2016), arXiv:1611.04831.
[39] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, How to escape saddle points efficiently, in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by D. Precup and Y. W. Teh (PMLR, International Convention Centre, Sydney, Australia, 2017) pp. 1724–1732.
[40] D. Shepherd and M. J. Bremner, Temporally unstructured quantum computation, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **465**, 1413 (2009).
[41] K. Fujii and T. Morimae, Commuting quantum circuits and complexity of ising partition functions, New Journal of Physics **19**, 033003 (2017).
[42] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).
[43] X.-M. Zhang, M.-H. Yung, and X. Yuan, Low-depth quantum state preparation, arXiv preprint arXiv:2102.07533 (2021).
[44] R. M. Karp, Reducibility among combinatorial problems, in *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*, edited by R. E. Miller, J. W. Thatcher, and J. D. Bohlinger (Springer US, Boston, MA, 1972) pp. 85–103.
[45] C. H. Papadimitriou and M. Yannakakis, Optimization, approximation, and complexity classes, Journal of Computer and System Sciences **43**, 425 (1991).
[46] M. X. Goemans and D. P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, J. ACM **42**, 1115 (1995).
[47] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, arXiv (2020), 2012.09265 [quant-ph].
[48] A. Lucas, Ising formulations of many np problems, Frontiers in Physics **2**, 5 (2014).
[49] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 1 (2018).
[50] C. Arenz and H. Rabitz, Drawing together control landscape and tomography principles, Phys. Rev. A **102**, 042207 (2020).
[51] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nature Communications **12**, 1791 (2021).
[52] T. S. Bolis, Degenerate critical points, Mathematics Magazine **53**, 294 (1980).
[53] F. Tacchino, A. Chiesa, S. Carretta, and D. Gerace, Quantum computers as universal quantum simulators: State-of-the-art and perspectives, Adv. Quantum Technol. **3**, 1900052 (2020).
[54] S. Kahruman-Anderoglu, E. Kolotoglu, S. Butenko, and I. Hicks, On greedy construction heuristics for the max-cut problem, Int. J. Comput. Sci. Eng. **3**, 211 (2007).
[55] 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 preprint arXiv:1712.05771 (2017).
[56] 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. Photonics **12**, 534 (2018).
[57] M. Willsch, D. Willsch, F. Jin, H. De Raedt, and K. Michielsen, Benchmarking the quantum approximate optimization algorithm, Quantum Inf. Process. **19**, 197 (2020).
[58] D. M. Abrams, N. Didier, B. R. Johnson, M. P. da Silva, and C. A. Ryan, Implementation of the XY interaction family with calibration of a single pulse, Nat. Electron. **3**, 744 (2020).
[59] 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).
[60] G. Pagano, A. Bapat, P. Becker, K. S. Collins, A. De, P. W. Hess, H. B. Kaplan, A. Kyprianidis, W. L. Tan, C. Baldwin, and et al., Quantum approximate optimization of the long-range ising model with a trapped-ion quantum simulator, Proc. Natl. Acad. Sci. U.S.A **117**, 25396 (2020).
[61] M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya, J. Atalaya, J. C. Bardin, R. Barends, S. Boixo, and et al., Quantum approximate optimization

of non-planar graph problems on a planar superconducting processor, Nat. Phys. **17**, 332 (2021).
[62] S. Hadfield, Z. Wang, B. O'Gorman, E. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019).
[63] J. R. McClean, M. P. Harrigan, M. Mohseni, N. C. Rubin, Z. Jiang, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Low depth mechanisms for quantum optimization, arXiv preprint arXiv:2008.08615 (2020).
[64] L. Zhu, H. L. Tang, G. S. Barron, 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 [quant-ph] (2020), arXiv: 2005.10258.
[65] 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 ansatze on a quantum processor, arXiv:1911.10205 [quant-ph] (2020), arXiv: 1911.10205.
[66] Z.-J. Zhang, T. H. Kyaw, J. S. Kottmann, M. Degroote, and A. Aspuru-Guzik, Mutual information-assisted Adaptive Variational Quantum Eigensolver Ansatz Construction, arXiv:2008.07553 [quant-ph] (2020), arXiv: 2008.07553.
[67] A. B. Magann, K. M. Rudinger, M. D. Grace, and M. Sarovar, Feedback-based quantum optimization (2021), arXiv:2103.08619 [quant-ph].
[68] T. Haug, K. Bharti, and M. Kim, Capacity and quantum geometry of parametrized quantum circuits, arXiv:2102.01659 [quant-ph] (2021), arXiv: 2102.01659.
[69] J. Stewart, *Multivariable Calculus: Concepts and Contexts*, Available 2010 Titles Enhanced Web Assign Series (Cengage Learning, 2009).
[70] J. J. Meyer, Gradients just got more flexible, Quantum Views **5**, 50 (2021).
[71] A. Vardy, The intractability of computing the minimum distance of a code, IEEE Transactions on Information Theory **43**, 1757 (1997).
[72] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-lbfgsb.html.
[73] https://pypi.org/project/cvxgraphalgs/.

Appendix A: Proof of Lemma (1)

We first provide an outline of the proof, including major components and equations, in appendix A 1, followed by a detailed proof in appendix A 2.

1. Proof Outline

For ease and consistency of notation throughout the proof, we implicitly associate H_j with the vertex subset S_j on which H_j acts. The first observation is that for all j such that $(a,b) \notin \mathsf{Cut}(H_j)$, we have that $e^{-i\theta_j H_j}$ commutes with $Z_a Z_b$. This observation motivates us to define the set $\mathcal{C}_{(a,b)} = \{H_j \in \mathcal{A} \mid (a,b) \in \mathsf{Cut}(H_j)\}$ and $\mathcal{K}_{(a,b)} = \{K \subset \mathcal{C}_{(a,b)} \mid \oplus\{H \in K\} = \emptyset\}$, which allows for rewriting the objective function $J(\boldsymbol{\theta})$ as:

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}} \left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K} \cos(2\theta_j) \prod_{H_j\in K} i\sin(2\theta_j) \right] \tag{A1}$$

Notice that $\cos(2\theta_k)$ and $\sin(2\theta_k)$ never appear in the same product together, such that for a certain k we can write:

$$J(\boldsymbol{\theta}) = \cos(2\theta_k)S_k + \sin(2\theta_k)T_k + V_k, \tag{A2}$$

where S_k, T_k, V_k do not depend on θ_k. This result allows for easy expressions for the gradient and Hessian (where for ease of notation we use $\partial_k J(\boldsymbol{\theta})$ for the k-th gradient element and $\partial^2_{j,k} J(\boldsymbol{\theta})$ for the j,k element of the Hessian):

$$\partial_k J(\boldsymbol{\theta}) = 2\Big[-\sin(2\theta_k)S_k + \cos(2\theta_k)T_k \Big] \tag{A3}$$

$$\partial^2_{k,k} J(\boldsymbol{\theta}) = -4\Big[\cos(2\theta_k)S_k + \sin(2\theta_k)T_k \Big] \tag{A4}$$

The condition that $\partial_k J(\boldsymbol{\theta}) = 0$ at a critical point yields:

$$\sin(2\theta_k)S_k = \cos(2\theta_k)T_k \tag{A5}$$

Extensive analysis on the relative signs of $\sin(2\theta_k), \cos(2\theta_k), S_k, T_k$ yields that nondegenerate critical points are either eigenstates of H_p or saddle points, where we use the following definition of a saddle point:

Definition 3. *A parameter configuration $\boldsymbol{\theta}^* \in \mathbb{R}^M$ is a saddle point of J if it is a critical point, but for all $\epsilon > 0$ there exist $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ with $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_1\|, \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_2\| < \epsilon$ and $J(\boldsymbol{\theta}_1) < J(\boldsymbol{\theta}^*) < J(\boldsymbol{\theta}_2)$. Furthermore, it is well-known [69] that a sufficient condition for $\boldsymbol{\theta}^*$ being a saddle point is that the set of scalars $\{z^T\Big(\nabla^2 J(\boldsymbol{\theta}^*)\Big)z : z \in \mathbb{R}^M\}$ contains both positive and negative elements, where $\nabla^2 J(\boldsymbol{\theta}^*)$ is the Hessian matrix of J evaluated at $\boldsymbol{\theta}^*$.*

Throughout the proof, we utilize either the former definition or the latter sufficient condition to show that a particular nondegenerate critical point is either a saddle point or an eigenstate of H_p.

For the case of degenerate critical points, we justify and apply Proposition 3 from [52], which states (reworded here from [52] to fit our context and notation):

Proposition 1. *Let $f(x) = \sum_{j=0}^{\infty} \frac{1}{j!} F_j(x)$ be the Taylor expansion of a function $f : \mathbb{R}^M \to \mathbb{R}$, where F_j is the j-th Taylor form. Let F_p denote the first nonzero Taylor form of f at a critical point a of f, let K_p denote the kernel of F_p, and let F_s be the first Taylor form that does not vanish identically on K_p (noting that at a critical point $2 \leq p < s$, and s may not exist). Now, suppose that F_p is positive semi-definite but not positive definite. If s exists and F_s takes a negative value on K_p, or if s does not exist, then a is a saddle point of f.*

Applying Proposition (1) yields that any degenerate critical point is a saddle, so that combined with the cases for nondegenerate critical points, we obtain that any parameter configuration at a critical point not corresponding to an eigenstate of H_p is a saddle. □

2. Details of the Proof

a. *Computation of the objective function*

We first remark that much of the computations of this section are only to derive the form of $J(\boldsymbol{\theta})$ as in equation (A24), which has also been shown in similar forms with respect to the well-known parameter shift rule [70]. Readers interested in the main portion of the proof can thus skip straight to equation (A24).

To derive a manipulable form of the cost function, we first compute $J(\boldsymbol{\theta})$, which we expand by utilizing linearity, to yield a form that is easier to manipulate:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \langle \varphi(\boldsymbol{\theta}) | \, H_p \, | \varphi(\boldsymbol{\theta}) \rangle \\ &= \sum_{(a,b)\in E} w_{a,b} \, \langle \varphi(\boldsymbol{\theta}) | \, Z_a Z_b | \varphi(\boldsymbol{\theta}) \rangle \\ &= \sum_{(a,b)\in E} w_{a,b} \, \langle \mathbf{0} | \left[(\prod_{j=1}^{M} e^{i\theta_j H_j}) Z_a Z_b (\prod_{j=1}^{M} e^{-i\theta_j H_j}) \right] | \mathbf{0} \rangle \end{aligned} \tag{A6}$$

Here, notice that for all j such that $(a,b) \notin \mathsf{Cut}(H_j)$, we have that $e^{i\theta_j H_j}$ commutes with $Z_a Z_b$. For ease of notation, let $\mathcal{C}_{(a,b)} = \{H_j \in \mathcal{A} \mid (a,b) \in \mathsf{Cut}(H_j)\}$, corresponding to the elements that do not commute with $Z_a Z_b$. Thus:

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} w_{a,b} \, \langle \mathbf{0} | \left[(\prod_{H_j \in \mathcal{C}_{(a,b)}} e^{i\theta_j H_j}) Z_a Z_b (\prod_{H_j \in \mathcal{C}_{(a,b)}} e^{-i\theta_j H_j}) \right] | \mathbf{0} \rangle \tag{A7}$$

Now, we first recall the well-known formula for matrices A, B with $B^2 = \mathbb{1}$:

$$e^{i\alpha B} A e^{-i\alpha B} = \cos^2(\alpha) A + \sin^2(\alpha) BAB + i \sin(\alpha)\cos(\alpha)[B, A] \tag{A8}$$

Setting $A = Z_a Z_b$, $B = H_j$, and $\alpha = \theta_j$ yields:

$$e^{i\theta_j H_j} Z_a Z_b e^{-i\theta_j H_j} = \cos^2(\theta_j) Z_a Z_b + \sin^2(\theta_j) H_j Z_a Z_b H_j + i \sin(\theta_j)\cos(\theta_j)[H_j, Z_a Z_b] \tag{A9}$$

For $H_j \in \mathcal{C}_{(a,b)}$, without loss of generality assume $\{a,b\} \cap H_j = \{a\}$. Then, using simple algebra we obtain:

$$\begin{aligned} e^{i\theta_j H_j} Z_a Z_b e^{-i\theta_j H_j} &= \cos^2(\theta_j) Z_a Z_b + \sin^2(\theta_j) H_j Z_a Z_b H_j + i \sin(\theta_j)\cos(\theta_j)[H_j, Z_a Z_b] \\ &= \cos^2(\theta_j) Z_a Z_b + \sin^2(\theta_j) X_a Z_a Z_b X_a + i \sin(\theta_j)\cos(\theta_j)(H_j X_a Z_b)[X_a, Z_a] \\ &= \cos(2\theta_j) Z_a Z_b + \sin(2\theta_j) H_j X_a Z_b Y_a \\ &= \cos(2\theta_j) Z_a Z_b + i \sin(2\theta_j) H_j Z_a Z_b \\ &= (\cos(2\theta_j)\mathbb{1} + i \sin(2\theta_j) H_j) Z_a Z_b \end{aligned} \tag{A10}$$

Thus, since each of the H_j's commute and $Z_a Z_b \, |\mathbf{0}\rangle = |\mathbf{0}\rangle$ for all $(a,b) \in E$, we arrive at:

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} w_{a,b} \, \langle \mathbf{0} | \left[\prod_{H_j \in \mathcal{C}_{(a,b)}} \Big(\cos(2\theta_j)\mathbb{1} + i \sin(2\theta_j) H_j \Big) \right] | \mathbf{0} \rangle \tag{A11}$$

Now, notice that the summands in the expansion of $\prod_{H_j \in \mathcal{C}_{(a,b)}}$ are products over all $H_j \in \mathcal{C}_{(a,b)}$ of either $\cos(2\theta_j)\mathbb{1}$ or $i\sin(2\theta_j)H_j$. Taking the expectation of a particular summand with respect to the state $|\mathbf{0}\rangle$ yields a non-zero value if and only if the product of the H_j's that are included is the identity; since each H_j corresponds to a particular vertex subset, this condition can also be expressed as the symmetric difference of all included H_j's being the empty set. This leads us to define $\mathcal{K}_{(a,b)} = \{K \subset \mathcal{C}_{(a,b)} \mid \oplus\{H \in K\} = \emptyset\}$, where for a particular $K \in \mathcal{K}_{(a,b)}$ the elements $H_j \in K$ are those corresponding to the $i\sin(2\theta_j)$ terms, and the elements $H_j \in \mathcal{C}_{(a,b)} - K$ correspond to the $\cos(2\theta_j)$ terms. This yields:

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)}} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \tag{A12}$$

While the presence of the imaginary unit i in the product may seem problematic given that the expectation J must be real here, we can see that $|K|$ is even for all $K \in \mathcal{K}_{(a,b)}$ (since in the construction of $\mathcal{K}_{(a,b)}$ we note that elements in K must have exactly one of either a or b, each of which requires an even number of elements in order to attain an empty symmetric difference), so that the product remains real.

Now, consider the gradient element $\partial_k J(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial\theta_k} J(\boldsymbol{\theta})$ corresponding to an ansatz element defined by some H_k:

$$\partial_k J(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_k} \sum_{(a,b)\in E} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)}} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \tag{A13}$$

Since θ_k appears only if $H_k \in \mathcal{C}_{(a,b)}$, only edges (a,b) in $\mathsf{Cut}(H_k)$ have a non-zero partial derivative:

$$\partial_k J(\boldsymbol{\theta}) = \sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)}} \frac{\partial}{\partial\theta_k} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \tag{A14}$$

Since θ_k appears exactly once in the product $\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j)$ (either as $\cos(2\theta_k)$ or $i\sin(2\theta_k)$ according to whether $H_k \in K$ or not), we can split this as follows:

$$\begin{aligned} \partial_k J(\boldsymbol{\theta}) = \sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \Bigg(& \frac{\partial}{\partial\theta_k} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \notin K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \\ & + \frac{\partial}{\partial\theta_k} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \in K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \Bigg) \end{aligned} \tag{A15}$$

Now, notice that in order to replace $\cos(2\theta_k)$ with its derivative $-2\sin(2\theta_k)$, we can multiply by $\frac{-2\sin(2\theta_k)}{\cos(2\theta_k)}$, and similarly for the derivative of $\sin(2\theta_k)$:

$$\begin{aligned} \partial_k J(\boldsymbol{\theta}) = & \sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \Bigg(\frac{-2\sin(2\theta_k)}{\cos(2\theta_k)} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \notin K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \\ & + \frac{2\cos(2\theta_k)}{\sin(2\theta_k)} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \in K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \Bigg) \\ = & \frac{-2\sin(2\theta_k)}{\cos(2\theta_k)} \Bigg(\sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \notin K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \Bigg) \\ & + \frac{2\cos(2\theta_k)}{\sin(2\theta_k)} \Bigg(\sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \in K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \Bigg) \end{aligned} \tag{A16}$$

Now, for ease of notation in arguing the remainder of this proof, define:

$$S_k = \frac{1}{\cos(2\theta_k)} \sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \notin K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \tag{A17}$$

$$T_k = \frac{1}{\sin(2\theta_k)} \sum_{(a,b)\in \mathsf{Cut}(H_k)} w_{a,b} \sum_{K \in \mathcal{K}_{(a,b)} \ s.t. \ H_k \in K} \left[\prod_{H_j \in \mathcal{C}_{(a,b)} - K} \cos(2\theta_j) \prod_{H_j \in K} i\sin(2\theta_j) \right] \tag{A18}$$

Notice in particular that S_k, T_k do not depend on θ_k, by construction since the prefactor $\frac{1}{\cos(2\theta_k)}$ cancels the existing $\cos(2\theta_k)$ term in the product for S_k, and similarly for T_k. Thus, we have:

$$\partial_k J(\boldsymbol{\theta}) = -2\sin(2\theta_k)S_k + 2\cos(2\theta_k)T_k = 2\Big[-\sin(2\theta_k)S_k + \cos(2\theta_k)T_k\Big] \tag{A19}$$

We can then easily compute the diagonal Hessian elements $\partial^2_{k,k} J(\boldsymbol{\theta}) \equiv \frac{\partial^2}{\partial\theta_k^2} J(\boldsymbol{\theta})$ as well:

$$\partial^2_{k,k} J(\boldsymbol{\theta}) = 2\Big[-2\cos(2\theta_k)S_k - 2\sin(2\theta_k)T_k\Big] = -4\Big[\cos(2\theta_k)S_k + \sin(2\theta_k)T_k\Big] \tag{A20}$$

Using this result, we can expand S_k, T_k in the formula for the diagonal Hessian element to relate its value to the objective function $J(\boldsymbol{\theta})$:

$$\begin{aligned}
-\frac{\partial^2_{k,k} J(\boldsymbol{\theta})}{4} &= \cos(2\theta_k)S_k + \sin(2\theta_k)T_k \\
&= \cos(2\theta_k)\cdot\frac{1}{\cos(2\theta_k)}\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\notin K}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \\
&\quad + \sin(2\theta_k)\cdot\frac{1}{\sin(2\theta_k)}\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\in K}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \\
&= \sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\notin K}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \\
&\quad + \sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\in K}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right]
\end{aligned} \tag{A21}$$

Since we sum over all $K \in \mathcal{K}_{(a,b)}$ such that $H_k \notin K$ and those such that $H_k \in K$, this is in fact a complete sum:

$$\begin{aligned}
-\frac{\partial^2_{k,k} J(\boldsymbol{\theta})}{4} &= \sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \\
&= \sum_{(a,b)\in E} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \\
&\quad - \sum_{(a,b)\notin\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \\
&= J(\boldsymbol{\theta}) - \sum_{(a,b)\notin\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right]
\end{aligned} \tag{A22}$$

Thus, we have:

$$J(\boldsymbol{\theta}) = -\frac{\partial^2_{k,k} J(\boldsymbol{\theta})}{4} + \sum_{(a,b)\notin\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos(2\theta_j)\prod_{H_j\in K} i\sin(2\theta_j)\right] \tag{A23}$$

Here, since as mentioned above θ_k appears only if $H_k \in \mathcal{C}_{(a,b)}$, the second summand does not depend on θ_k; this allows us to proceed with our analysis easily.

We now can summarize the main ingredients required here:

$$J(\boldsymbol{\theta}) = \cos(2\theta_k)S_k + \sin(2\theta_k)T_k + (\text{term not dependent on } \theta_k) \tag{A24}$$

$$\partial_k J(\boldsymbol{\theta}) = 2\Big[-\sin(2\theta_k)S_k + \cos(2\theta_k)T_k\Big] \tag{A25}$$

$$\partial^2_{k,k} J(\boldsymbol{\theta}) = -4\Big[\cos(2\theta_k)S_k + \sin(2\theta_k)T_k\Big] \tag{A26}$$

At a critical point $\boldsymbol{\theta}^*$, we have that for each k that $\partial_k J(\boldsymbol{\theta}) = 0$. From (A25) we thus have that at a critical point the condition

$$\sin(2\theta_k)S_k = \cos(2\theta_k)T_k \tag{A27}$$

has to hold for all k.

b. *Case considerations*

In order to prove that any critical point $\boldsymbol{\theta}^*$ not corresponding to an eigenstate of H_p is a saddle point, we consider the cases (a)-(c) below. We first consider case **(a)**, and show that if for some k both sides of (A27) are equal but non-zero, meaning that $\sin(2\theta_k) \neq 0$, $\cos(2\theta_k) \neq 0$, $S_k \neq 0$, and $T_k \neq 0$, the corresponding critical points correspond to saddle points. We then go on to consider the case **(b)** where for some k both sides are zero with $S_k = T_k = 0$, showing that in this case the corresponding critical points must be saddle points too. Here we will distinguish between (i) non-degenerate and (ii) degenerate critical points, where for case (ii) we utilize Proposition 1. As the cases (a) and (b) correspond to saddle points, the only case left for which a critical point $\boldsymbol{\theta}^*$ could potentially not correspond to a saddle point is **(c)**, where for each k either $\cos(2\theta_k) = S_k = 0$ or $\sin(2\theta_k) = T_k = 0$, so that together condition (A27) is satisfied for all k. We finally show that such critical points satisfying (c) correspond to eigenstates of H_p.

Case (a): There exists k such that $\sin(2\theta_k) \neq 0$, $\cos(2\theta_k) \neq 0$, $S_k \neq 0$, $T_k \neq 0$

If for some k the above holds, we can rewrite condition (A27) as $\frac{\sin(2\theta_k)}{\cos(2\theta_k)} = \frac{T_k}{S_k}$. The objective function $J(\boldsymbol{\theta})$ given by (A24) can then be rewritten as:

$$J(\boldsymbol{\theta}) = \cos(2\theta_k) \cdot \frac{S_k^2 + T_k^2}{S_k} + (\text{term not dependent on } \theta_k) \tag{A28}$$

We note again that S_k and T_k are independent of θ_k. Moreover, since $\sin(2\theta_k) \neq 0$ we have $\cos(2\theta_k) \neq \pm 1$. As such, for all $\epsilon > 0$ the ϵ-ball around θ_k both increases and decreases the value of $\cos\Big(2(\theta_k \pm \epsilon)\Big)$. Thus, by definition (3) all critical points corresponding to case (a) are saddle points.

Case (b): There exists k such that $S_k = T_k = 0$

(i) Non-degenerate case (invertible Hessian)
We first note that if $S_k = T_k = 0$ for some k, from (A26) we see that then $\partial^2_{k,k} J(\boldsymbol{\theta}) = 0$. Now, consider the vector u with 1 in the k-th entry and 0 everywhere else. Since the Hessian matrix $\nabla^2 J(\boldsymbol{\theta})$ is invertible by definition, so that $\Big(\nabla^2 J(\boldsymbol{\theta})\Big)u \neq \mathbf{0}$, there exists some l such that the off-diagonal Hessian element $\partial^2_{k,l} J(\boldsymbol{\theta}) \neq 0$. For $t \in \mathbb{R}$, consider the set of vectors v_t with value t in the l-th position, 1 in the k-th position, and 0 everywhere else, so that

$$v_t^T \Big(\nabla^2 J(\boldsymbol{\theta})\Big) v_t = t^2 \partial^2_{l,l} J(\boldsymbol{\theta}) + 2t \partial^2_{k,l} J(\boldsymbol{\theta}). \tag{A29}$$

If $\partial^2_{l,l} J(\boldsymbol{\theta}) = 0$, then (A29) is linear in t. Thus, in this case the left-hand side attains for $t \in \mathbb{R}$ both positive and negative values. If $\partial^2_{l,l} J(\boldsymbol{\theta}) \neq 0$, then (A29) is quadratic in t with roots at $t = 0$ and $t = -\frac{2\partial^2_{k,l} J(\boldsymbol{\theta})}{\partial^2_{l,l} J(\boldsymbol{\theta})} \neq 0$. As such, here the left-hand side attains for $t \in \mathbb{R}$ both positive and negative values too. From the sufficient condition in the definition (3) of a saddle point we conclude that in the case (b)(i) the corresponding critical points correspond to saddle points.

(ii) Degenerate case (non-invertible Hessian)
We first show that in order to obtain a critical point not corresponding to a saddle point, the Hessian $\nabla^2 J(\boldsymbol{\theta})$ at these points must be diagonal. As above, since $S_k = T_k = 0$ we know that $\partial^2_{k,k} J(\boldsymbol{\theta}) = 0$. If the Hessian is not diagonal, then there exists some l such that the off-diagonal Hessian element $\partial^2_{k,l} J(\boldsymbol{\theta}) \neq 0$, which allows for proceeding as in case (b)(i). We conclude that critical points yielding a non-diagonal Hessian correspond to saddle points.

Therefore, at critical points that do not immediately correspond to saddle points the Hessian must be diagonal with at least one element being 0. We treat this case by applying Proposition (1) from above, identifying

the function $f(x)$ as the objective function $J(\boldsymbol{\theta})$. Without loss of generality, assume the Hessian is positive semi-definite (as an analogous argument holds for the negative semi-definite case). Notice that the kernel of the Hessian (which we denote by K_p as in the setting of Proposition (1)) is the set of vectors with zeros in all indices j for which the j-th diagonal Hessian element $\partial^2_{j,j}J(\boldsymbol{\theta})$ is non-zero, and any real values for elements corresponding to other indices. Furthermore, as in the setting of Proposition (1) consider the case where s exists (since otherwise we immediately obtain that we have a saddle) such that F_s is the first Taylor form that does not vanish identically on K_p. Without loss of generality, assume $s = 3$ (as an analogous argument holds for $s > 3$), and let $\mathfrak{T}$ be the order-3 tensor representing the third-derivatives of J. First, notice that the "diagonal" elements satisfy $\partial^3_{k,k,k}J(\boldsymbol{\theta}) = -8\Big[-\sin(2\theta_k)S_k + \cos(2\theta_k)T_k\Big] = -4\partial_k J(\boldsymbol{\theta}) = 0$, which implies that if the kernel K_p is one-dimensional, $\mathfrak{T}$ is zero identically on K_p. As such, in order for $\mathfrak{T}$ not to vanish identically on K_p, there must be some set of indices j, k, l such that $\partial^2_{j,j}J(\boldsymbol{\theta}) = \partial^2_{k,k}J(\boldsymbol{\theta}) = \partial^2_{l,l}J(\boldsymbol{\theta}) = 0$ but $\partial^3_{j,k,l}J(\boldsymbol{\theta}) \neq 0$. Now, let $u \in K_p$ be the vector with 1 in the j-th, k-th, and l-th entries and 0 everywhere else. Notice that by construction, applying $\mathfrak{T}$ to u and $-u$ yields non-zero values with opposite signs, so in particular $\mathfrak{T}$ takes a negative value on K_p. We can now apply Proposition (1), which states that we have a saddle point in this case.

Case (c) For all k, either $\sin(2\theta_k) = 0$ and $T_k = 0$, or $\cos(2\theta_k) = 0$ and $S_k = 0$

From (A12) we recall the form of the objective function:

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}} \left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K} \cos(2\theta_j) \prod_{H_j\in K} i\sin(2\theta_j)\right] \tag{A30}$$

Within the set $\mathcal{C}_{(a,b)}$, each H_j either has $\sin(2\theta_j) = 0$ or $\cos(2\theta_j) = 0$. This means that at most one of the products can be non-zero, since all other products will contain at least one configuration with $\sin(2\theta_j)$ or $\cos(2\theta_j)$ being 0. This non-zero product therefore either takes the value 1 or -1 (as it is a product of sin and cos that are each either 1 or -1), yielding:

$$J(\boldsymbol{\theta}) = \sum_{(a,b)\in E} (\pm w_{a,b}), \tag{A31}$$

where the sign of a particular $w_{a,b}$ is determined by whether the number of $H_j \in \mathcal{C}_{(a,b)}$ that have $\cos(2\theta_j) = -1$ or $\sin(2\theta_j) = -1$ is even or odd. As discussed in section II, this assignment of signs precisely corresponds to a cut of the graph, which in turn corresponds to an eigenstate of H_p, as desired.

At this juncture we also remark that an alternate proof of case (c) can be seen by noticing that $\sin(2\theta_k) = 0$ or $\cos(2\theta_k) = 0$, implying $\theta_k = n\pi$ and $\theta_k = (2n+1)/2\pi$ respectively, such that all gates belong to full flips of qubits or a global phase. As the circuit starts in an eigenstate (namely, $|\mathbf{0}\rangle$), the circuit also ends in an eigenstate, as desired. □

Appendix B: Barren Plateaus

The phenomenon of barren plateaus has recently been considered as one of the main bottlenecks for VQAs [47]. A barren plateau appears if the gradient of the objective function becomes exponentially small, which is typically analyzed by randomly "sampling" the optimization landscape given by $J(\boldsymbol{\theta})$. That is, a barren plateau appears if the variance $\mathrm{Var}\Big(\partial_k J(\boldsymbol{\theta})\Big)$ of the components of the gradient becomes exponentially small in the number of qubits n, while the expectation $\mathbb{E}[\partial_k J(\boldsymbol{\theta})]$ vanishes for all k. Here, we compute the variances explicitly for the $\mathbb{X}$-ansatz. The expectations are taken over the parameters $\boldsymbol{\theta}$, each of which are considered to be independently and identically uniformly distributed in $[0, 2\pi)$. Throughout the computations, we utilize the well-known identities $\mathbb{E}[\sin(x)] = \mathbb{E}[\cos(x)] = 0$ and $\mathbb{E}[\sin^2(x)] = \mathbb{E}[\cos^2(x)] = \frac{1}{2}$.

In order to compute the variances, we first recall from (A25) the form of the gradient element $\partial_k J(\boldsymbol{\theta})$:

$$\begin{aligned}\partial_k J(\boldsymbol{\theta}) = \frac{-2\sin(2\theta_k)}{\cos(2\theta_k)}&\left(\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\notin K} \left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K} \cos(2\theta_j) \prod_{H_j\in K} i\sin(2\theta_j)\right]\right) \\ &+ \frac{2\cos(2\theta_k)}{\sin(2\theta_k)}\left(\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b} \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\in K} \left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K} \cos(2\theta_j) \prod_{H_j\in K} i\sin(2\theta_j)\right]\right) \qquad \text{(B1)} \\ &= 2\Big[-\sin(2\theta_k)S_k + \cos(2\theta_k)T_k\Big] \qquad \text{(B2)}\end{aligned}$$

Thus, since S_k, T_k are independent from θ_k and each have finite expectation, we can write:

$$\mathbb{E}[\partial_k J(\boldsymbol{\theta})] = 2\Big[-\mathbb{E}[\sin(2\theta_k)]\mathbb{E}[S_k] + \mathbb{E}[\cos(2\theta_k)]\mathbb{E}[T_k]\Big] = 2\cdot(0\cdot\mathbb{E}[S_k] + 0\cdot\mathbb{E}[T_k]) = 0, \tag{B3}$$

where we used the identities described above. For the variance, we can first write:

$$\begin{aligned}
\mathbb{E}[(\partial_k J(\boldsymbol{\theta}))^2] &= \mathbb{E}[4(\sin^2(2\theta_k)S_k^2 - 2\sin(2\theta_k)\cos(2\theta_k)S_kT_k + \cos^2(2\theta_k)T_k^2)] \\
&= 4\cdot\Big[\mathbb{E}[\sin^2(2\theta_k)]\mathbb{E}[S_k^2] - \mathbb{E}[\sin(4\theta_k)]\mathbb{E}[S_kT_k] + \mathbb{E}[\cos^2(2\theta_k)]\mathbb{E}[T_k^2]\Big] \\
&= 4\cdot\frac{1}{2}\cdot\mathbb{E}[S_k^2] - 0 + 4\cdot\frac{1}{2}\cdot\mathbb{E}[T_k^2] \\
&= 2\cdot\mathbb{E}[S_k^2 + T_k^2]
\end{aligned} \tag{B4}$$

where we used linearity and the identities described above. Since in the expansion of S_k^2, T_k^2 the expectation of any single sine or cosine is zero, only the square of the terms survive. This yields:

$$\begin{aligned}
\mathbb{E}[(\partial_k J(\boldsymbol{\theta}))^2] &= \mathbb{E}\left[\frac{2}{\cos^2(2\theta_k)}\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b}^2 \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\notin K}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos^2(2\theta_j)\prod_{H_j\in K} -\sin^2(2\theta_j)\right]\right] \\
&\quad + \mathbb{E}\left[\frac{2}{\sin^2(2\theta_k)}\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b}^2 \sum_{K\in\mathcal{K}_{(a,b)}\ s.t.\ H_k\in K}\left[\prod_{H_j\in\mathcal{C}_{(a,b)}-K}\cos^2(2\theta_j)\prod_{H_j\in K} -\sin^2(2\theta_j)\right]\right] \qquad \text{(B5)} \\
&= 2\cdot\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b}^2 \sum_{K\in\mathcal{K}_{(a,b)}}\left(\frac{1}{2}\right)^{|\mathcal{C}_{(a,b)}|-1} \qquad \text{(B6)} \\
&= 4\cdot\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b}^2\cdot\frac{|\mathcal{K}_{(a,b)}|}{2^{|\mathcal{C}_{(a,b)}|}} \qquad \text{(B7)}
\end{aligned}$$

where from (B5) to (B6) we used the identities $\mathbb{E}[\sin^2(2\theta_j)] = \mathbb{E}[\cos^2(2\theta_j)] = \frac{1}{2}$ as described above, as well as noticing that since each θ_j are independent, each $H_j \in \mathcal{C}_{(a,b)}$ contributes a factor of $\frac{1}{2}$ in the expectation except for H_k, as the values corresponding to H_k are cancelled out. From (B6) to (B7), we applied simple counting and re-arranging, yielding the relatively simple form in (B7).

As such, whether the variance of the gradient vanishes exponentially depends on the quantity $|\mathcal{K}_{(a,b)}|/2^{|\mathcal{C}_{(a,b)}|}$. For the classical ansatz (10), we see that $|\mathcal{C}_{(a,b)}| = 2$ and $|\mathcal{K}_{(a,b)}| = 1$, so that the variance is simply $\sum_{(a,b)\in\mathsf{Cut}(H_k)} w_{a,b}^2 > \mathcal{O}(1)$. Consequently, the classical ansatz (10) does not exhibit barren plateaus. For arbitrary $\mathbb{X}$-ansätze, it remains an open question how the scaling of $|\mathcal{K}_{(a,b)}|$ compares to that of $2^{|\mathcal{C}_{(a,b)}|}$. We remark that while the quantity $|\mathcal{C}_{(a,b)}|$ can be computed in linear time, determining $|\mathcal{K}_{(a,b)}|$ constitutes a bottleneck in evaluating the objective function (A12). In fact, determining the set $\mathcal{K}_{(a,b)}$ can be shown to be NP-hard: for any $\mathbb{X}$-ansatz element $H_j = \bigotimes_{i\in S_j} X_i$ for $S_j \subset V$, represent H_j as an n-bit binary string $\boldsymbol{x}_j$ with $x_{j,k} = 1$ if and only if $k \in S_j$. Then, the condition $\oplus\{H \in K\} = \emptyset$ is equivalent to the condition $\oplus\{\boldsymbol{x}_j \in K\} = \mathbf{0}$ for some subset $K \subset \mathcal{C}_{(a,b)}$. This is a well-known problem referred to as computing the minimum distance of a binary linear code, which was shown to be NP-hard in [71], thus showing that determining $\mathcal{K}_{(a,b)}$ is also NP-hard.

Thus, we see that determining the existence of barren plateaus in $\mathbb{X}$-ansätze is related to whether the objective function can be evaluated efficiently on a classical computer.

Appendix C: Comparison between the GW algorithm and BFGS for solving MaxCut using the ansatz (10)

Here, we numerically compare the effectiveness of the classical ansatz (10), consisting of local rotations around X only, and the GW algorithm for solving MaxCut. In particular, we compare the approximation ratios α_{grad} obtained from solving (9) using BFGS and the classical ansatz (10) against the approximation ratios α_{GW} obtained from the GW algorithm. In all simulations we used the scipy optimizer L-BFGS-B [72] with hyperparameters gtol $= 10^{-6}$ and ftol $= 10^{-5}$. We used the GW algorithm implemented in the python package cvxgraphalgs [73].

Fig. 6 shows the ratio $\alpha_{grad}/\alpha_{GW}$ as a function of the vertex degree of a randomly chosen k-regular graph with $n = 30$ (red), $n = 50$ (blue), and $n = 70$ (green) vertices, for edge weights $w_{a,b} \in [0, 5]$.

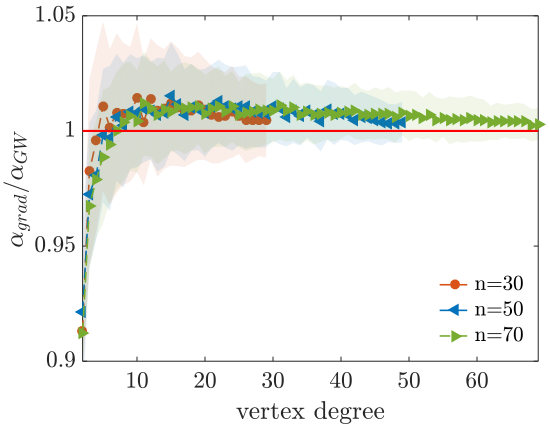

Figure 6. Comparison between the performance of the classical ansatz (10) coupled with a BFGS algorithm and the performance of the GW algorithm. The ratio between the two corresponding approximation ratios α_{grad} and α_{GW} is shown as a function of the vertex degree of a k-regular graph for $n = 30, 50, 70$. Each data point shows the average value of $\alpha_{grad}/\alpha_{GW}$ taken over 200 different graph realizations, and the associated shaded area shows the standard deviation. The straight red line indicates when BFGS and (10) performs better, on average, than the GW algorithm.