Exploring the optimization landscape of variational quantum algorithms

by

Raimel A. Medina Ramos

July, 2024

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Prof. Onur Hosten, Chair
Prof. Maksym Serbyn
Prof. Mikhail Lemeshko
Prof. Marcus Huber

The thesis of Raimel A. Medina Ramos, titled *Exploring the optimization landscape of variational quantum algorithms*, is approved by:

Supervisor: Prof. Maksym Serbyn, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. Mikhail Lemeshko, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. Marcus Huber, Atominstitut, Technische Universitat Wien, Vienna, Austria

Signature: _____

Defense Chair: Prof. Onur Hosten, ISTA, Klosterneuburg, Austria

Signature: _____

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Raimel A. Medina Ramos
July, 2024

Signed page is on file

Abstract

Can current quantum computers provide a speedup over their classical counterparts for some kinds of problems? In this thesis, with a focus on ground state search/preparation, we address some of the challenges that both quantum annealing and variational quantum algorithms suffer from, hindering any possible practical speedup in comparison to the best classical counterparts.

In the first part of the thesis, we study the performance of quantum annealing for solving a particular combinatorial optimization problem called 3-XOR satisfability (3-XORSAT). The classical problem is mapped into a ground state search of a 3-local classical Hamiltonian H_C. We consider how modifying the initial problem, by adding more interaction terms to the corresponding Hamiltonian, leads to the emergence of a first-order phase transition during the annealing process. This phenomenon causes the total annealing duration, T, required to prepare the ground state of H_C with a high probability to increase exponentially with the size of the problem. Our findings indicate that with the growing complexity of problem instances, the likelihood of encountering first-order phase transitions also increases, making quantum annealing an impractical solution for these types of combinatorial optimization problems.

In the second part, we focus on the problem of barren plateaus in generic variational quantum algorithms. Barren plateaus correspond to flat regions in the parameter space where the gradient of the cost function is zero in expectation, and with the variance decaying exponentially with the system size, thus obstructing an efficient parameter optimization. We propose an algorithm to circumvent Barren Plateaus by monitoring the entanglement entropy of k-local reduced density matrices, alongside a method for estimating entanglement entropy via classical shadow tomography. We illustrate the approach with the paradigmatic example of the variational quantum eigensolver, and show that our algorithm effectively avoids barren plateaus in the initialization as well as during the optimization stage.

Lastly, in the last two Chapters of this thesis, we focus on the quantum approximate optimization algorithm (QAOA), originally introduced as an algorithm for solving generic combinatorial optimization problems in near-term quantum devices. Specifically, we focus on how to develop rigorous initialization strategies with guarantee improvement. Our motivation for this study lies in that for random initialization, the optimization typically leads to local minima with poor performance. Our main result corresponds to the analytical construction of index-1 saddle points or transition states, stationary points with a single direction of descent, as a tool for systematically exploring the QAOA optimization landscape. This leads us to propose a novel greedy parameter initialization strategy that guarantees for the energy to decrease with an increasing number of circuit layers. Furthermore, with precise estimates for the negative Hessian eigenvalue and its eigenvector, we establish a lower bound for energy improvement following a QAOA iteration.

Acknowledgements

As I stand on the verge of completing this long journey that started 5 years ago, I owe a tremendous debt of gratitude to far too many people. I would like to dedicate a few words to those who have impacted me in one way or another, be it professionally, personally, or both.

I wholeheartedly thank my supervisor, Maksym Serbyn, for accepting my request to work with him, for his outstanding physical intuition, and his contagious passion for physics. His willingness to help and patience that exceeds understanding have been invaluable. Without his support, which allowed me to freely pursue my research interests, this work would not have been possible. I'm also grateful to Richard Kueng for his crucial insights on our projects, shared through numerous enlightening conversations over the years. Working with someone as talented as him has been a genuine privilege.

The time I spent within my office's walls would not have been the same without my colleagues and good friends Pietro and Stefan. For the countless coffees, and the long discussions about life, happiness, research, ambitions, and more, I deeply thank you both. Similarly, I want to thank Alex and Marko for teaching me so many things about physics, numerical methods, and coding in general. I also want to thank all the past and present group members with whom I have shared enjoyable times. Lastly, I wish to express my appreciation to all the fellow scientists I've had the pleasure of meeting and interacting with during my time at ISTA.

Most of the research presented in this thesis has been funded by the European Research Council, under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 850899). I also want to acknowledge the help and patience I received from Alois, the Scientific Service Units (SSUs), and Caro from the A2P office at ISTA.

Since there is, thankfully, more to life than research, I want to thank all the people and friends I'm grateful to have met during all these years in Austria. To Volker and Jo, to Alex and Maria, to Mel and Luis, and Ruben, I wholeheartedly thank you all. I'm thankful for "el Carva," "el Truji," and "el Willy," my friends from back home. Even though we met in Argentina, it feels like they're going to be with me forever, making this journey so much better. I also want to thank Marcelo and Lidia, "el Padrino y la Madrina," for all the endless love, support, and encouragement. Last, but not least, I want to thank my friends from Cuba, Jorge, and "el grupo de los jodidos": Javier, Oscarin, and Ruly. You guys truly know!

To my parents, Raúl and Laritza. A person's upbringing begins in the family, and I do not doubt that without them, I could never have gotten to where I am now. To them, I am eternally and wholly grateful. I hope this achievement, in some way, repays all the effort and sacrifice they have made for me. "Gracias por todo, mamá y papá."

To my brother, Riguito, "mi bito" for giving me endless smiles.

Finally, I want to dedicate this last, personal paragraph to my wife, "mi tilla". I don't think I dominate English well enough to express how lucky I feel to have you by my side. These last

years have been, in many aspects, the most intense in my life. Even in the toughest moments, I did not waver because you were there with me. Thank you for all your love.

About the Author

Raimel A. Medina started his B.Sc in Physics at the Universidad Central Marta Abreu de las Villas in Cuba. He then moved to the Instituto Balseiro in Argentina to finish his degree and to pursue an M.Sc in Physics. After that, Raimel moved to ISTA in February 2019 as an intern at Prof. Serbyn's group where he became a Ph.D. student in September 2019. During his Ph.D. studies, Raimel published four peer-reviewed papers in PRB, PRA, and PRXQ as well as a preprint, and presented the work at conferences and invited talks. Lastly, in the period of September to December of 2022, Raimel joined Pasqal for a research internship to work on variational algorithms for solving differential equations.

List of Collaborators and Publications

Raimel Medina and Maksym Serbyn. Duality approach to quantum annealing of the 3-variable exclusive-or satisfiability problem (3-XORSAT). *Phys. Rev. A*, 104:062423, Dec 2021

Raimel Medina, Romain Vasseur, and Maksym Serbyn. Entanglement transitions from restricted boltzmann machines. *Phys. Rev. B*, 104:104205, Sep 2021

Stefan H. Sack, Raimel A. Medina, Alexios A. Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3:020365, Jun 2022

Stefan H. Sack, Raimel A. Medina, Richard Kueng, and Maksym Serbyn. Recursive greedy initialization of the quantum approximate optimization algorithm with guaranteed improvement. *Phys. Rev. A*, 107:062404, Jun 2023

Raimel A. Medina and Maksym Serbyn. A Recursive Lower Bound on the Energy Improvement of the Quantum Approximate Optimization Algorithm. *arXiv*, 2405.10125, May 2024

Table of Contents

List of Figures

List of Algorithms

CHAPTER 1

Introduction

1.1 Quantum mechanics and a new form of computing

Richard Feynman famously stated, "Nature isn't classical, dammit, and if you want to make a simulation of nature, you'd better make it quantum mechanical..." This highlights a crucial insight: everything, at its core, is built from atoms—nuclei and electrons operating under quantum mechanics. While the peculiarities of the quantum realm may not be immediately evident, a deeper examination uncovers that the influence of quantum mechanics is pervasive in our everyday technology. Particularly, without our quantum understanding of the solid state physics and band theory of metals, insulators, and semiconductors, the semiconductor industry, with its foundational transistors and integrated circuits, would not have blossomed. Likewise, the vast fields of quantum optics and lasers underpin industries ranging from optical communications to the digital arts, showcasing their basis in quantum technologies.

Given that quantum mechanics underlies all, it is logical to imagine information storage on individual atoms, electrons, or photons, urging us to rethink information beyond the traditional binary system. We should instead contemplate the consequences of media's quantum nature on information storage and processing. The emerging field of quantum information theory is teeming with ongoing explorations, marked by numerous breakthroughs and advances each year.

In this introduction, we provide a brief overview of the field of quantum computing with a focus on variational algorithms for the goal of optimization and simulation of physical systems. We first review the early days of quantum computing, emphasizing known quantum algorithms with provable speedup over their classical counterparts. We next discuss the present capabilities of existing devices and focus on recent efforts on variational quantum algorithms for optimization, machine learning, and simulation of physical systems. We finish with an overview of the thesis and a discussion on relevant challenges and open questions.

1.1.1 The past: Early developments in quantum algorithms

Quantum computation (QC) originated with Benioff's proposals for quantum Turing machines [Ben80, Ben82], and Feynman's ideas for circumventing the difficulty of simulating quantum mechanics by classical computers [Fey82]. This led to Deutsch's proposal for universal QC in terms of what has become the "standard" model: the circuit, or gate model of QC [DP89].

Quantum algorithms then emerged for solving oracle problems, such as Deutsch's algorithm in 1985 [DP85], the Bernstein–Vazirani algorithm in 1993 [BV93], and Simon's algorithm in 1994 [Sim94]. Even though these algorithms did not solve practical problems, they demonstrated mathematically that one could gain more information by querying a black box with a quantum state in superposition, sometimes referred to as quantum parallelism. Arguably, the drive for quantum computing took off in 1994 when Peter Shor, building on these previous results, provided an efficient quantum algorithm for finding prime factors of composite integers, rendering most classical cryptographic protocols unsafe [Sho94, Sho97].

Shortly after, in 1996, Grover's algorithm established a quadratic quantum speedup $O(\sqrt{N})$ for the widely applicable unstructured search problem [Gro96], which typically require $O(N)$ time, with N the size of the problem. Although this quantum algorithm does not change the complexity class it still provides significant speed-up for large N. That same year, Seth Lloyd proved that quantum computers could simulate quantum systems without the exponential overhead present in classical simulations [Llo96], validating Feynman's 1982 conjecture.

Since then, the study of quantum algorithms has matured as a sub-field of quantum computing with applications in search and optimization, machine learning, simulation of quantum systems, and cryptography. For a more detailed overview, see [Mon16, DMB+23].

1.1.2 The present: NISQ era and the search for a quantum advantage

It took roughly 40 years after Feynman's groundbreaking idea, to be in a state where the current quantum devices can provide useful solutions to hard quantum problems. The reason behind this lies in that the implementation of quantum algorithms requires that the minimal quantum information units, qubits, are as reliable as classical bits. Qubits need to be protected from environmental noise that induces decoherence but, at the same time, their states have to be controlled by external agents. This control includes the interaction that generates entanglement between qubits and the measurement operation that extracts the output of the quantum computation.

Eventually, we expect to be able to protect quantum systems and scale up quantum computers using the principle of quantum error correction (QEC) [Got09]. Unfortunately, the overhead of QEC in terms of the number of qubits is, at the present day, still far from current experimental capabilities. To achieve the goal of fault-tolerant quantum computation, the challenge is to scale up the number of qubits with sufficiently high qubit quality and fidelity in operations such as quantum gate implementation and measurement.

Most of the originally proposed quantum algorithms require millions of physical qubits to incorporate these QEC techniques successfully. Existing quantum devices nowadays contain on the order of 100 physical qubits and they are sometimes denoted as "noisy intermediate-scale quantum (NISQ)" devices [Pre18], meaning their qubits and quantum operations are not quantum error corrected and, therefore, imperfect. For example, on these devices, two-qubit gates have an error rate of $\sim 1\%$, while the errors for single-qubit gates are $\sim 0.1\%$. This in turn, severely limits the number of gates that we can coherently apply to current quantum hardware.

One of the goals in the NISQ era is to extract the maximum quantum computational power from current devices while developing techniques that may also be suited for the long-term goal of fault-tolerant quantum computation. In the next two sections, we will revise the paradigm

of adiabatic quantum computing (AQC), and variational quantum algorithms using the gate model of quantum computing, with a focus on optimization.

On this note, it is important to comment that it is not expected that quantum computers will be able to solve efficiently worst-case instances of nondeterministic-polynomial-time (NP) hard problems [MM09] like combinatorial optimization problems. However, it is conceivable (though not proven so far) that quantum devices will be able to find better approximate solutions or find such approximate solutions parametrically faster.

1.2 Adiabatic Quantum Computing

In adiabatic quantum computing (AQC) the computation proceeds from an initial Hamiltonian whose ground state is known and easy to prepare, to a final Hamiltonian whose ground state encodes the solution to the computational problem. The adiabatic theorem guarantees that the system will track the instantaneous ground state provided the Hamiltonian varies sufficiently slowly. It turns out that this approach to QC has deep connections to condensed matter physics, computational complexity theory, and heuristic algorithms. It is important to remark that even though AQC is based on an idea that is quite distinct from the circuit or gate-based model since in the latter a computation may in principle evolve in the entire Hilbert space and is encoded into a series of unitary quantum logic gates, it has been shown that both paradigms are equivalent [AvDK$^+$07]. In other words, AQC and all other models for universal quantum computation can simulate one another with at most polynomial resource overhead. For more details on the topic, see reviews [AL18, LMSS15, HKL$^+$20].

Focusing on combinatorial optimization, there the classical problem is embedded in a "cost Hamiltonian" H_C which is diagonal in the computational basis

$$H_C = \sum_n E_n \left|n\right\rangle \left\langle n\right| . \tag{1.1}$$

Being diagonal in the computational basis the problem Hamiltonian can be written in terms of the action of σ^z and I operators only. Thus we can expand:

$$H_C = hI + \sum_i h_i \sigma_i^z + \sum_{ij} h_{ij} \sigma_i^z \sigma_j^z + \sum_{ijk} h_{ijk} \sigma_i^z \sigma_j^z \sigma_k^z + \dots \tag{1.2}$$

One then considers a time-dependent Hamiltonian which extrapolates between an "initial Hamiltonian" H_B (also called "driver/mixing Hamiltonian") and the cost Hamiltonian according to a predetermined schedule

$$H(t) = f(t)H_B + g(t)H_C, \tag{1.3}$$

where $f(0) = g(T) = 1$ and $f(T) = g(0) = 0$ and T is the duration of computation. The adiabatic algorithm works as follows [FGGS00]: the initial state is prepared to be the ground state of H_B. If the duration of computation T is long enough, the adiabatic theorem guarantees that the system will stay arbitrarily close to the instantaneous ground state at all times. The ground state of the final Hamiltonian encodes the solution of the optimization problem and can be read via measurements on the computational basis.

The runtime T of an adiabatic algorithm scales at worst as $1/\Delta^3$, where Δ is the minimum eigenvalue gap between the ground state and the first excited state of the Hamiltonian $H(t)$.

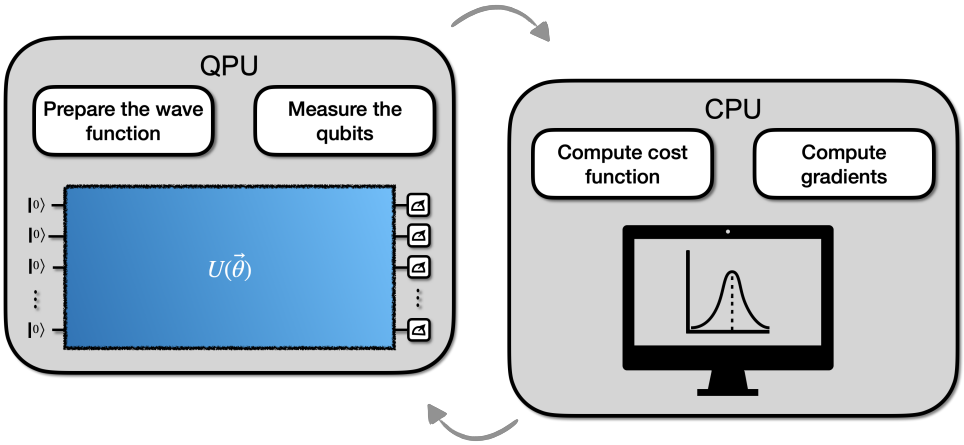

Figure 1.1: Illustration of a generic VQA. The Quantum Processing Unit (QPU) is used to implement the parameterized quantum state $|\psi(\vec{\theta})\rangle = U(\vec{\theta})|0\rangle$ and to measure the qubits in the computational basis. The output from the QPU is fed back to the Classical Processing Unit (CPU) to compute the value of the cost function as well as the gradient of the parameters. The arrows indicate the iterative nature of this process.

If the Hamiltonian is varied sufficiently smoothly, the runtime can be improved to $O(1/\Delta^2)$ up to a polylogarithmic factor in Δ [EH12]. Because of the dependence of the run time on the gap, the performance of quantum adiabatic algorithms is strongly influenced by the type of quantum phase transition the same system would undergo in the thermodynamic limit [vDMV01, FGGN05, LO04]. In Chapter. 2, we present the progress we made on this issue, applied to the particular case of the 3-XOR satisfiability problem [Sch78] with a focus on instances with a highly degenerate ground state manifold.

1.3 Variational Quantum Algorithms

Most of the current NISQ algorithms harness the power of quantum computers in a hybrid quantum-classical arrangement. Such algorithms delegate the classically difficult part of a computation to the quantum computer and perform the other on a sufficiently powerful classical device. These algorithms variationally update the variables of a parameterized quantum circuit and hence are referred to as variational quantum algorithms (VQA) [CRO$^+$19, CAB$^+$21, BCLK$^+$22], see Fig. 1.1 for high-level illustration of a generic VQA. This approach has the added advantage of keeping the quantum circuit depth shallow and hence mitigating noise, in contrast to quantum algorithms developed for the fault-tolerant era.

The first proposals of VQA were the variational quantum eigensolver (VQE) [PMS$^+$14, MRBAG16], originally proposed to solve quantum chemistry problems, and the quantum approximate optimization algorithm (QAOA) [FGG14], proposed to solve combinatorial optimization problems. The variational hybrid approach has seen a wide range of proof of concept applications on NISQ devices ranging from quantum chemistry [KMT$^+$17, Aru20] to quantum optimization [Har21a, LHA$^+$20] and quantum machine learning [HCT$^+$19, JDM$^+$21]. In this work, however, we will only focus on finding ground states of chemistry Hamiltonians using the VQE, and solving classical combinatorial optimization problems using the QAOA, both of which we introduce below.

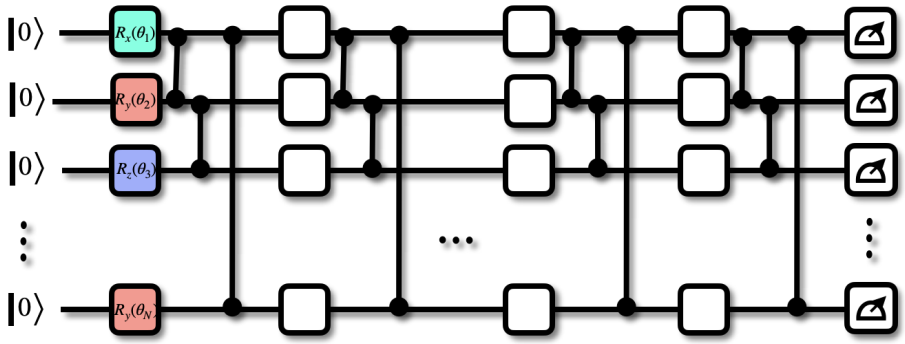

Figure 1.2: Illustration of a Hardware Efficient Ansatz circuit. Single qubit gates correspond to rotation gates around the X, Y, and Z axis, while CZ are used as entangling gates

1.3.1 Variational Quantum Eigensolver

The aim of the VQE, initially introduced by [PMS+14], is to estimate the ground state energy E_0 of a molecule. The interactions of the system are encoded in a Hamiltonian H, usually expressed as a linear combination of simple operators h_k with coefficients c_k. Taking the Hamiltonian H as input, here one defines the cost function $E(\theta) = \langle\psi_0|U^\dagger(\theta)HU(\theta)|\psi_0\rangle$, where $|\psi_0\rangle$ is the initial state that is typically assumed to be a product state. According to the Rayleigh-Ritz variational principle, the cost is meaningful and faithful as $E(\theta) > E_0$, with the equality holding if $|\psi(\theta)\rangle$ is the ground state $|E_0\rangle$ of H.

Employing problem-inspired ansatzes for the ground state search in quantum chemistry systems has proven to be a promising approach due to their capability to ensure rapid convergence towards an optimal solution. This is exemplified by the unitary coupled cluster (UCC) method [TB06], which refines the Hartree-Fock approximation by accounting for quantum correlations. Nonetheless, despite their theoretical appeal, these ansatzes pose a challenge for current quantum devices, primarily due to the requisite deep circuits for their implementation [MBB+18]. Consequently, hardware-efficient ansatzes [KMT+17] have gained prominence as a more practical solution for near-term quantum devices, balancing between implementational feasibility and the ability to capture essential quantum characteristics of the system.

The quantum circuit of a hardware-efficient ansatz with p layers is usually given by a unitary circuit $U(\boldsymbol{\theta})$ [KMT+17]

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{p} W_l \left(\prod_{i=1}^{N} R_l^i(\theta_l^i) \right), \tag{1.4}$$

where $\theta_l^i \in [-\pi, \pi)$ are pN variational angles, concisely denoted as $\boldsymbol{\theta}$. In this expression, the product goes over the spatial dimension $i = 1, \ldots, N$ that labels individual qubits and the "time dimension", $l = 1, \ldots, p$, with p specifying the number of layers, see Fig. 1.2. We choose the single-qubit gates to be rotations $R_l^i(\theta_l^i) = \exp\left(-\frac{i}{2}\theta_l^i G_{l,i}\right)$ with random directions given by $G_{l,i} \in \{\sigma^x, \sigma^y, \sigma^z\}$. W_l is an entangling layer that usually consists of two-qubit entangling gates represented by nearest-neighbor controlled-Z (CZ) or controlled-NOT (CNOT) gates (see Fig. 1.2), depending on the type of hardware used and the corresponding set of native gates.

The VQE has been the subject of extensive theoretical and experimental scrutiny, with numerous adaptations and enhancements proposed in the literature [PMS+14, KMvB+19]. A critical point to consider is that, despite its promising experimental realizations, VQE has not yet

surpassed the performance of the best classical algorithms available. Moreover, unlike the quantum approximate optimization algorithm (QAOA)—which we will discuss in the following section—VQE does not provide analytical performance guarantees, even under ideal conditions. In real-world applications, both QAOA and VQE are confronted with trainability challenges, a topic that will be central to the discussions in this work.

1.3.2 Quantum Approximate Optimization Algorithm

The QAOA was first introduced by Farhi et al. [FGG14] as a near-term algorithm for approximately solving classical combinatorial optimization problems. The first application of the algorithm was for solving the maximum-cut MaxCut problem. MaxCut is an important combinatorial optimization problem with applications in diverse fields such as theoretical physics and circuit design. MaxCut seeks to partition a given (un)weighted graph $\mathcal{G}$ into two groups such that the number of edges $n_{\mathcal{E}}(\mathcal{G})$ (or the sum of their weights, for weighted problems) that connect vertices from different groups are maximized, see Fig. 1.3(a) for an example. Finding the MaxCut for a graph with N vertices is equivalent to finding a ground state for the N-qubit classical Hamiltonian

$$H_C = \sum_{\langle i,j\rangle \in \mathcal{E}} J_{ij}\sigma_i^z\sigma_j^z, \tag{1.5}$$

with the sum running over a set of graph edges $\mathcal{E}$ with weights J_{ij} and σ_i^z being the Pauli-z matrix acting on the i-th qubit. The full spectrum of H_C consists of all product states ordered according to their energies and will be used in what follows as a complete basis, $|E_0\rangle, |E_1\rangle, \ldots, |E_{2^N-1}\rangle$.

The depth-p QAOA algorithm [FGG14], denoted in what follows as QAOA$_p$, minimizes the expectation value of the classical Hamiltonian over the variational state $|\Gamma^p\rangle$ where $\Gamma^p = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ encodes variational angles $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ shown in Fig. 1.3:

$$|\Gamma^p\rangle = \prod_{i=1}^{p} e^{-\beta_i H_B} e^{-\gamma_i H_C} |+\rangle \tag{1.6}$$

Here

$$H_B = -\sum_{i=1}^{N} \sigma_i^x, \tag{1.7}$$

is the mixing Hamiltonian, and the circuit depth p controls the number of applications of the classical and mixing Hamiltonian. The initial product state $|+\rangle = \otimes_{i=1}^{N} |+\rangle_i$, where all qubits point in the x-direction is an equal superposition of all possible graph partitions which is also the ground state of H_B.

Finding the minimum of

$$E(\Gamma^p) = \langle \Gamma^p | H_C | \Gamma^p \rangle \tag{1.8}$$

over angles $(\beta_1, \ldots, \beta_p)$ and $(\gamma_1, \ldots, \gamma_p)$ that form a set of $2p$ variational parameters, $\Gamma^p = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, yields a desired approximation to the ground state of H_C, equivalent to an approximate a solution of MaxCut. The scalar function $E(\Gamma^p)$ thus defines a $2p$-dimensional energy landscape where the global minimum yields the best set of QAOA parameters. The performance of the QAOA is typically reported in terms of how close is the approximation ratio to one,

$$1 - r(\Gamma^p) = \frac{E_0 - E(\Gamma^p)}{E_0}, \tag{1.9}$$

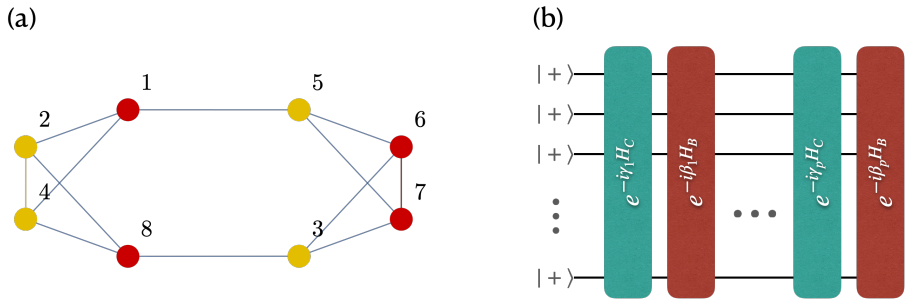

Figure 1.3: (a) Example of a maximum-cut solution for a 3-regular graph composed of 3 vertices. (b) Illustration of the QAOA circuit with p layers. Each layer is composed of a unitary rotation $U_C(\gamma_i) = e^{-i\gamma_i H_C}$ with the cost Hamiltonian H_C, followed by another unitary operator $U_B(\beta_i) = e^{-i\beta_i H_C}$ with the mixing Hamiltonian H_B.

where E_0 is the ground state of the classical Ising Hamiltonian (1.5) assumed to be unique for simplicity. From here we see that a decrease in $1 - r$ implies that the expectation value of the cost function is approaching the ground state energy of classical Hamiltonian.

In this work, we restrict our attention to MaxCut on 3-regular graphs, where every vertex is connected to exactly 3 other vertices. In this regard, it is known to be NP-hard to design an algorithm that guarantees a minimum approximation ratio of $r_\star \geq 16/17$ on MaxCut for all graphs [Hůl01], or $r_\star \geq 331/332$ when restricted to unweighted 3-regular graphs [BK99]. The best classical algorithms to date give the approximation ratio of $r_\star \approx 0.8786$ for general graphs [GW95], and $r_\star = 0.9326$ for unweighted 3-regular graphs [HLZ04] using semidefinite programming. While QAOA for $p = 1$, with an approximation ratio of 0.692 for unweighted 3-regular graphs, does not outperform its classical counterparts for the MaxCut problem, QAOA has been found to surpass the Goemans-Williamson bound for larger values of p [Cro18]. This result is however purely heuristic and it remains to be shown if it holds beyond the system sizes that can be simulated classically.

In summary, both heuristic [Cro18, ZWC+20, SS21a, SMKS23] and experimental results [Har21a, WVG+22, WSW24, E+22] give hope that the QAOA might be a promising algorithm for achieving a quantum advantage on real quantum hardware in the near future.

1.4 Challenges for Variational Quantum Algorithms

The progress of variational quantum algorithms (VQAs) also reveals significant challenges. Addressing these is vital to reach quantum advantage with scalable devices. A deep understanding of VQA limits is key to creating improved algorithms, ensuring reliable performance, and advancing quantum hardware design.

As for any variational approach, the success of a variational quantum algorithm (VQA) depends on the efficiency and reliability of the ansatz and the optimization method used. The choice of ansatz determines what kind of quantum states the parameterized quantum circuit can effectively prepare. Thus, given the shallow-depth nature of present devices, it is important to design smart ansatze that take advantage of the specific details (for example, presence of symmetries) of the problem of interest. When investigating the VQE, we will use the hardware-efficient ansatz, which is known to be universal. This means that for any arbitrary state $|\psi\rangle$, there exists a finite-depth p circuit that can prepare it exactly, with p possibly scaling exponentially with the Hilbert space dimension.

Classical optimization in VQAs is complex, as it tends to be NP-hard due to the presence of many local minima in the cost function "landscape" [SJAG19]. Additionally, training VQAs faces the stochastic nature of quantum measurements, hardware noise, and barren plateaus—regions in parameter space where gradients are near zero and hinder optimization. These challenges, coupled with the exponential increase of local minima with parameter count, are critical to addressing for VQAs to be successful. In this work, we present the progress that we made on these issues and discuss the impact of our results, leading to a better understanding of the capabilities and limitations of VQAs.

1.5 Overview of the thesis

1.5.1 Contents of Chapter 2

In Chapter 2, we evaluate the effectiveness of adiabatic quantum computing on the 3-XOR satisfiability problem when new interactions are systematically introduced. Through a duality transformation, we analyze both analytically and numerically how phase transitions arise with increasing problem complexity. Moreover, the discussed duality transformation enables the exploration of problem instances with a highly degenerate ground state manifold. Our findings indicate that first-order phase transitions can occur, potentially causing the annealing time to scale exponentially with system size and thus making quantum annealing impractical for these optimization tasks. The content of this Chapter is based on the published work [MS21].

1.5.2 Contents of Chapter 3

Chapter 3 investigates the emergence of barren plateaus (BPs) in generic variational quantum algorithms and introduces a strategy to mitigate this issue by monitoring local entanglement during classical optimization. We employ classical shadow tomography, a method efficient in estimating local observable expectations, to detect and navigate around regions with negligible gradients. This enhances the optimization trajectory of VQAs. Additionally, our findings emphasize the importance of tailored initialization strategies that exploit problem-specific features. The content of this Chapter is based on the published work [SMM+22].

1.5.3 Contents of Chapter 4

In Chapter 4 we focus on the quantum approximate optimization algorithm. We introduce a greedy initialization of QAOA which guarantees improving performance with an increasing number of layers. Our main result is an analytic construction of $2p+1$ *transition states* — saddle points with a unique negative curvature direction — for QAOA with $p+1$ layers that use the local minimum of QAOA with p layers. Transition states connect to new local minima, which are guaranteed to lower the energy compared to the minimum found for p layers. We use the Greedy procedure to navigate the exponentially increasing with p number of local minima resulting from the recursive application of our analytic construction. The performance of the Greedy procedure matches available initialization strategies while providing a guarantee for the minimal energy to decrease with an increasing number of layers p. The content of this Chapter is based on the published work [SMKS23].

1.5.4 Contents of Chapter 5

In Chapter 5 building on the results from Chapter 4 we obtain insights into the large-depth regime of the QAOA using an analytic expansion of the cost function around the so-called transition states. We construct an analytic estimate of the negative Hessian eigenvalue and corresponding eigenvector at each transition state, which enables us to obtain an analytical lower bound on the improvement of the cost function, and to reduce the cost of optimization by bypassing the need to construct and diagonalize the Hessian matrix. Finally, we numerically verify the accuracy of our estimates. Although the obtained energy lower bound underestimates the improvement of the cost function, we find it shows an exponential decrease with the number of layers p, similar to the heuristically observed behavior. The content of this Chapter is based on the preprint [MS24].

CHAPTER 2

Duality approach to quantum annealing of the 3-XORSAT problem

In this Chapter, we investigate the performance of quantum annealing on two specific instances of the 3-XOR satisfiability problem. We investigate how the performance of the algorithm, as captured by the presence of first/second order quantum phase transitions, is affected as the number of interaction terms in the classical Hamiltonian is increased. This Chapter is based on the paper:

Raimel Medina and Maksym Serbyn. Duality approach to quantum annealing of the 3-variable exclusive-or satisfiability problem (3-XORSAT). *Phys. Rev. A*, 104:062423, Dec 2021

2.1 Introduction

The quantum adiabatic algorithm [FGGS00], which can be viewed as a generalization of quantum annealing [ACdF89, FGS$^+$94, KN98, BBRA99, ST06], was considered as a perspective quantum algorithm since early days of quantum computing. In this algorithm, the solution of a classically hard combinatorial optimization problem [MM09] is mapped onto a problem of finding a ground state of a classical spin Hamiltonian. Such ground state is in turn obtained by initializing a quantum spin system in a ground state of a simple quantum Hamiltonian and then adiabatically interpolating between the quantum and classical Hamiltonians. The success of this algorithm, which is quantified by the overlap between the final state after the evolution and the ground state, is guaranteed, provided the spectrum features a finite gap throughout the adiabatic evolution, see Refs. [BFK$^+$13, LMSS15, AL18, HKL$^+$20] for recent reviews.

The performance of the algorithm was studied theoretically for several optimization problems [JKSZ10, FGH$^+$12]. Remarkably, in many cases the gap was shown to vanish polynomially or even exponentially in the problem size [JKSZ10, FGH$^+$12], giving evidence of the phase transition encountered in the annealing process. The majority of models studied to date featured a *unique* ground state. While such problems are convenient for numerical studies, in many interesting combinatorial problems one often encounters a degenerate space of solutions. Classical problems with many possible solutions, where some are similar to each other, while others are globally different, are said to be in a "clustering phase" [MRZ02]. Classical optimization problems in the clustering phase correspond to the spin Hamiltonians with *degenerate*

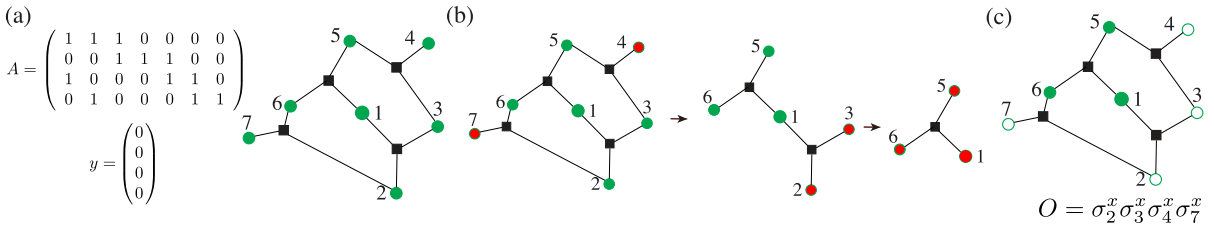

Figure 2.1: (a) Matrix A that specifies 3-XORSAT problem with $N = 7$ variables and $M = 5$ conditions, and corresponding hypergraph where vertices shown by green dots denote spins and black squares are the edges that correspond to three-spin interaction terms. (b) Illustration of leaf removal algorithm that can find the solution to the classical problem. Starting from the original hypergraph in (a) at each step one removes spins that enter in just one interaction (equivalently, are included only in one edge). In the first step, one removes spins 4 and 7. Then we can remove either spins $2, 3$ or spins $5, 6$. In the last step, all three remaining spins can be removed. (c) A simultaneous flip of spins $2, 3, 4, 7$ (white-filled circles) does not change the energy of the system. Such degeneracy corresponds to the operator O that commutes with the classical Hamiltonian.

ground state manifold, a situation often explicitly ruled out in quantum adiabatic algorithm performance studies.

In this work, we specifically focus on classical optimization problems with degenerate space of solutions. To this end, we use the "exclusive-or" satisfiability (XORSAT) problem [Sch78, CDD01] for studies of quantum algorithm performance in the clustering phase. XORSAT is equivalent to a boolean linear algebra problem, hence it is easily verifiable and solvable in satisfiable cases. Restricting to the case where each exclusive or condition involves exactly 3 variables, we obtain the so-called 3-XORSAT problem, which maps onto a classical spin Hamiltonian with three-spin interactions specified by a certain hypergraph. This spin model was studied in the literature, where the existence of clustering phase was established for random hypergraphs ensembles [MRZ02, CDD01].

We focus on particular instances of the 3-XORSAT problem, which provide an example of classically solvable instances, yet feature a large degeneracy in the solutions space. We show that such degeneracy in the solution space can be recast into the emergence of a set of Z_2 conserved charges that persists in the quantum model. To restrict the problem to a particular sector, we generalize the duality introduced in Ref. [FGH$^+$12]. Applying the duality to the spin model on a tree hypergraph results in an Ising-type model, facilitating numerical and analytical understanding. In particular, we establish that the 3-XORSAT model on a tree hypergraph does not feature a phase transition, guaranteeing the success of the quantum adiabatic algorithm. On the other hand, the closure of the tree hypergraph leads to an emergence of the second-order phase transition encountered throughout adiabatic evolution.

The structure of this paper is organized as follows. In Sec. 2.2 we briefly review the 3-XORSAT problem as well as the quantum adiabatic algorithm. In Sec. 2.3 we illustrate the duality mapping using specific instances of the 3-XORSAT problem. For each of these instances, we find the dual Hamiltonian, as well as discuss its energy spectrum and minimal gap dependence with system size. We conclude in Sec. 2.4 with a brief discussion of our results and a summary of interesting directions for future work.

2.2 Classical and quantum 3-XORSAT model

In this section, we introduce the classical 3-XORSAT model and associated spin Hamiltonian. We briefly review the application of the so-called "leaf removal algorithm" [MRZ02] to find the solution of a classical problem and highlight the emergent degeneracy of the classical energy landscape. Finally, we discuss the application of the quantum adiabatic algorithm for finding the ground state of the classical 3-XORSAT model. We show that even though the degeneracy of the classical energy landscape is lifted in the presence of a transverse field, a set of commuting integrals of motion remains.

2.2.1 Classical 3-XORSAT

Classical 3-XORSAT problem [Sch78] consists in finding the arrangements of binary variables $x_1, \dots x_N$ that satisfy the set of M distinct "exclusive-or" (XOR) clauses with only three variables participating in each condition. Using equivalence between XOR operator and binary addition, we can rewrite the XOR clause $x_1 \oplus x_2 \oplus x_3 = b$ where $b = 0, 1$ as $x_1 + x_2 + x_3 = b \, mod \, 2$. This allows us to map a 3-XORSAT problem onto a system of linear equations:

$$A \cdot x = y \pmod 2, \tag{2.1}$$

where A is a $M \times N$ matrix and y is a $M-$component vector with binary entries, $A_{ai} \in \{0, 1\}$, $y_a \in \{0, 1\}$. Since we are restricted to clauses with only three variables, each row of the matrix A contains exactly three ones with all other entries being zero, see example in Fig. 2.1(a). Determining whether the Boolean system of equations (2.1) admits an assignment of the Boolean variables satisfying all the equations constitutes the decision version of the 3-XORSAT problem. In general, one is also interested in the set of solutions and its size. Throughout this work, our focus will be on quantum annealing approach to finding the solution of the XORSAT problem.

The 3-XORSAT problem defined by means of a linear system of equations with N variables and M equations can be naturally mapped to the problem of energy minimization for an ensemble of N classical spins, σ_i^z, with M three-spin interactions [MM09]. Defining $\sigma_i^z = (-1)^{x_i}$ and $J_a = (-1)^{y_a}$ one can demonstrate that solution of Eq. (2.1) corresponds to a zero-energy ground state of the following classical Hamiltonian:

$$H_{\mathsf{c}} = \sum_{\alpha=1}^{M} (1 - J_\alpha \sigma_{i_\alpha}^z \sigma_{j_\alpha}^z \sigma_{k_\alpha}^z). \tag{2.2}$$

In case when the system of equations (2.1) does not admit a solution that satisfies all conditions (it is said to be UNSAT), the ground state of the H_{c} corresponds to a bit assignment that violates the minimal possible number of conditions.

The 3-XORSAT problem and corresponding classical Hamiltonian are fully fixed by the pair of (A, y), or, equivalently the choice of three-spin interactions and a value of couplings, $J_\alpha = \pm 1$. Interactions between spins can be conveniently visualized using the hypergraph, where vertices correspond to spins, and edges (which now join three spins, hence these are in fact hyperedges) correspond to interactions. A particular instance of the 3-XORSAT problem and corresponding hypergraph is illustrated in Fig. 2.1(a).

The hypergraph representation provides a visual way to find the solution to the 3-XORSAT problem. The so-called leaf removal algorithm [PSW96] is illustrated in Fig 2.1(b) and consists of removing the spins that enter only in a single interaction. The insight is that if a given spin,

say σ_7^z, appears in the Hamiltonian only once, e.g. in the term $\sigma_2^z\sigma_6^z\sigma_7^z$ for the chosen example, we can always satisfy the corresponding interaction term by adjusting the value of σ_7^z. Thus we are allowed to erase this spin and the corresponding interaction term. Iterating such search and removal of spins that enter a single interaction term (so-called leaves) on the hypergraph is the essence of the leaf removal procedure. This procedure halts if one removes all vertices and edges as is shown in Fig. 2.1(b). This case corresponds to an instance of the 3-XORSAT problem that is completely solvable by the leaf removal algorithm. Another alternative is when in the process of iterating the leaf removal procedure one fails to find any leaves. The leaf removal algorithm halts at such an instance and the remaining hypergraph is typically dubbed a "glassy core" [MRZ02], see an example of such hypergraph in Fig. 2.4(a).

During the iterative process of the leaf removal algorithm, one may encounter instances when more than two spins participating in a given interaction term are simultaneously removed, see Fig 2.1(b). When such interaction edge and two associated spins are removed a degeneracy emerges. In the example in Fig. 2.1(b) we remove simultaneously σ_2^z and σ_3^z, hence flipping these spins simultaneously does not affect the energy of the given interaction edge. At the level of the full Hamiltonian, such instances lead to an emergence of global degeneracies: in the example that we show the total energy does not change if one flips spins 2, 3, 4, and 7. Depending on the geometry of the problem, one may encounter many such degeneracies with their number being a finite fraction of the total number of spins — this is characteristic of the so-called clustering phase [MRZ02, CDD01]. Some of this degeneracy though originates from the structure of the glassy core, which typically does not have a unique solution (UNSAT) but instead has multiple degenerate ground states.

2.2.2 Solving 3-XORSAT with quantum adiabatic algorithm

One approach to finding the ground state of the classical Hamiltonian (2.2) or, equivalently, to finding the bit assignment that violates the smallest possible number of equations in (2.1) is provided by quantum adiabatic algorithm [FGGS00]. Supposing that classical Hamiltonian (2.2) can be implemented on a quantum simulator, we initialize the system in the ground state of a quantum paramagnet Hamiltonian

$$H_{\mathsf{q}} = -\sum_{i=1}^{N} \sigma_i^x, \tag{2.3}$$

and evolve this state under the following time-dependent Hamiltonian:

$$H_T(t) = (1 - \frac{t}{T})H_{\mathsf{q}} + \frac{t}{T}H_{\mathsf{c}}, \tag{2.4}$$

from time $t = 0$ to T. According to the adiabatic theorem, if T is sufficiently large and H_{q} and H_{c} do not commute with each other, the quantum simulator will remain with high fidelity in the ground state for all times, resulting in the preparation of the ground state of H_{c} at time T.

The running time T, depends on the energy spectrum of $H_T(t)$. In particular, the time required for preparing the ground state with high fidelity is bounded from below by the inverse square of the minimum gap encountered throughout the time evolution, $T \gg \max_t |V_{10}(t)|/[\min_t \Delta(t)]^2$. Here the gap is defined as a difference between the energy of the ground state and the first excited state, $\Delta(t) = E_1 - E_0$, and $V_{10} = \langle 0|\partial_t H(t)|1\rangle$ is the matrix element of the time-dependent part of the Hamiltonian between ground state $|0\rangle$ and first excited state $|1\rangle$. Due to this bound, many theoretical studies of the efficiency of the quantum adiabatic algorithm focus on the behavior of the minimum gap of $H_T(t)$ [vDMV01, FGGN05].

2.2.3 Behavior of gap and degeneracies

The behavior of the gap for the so-called 3-regular 3-XORSAT Hamiltonian, where each spin enters in exactly three interaction terms, was considered previously [JKSZ10, FGH$^+$12]. It was found that the system goes through a first-order quantum phase transition, displaying an exponential decrease of the gap with system size. However, these studies were restricted to the instances of the classical 3-XORSAT problem that do not have any degeneracy in the ground state. These instances are said to have *unique satisfying assignment*, and their consideration simplifies the study of the gap behavior [JKSZ10, FGH$^+$12]. For the 3-XORSAT problem defined on a 3-regular ensemble of random hypergraphs in the $N \to \infty$ these instances form a non-zero fraction (~ 0.285) of the set of all instances [JKSZ10]. Yet, the behavior of instances that have degenerate ground state manifold was not studied.

In this work we (to the best of our knowledge) provide the first results relative to systems with degenerate ground states. We consider instances where degeneracy of the ground state originates from the existence of simultaneous spin flips that do not change the energy of the classical Hamiltonian (see discussion in Section 2.2.1). We note, that the ground state may have additional degeneracy due to the problem being UNSAT, which is not considered here. If simultaneous flipping of spins $\sigma^z_{i_1} \to -\sigma^z_{i_1}, \ldots, \sigma^z_{i_k} \to -\sigma^z_{i_k}$ does not change the energy of the system, the following operator

$$O = \sigma^x_{i_1}\sigma^x_{i_2}\ldots\sigma^x_{i_k}, \tag{2.5}$$

commutes with classical Hamiltonian, $[O, H_{\mathsf{c}}] = 0$. Since the quantum Hamiltonian, H_{q}, contains only σ^x terms, any such operator also commutes with the full $H_T(t)$,

$$[O, H_T(t)] = 0,$$

for any t, thus corresponding to an Abelian Z_2 *symmetry* present in the system. Moreover, as we mentioned above, many typical instances of the 3-XORSAT problem may contain a possibly extensive number of distinct operators $\{O_l\}$, $l = 1, \ldots q$ that commute not only with the Hamiltonian but also among themselves.

The presence of q distinct Abelian symmetries leads to spectral degeneracy only for the classical Hamiltonian, i.e. only for $H_T(t)$ at $t = T$. However, although these symmetries do not give rise to spectral degeneracy when $t < T$, their presence fragments the 2^N-dimensional Hilbert space of model (2.4) into 2^q distinct sectors, each labeled by ± 1 eigenvalues of corresponding O_l operator. The full Hamiltonian assumes block-diagonal form when written in the basis that diagonalizes operators $\{O_l\}$,

$$H_T(t) = \bigoplus_{\alpha=1}^{2^q} H_\alpha(t), \tag{2.6}$$

where α runs over all 2^q blocks.

The unitary evolution preserves the symmetries of the Hamiltonian. This implies that the search for the minimum gap is performed inside the block $H_\alpha(t)$, which contains the initial state, $|\psi(0)\rangle$. Due to the reduced dimensionality of H_α, we can perform exact numerical calculations for a wide range of system sizes.

One of the main results of this work is the *duality transformation* which allows us to explicitly obtain the form of the Hamiltonian $H_\alpha(t)$ restricted to a given sector. In the next section, we introduce this duality transformation using specific examples. This duality allows us to readily study the behavior of the gap even in the presence of extensive degeneracies in the system and understand the fate of the quantum adiabatic algorithm.

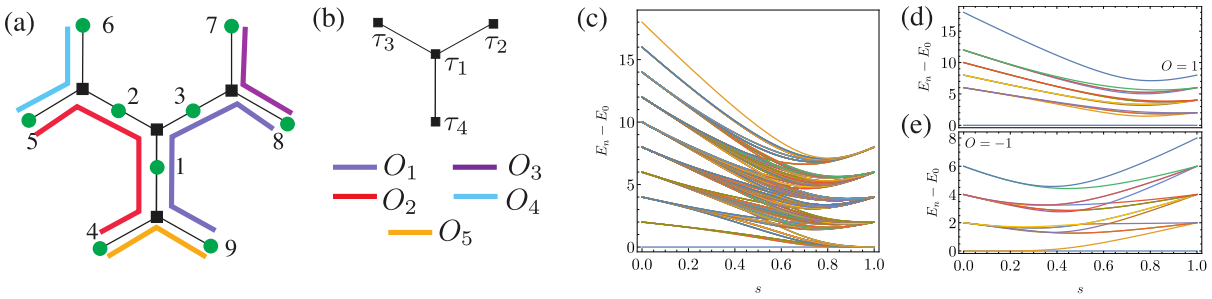

Figure 2.2: (a) Example of the Hushimi tree at the level $g = 2$. A convenient choice of the set of independent conserved quantities is shown by colored lines with different lines corresponding to individual conserved quantities, for instance, $O_1 = \sigma_1^x \sigma_3^x \sigma_8^x \sigma_9^x$. (b) Dual degrees of freedom live on the tree hypergraph with $g - 1$ generations. (c) The evolution of the low-lying spectrum as a function of parameter $s = t/T$ reveals many crossings and large degeneracy in the spectrum of classical Hamiltonian at $s = 1$. (d) Spectrum of dual Hamiltonian in the sector where all charges $O_l = 1$ has only avoided crossings demonstrating that application of duality resolves all symmetries. (e) The spectrum of the dual Hamiltonian in the sector $O_l = -1$, where the dual model has an additional emergent Z_2 symmetry, that is manifested in the degeneracy of ground state manifold of the dual model for small values of s.

2.3 Duality approach to quantum 3-XORSAT model

As discussed above, the duality provides a natural approach to the quantum 3-XORSAT Hamiltonian in the presence of conserved quantities. In this section, we illustrate duality using specific instances of 3-XORSAT model, whereas in the Appendix A.1 we formulate the duality using the language of linear algebra which allows us to apply such transformation to the 3-XORSAT problem on arbitrary graphs in an efficient manner.

2.3.1 Duality for tree hypergraph

The structure of degeneracies in the 3-XORSAT model is determined by its connectivity. While often the 3-XORSAT model is considered on random graphs [JKSZ10, FGH$^+$12], below we consider an instance of the 3-XORSAT problem that is fully solvable by the leaf removal algorithm. In particular, we consider a *tree hypergraph* that may be thought of as a toy example of the structure of the leaves of the generic 3-XORSAT instances. We find that the dual Hamiltonian is an Ising model and obtain that the energy gap remains constant in the thermodynamic limit.

Degeneracies and conserved charges

We consider the 3-XORSAT problem on the tree hypergraphs with connectivity 2 and a varying number of generations. An example of a tree hypergraph shown in Fig 2.2(a) has $g = 2$ generations of spins and contains $N = 3(2^g - 1) = 9$ vertices and $M = 4$ edges. Any such tree hypergraph corresponds to a trivial solvable instance of 3-XORSAT: application of leaf removal algorithm completely removes all vertices and results in a solution.

In the process of a leaf removal iteration, one always encounters pairs of spins that belong to the same edge and are removed simultaneously. As explained in Sec. 2.2.1, this leads to degeneracies. The tree hypergraph with g generations is characterized by $q = 3 \cdot 2^{g-1} - 1$ independent Z_2 charges (by independent Z_2 charges we refer to a minimal set of independent

operators O_l such that any other string of σ^x that commutes with the Hamiltonian can be expressed as a product of some of O_l from this set.). For the particular hypergraph in Fig. 2.2(a) this formula yields $q = 5$ charges, which are shown by different colors in Fig. 2.2. A given symmetry sector can be fixed by specifying the eigenvalues of all these charges. In particular, the ground state of the quantum part of the annealing Hamiltonian, H_{q} in Eq. (2.3), $|\psi(0)\rangle = |\rightarrow \ldots \rightarrow\rangle$ corresponds to the values of all charges $O_l = 1$. We are interested in performing a duality transformation that restricts the Hamiltonian to a particular symmetry sector. Taking into account that the ratio between the number of independent charges and the number of spins q/N tends to the value of $1/2$ in the thermodynamic limit $g, N \to \infty$, the duality is capable of drastically reducing the Hilbert space dimension from 2^N to approximately $2^{N/2}$.

Dual Hamiltonian

We explicitly construct the duality, by defining spins τ that live at the edges of the hypergraph, see Fig. 2.2(b). The τ^x operators are expressed via original spins as:

$$\tau^x_{(ijk)} = \sigma^z_i \sigma^z_j \sigma^z_k, \tag{2.7}$$

where $\tau^x_{(ijk)}$ is the dual-spin located at the edge that was connecting spins (i, j, k). To simplify notations, we label the edges and dual spin operators τ_α by greek indices as in Fig. 2.2(b); for instance, $\tau^x_{\alpha=1} = \tau^x_{(123)} = \sigma^z_1 \sigma^z_2 \sigma^z_3$. This mapping converts the classical Hamiltonian, H_{c} in Eq. (2.2) into the simple sum of τ^x_i operators,

$$\tilde{H}_{\mathsf{c}} = -\sum_{\alpha \in V} J_\alpha \tau^x_\alpha, \tag{2.8}$$

where we omitted a constant term from Eq. (2.2). Tilde emphasizes that this Hamiltonian acts in the Hilbert space of τ-spins and index α runs over all vertices of the dual graph, Fig. 2.2(b), denoted as V.

Similar to duality applied to discrete Abelian gauge theories [Fra13], the relation between the τ^z and σ^x is non-local. The τ^z operators are defined via product of σ^x operators on the path from a certain "root vertex",

$$\tau^z_\alpha = \prod_{m \in \text{path to } \alpha} \sigma^x_m. \tag{2.9}$$

This root vertex is chosen as $i = 9$ in Fig. 2.2(a). Then for the graph in Fig. 2.2(b) we have: $\tau^z_{\alpha=4} = \tau^z_{(149)} = \sigma^x_9$, $\tau^z_1 = \sigma^x_9 \sigma^x_1$, $\tau^z_2 = \sigma^x_9 \sigma^x_1 \sigma^x_3$, and $\tau^z_3 = \sigma^x_9 \sigma^x_1 \sigma^x_2$. This construction will result in the simple expression for original spins, $\sigma^x_1 = \tau^z_1 \tau^z_4$, unless they are located at the boundary of the graph. Thus, for the bulk spins the dual $\tilde{H}_{\mathsf{q}}$ of H_{q} coincides with an Ising model on a tree

However, the situation is different for the boundary spins. To obtain the expression for σ^x_i at the boundary, one must use the existence of the conserved charges. For example, the spin σ^x_4 cannot be expressed via the product of any of the four τ^z_α operators. However, we observe that $\sigma^x_4 = (\sigma^x_4 \sigma^x_9)\sigma^x_9 = O_1 \sigma^x_9 = O_1 \tau^z_4$. Remaining boundary spins σ^x_i with $i = 5, \ldots 8$ can be constructed in a similar way. Dual spin operators $\tau^{x,z}_\alpha$ defined in such way obey the standard Pauli commutation relations, $\{\tau^z_\alpha, \tau^x_\alpha\} = 0$ and $[\tau^z_\alpha, \tau^x_\beta] = 0$ for $\alpha \neq \beta$.

Collecting all terms together and denoting $s = t/T$ we obtain the dual of the full Hamiltonian, Eq. (2.4) as:

$$\tilde{H}_T(s) = -s \sum_{\alpha \in V} J_\alpha \tau^x_\alpha - (1-s) \sum_{\langle \alpha\beta \rangle \in V} \tau^z_\alpha \tau^z_\beta - (1-s) \sum_{\alpha \in \partial V} h^z_\alpha[O] \tau^z_\alpha. \tag{2.10}$$

The first two terms here correspond to the Ising model on a Cayley tree, see Fig. 2.2(b). The last term encodes the dependence of duality on the values of conserved charges and involves only τ-spins at the boundary of the Cayley tree ∂V ($\tau^z_{2,3,4}$ in the present example). The effective symmetry-breaking field coupled to boundary spins reads:

$$h^z_\alpha[O] = (1 + O_{m_\alpha}) \prod_{m \in \text{path from root}} O_m. \tag{2.11}$$

Here O_{m_α} is the charge that involves only two spins, including α, and the product is overall charges encountered on the path from the root. For instance, $h_4[O] = 1 + O_5$, $h_2[O] = (1 + O_3)O_1$ in notation of Fig. 2.2.

Remarkably, the first line in the dual Hamiltonian Eq. (2.10) is the Ising model on the Cayley tree with connectivity equal to three. This part of the dual Hamiltonian has global Z_2 symmetry $\tau^z \to -\tau^z$ and does not depend on the values of conserved charges (signs of J_α can be removed by the relabeling $\tau^x \to -\tau^x$ in the present case). However, in addition, we also have the second line in Eq. (2.10) that imposes a Z_2-symmetry breaking effective field on the dual degrees of freedom at the boundary. The strength of these symmetry-breaking fields depends on the sector of conserved charges as we discuss below.

Energy spectrum and minimal gap of dual Hamiltonian

To illustrate the advantage of describing the system with the dual Hamiltonian, we show the spectrum of the original Hamiltonian Eq. (2.4) as a function of s in Fig. 2.2(c). The low-lying energy levels become highly degenerate at $s = 1$, corresponding to the degeneracy of the ground state manifold of the classical problem. Moreover, we observe multiple level crossings between eigenstates that belong to different symmetry sectors. The level crossings and degeneracy complicate the determination of the minimal gap encountered throughout the adiabatic algorithm.

In comparison, Fig. 2.2(d-e) demonstrates the spectrum of the dual Hamiltonian (2.10) for particular values of conserved charges (also referred to as "sector") has much lower complexity. These energy levels are a subset of energy levels shown in Fig. 2.2(c). The sector of conserved charges is a *property of initial state*. The ground state of the quantum paramagnet $|\rightarrow\rightarrow \ldots \rightarrow\rangle$, is an eigenstate of all O_m operators with eigenvalue $O_m = 1$. Thus, from Eq. (2.11) we obtain a *uniform* magnetic field $h^z_\alpha[O] = 2$ for all α at the boundary of the tree. The presence of this magnetic field leads to a strong breaking of Z_2 symmetry that would be otherwise present in the dual Hamiltonian. Hence, it helps to avoid the second-order phase transition in an Ising model, and Fig. 2.2(d) shows that the finite gap of order one is present for all values of s.

The duality facilitates the determination of the gap on several levels. First, it decreases the number of degrees of freedom and allows us to study the problem in a smaller Hilbert space. Second, it removes the degeneracies and explicitly resolves all symmetries present in the problem, making the extraction of the energy gap more straightforward. As a result, the duality allows us to study the finite-size scaling of the gap for the family of tree hypergraphs with up to $g = 6$ generations with $N = 189$ spins. We use the density-matrix renormalization-group (DMRG) algorithm to obtain the ground state and energy gap as a function of the parameter s. Previous works have studied the transverse field Ising model on the Cayley tree [NFG$^+$08, LvDX12, LSS08] with a global symmetry breaking field. In our study, we apply the DMRG algorithm to an Ising model with symmetry-breaking fields at the boundary, corresponding to the energy spectrum encountered in the adiabatic algorithm launched from

the paramagnetic ground state. The resulting behavior of the gap for different system sizes, $N = 3(2^g - 1)$ is shown in Fig. 2.3. Note that we do not include data points corresponding to generations $g = 1, 2$ ($N = 3, 9$ spins) due to the presence of strong finite-size effects for such small system sizes. In the dual picture, the first case corresponds to a trivial system with a single degree of freedom. In the second case, the number of dual spins is 4, however, three dual degrees of freedom are located at the boundary of the tree.

The finite-size scaling of the gap, shown in the inset of Fig. 2.3 reveals that the gap approaches a constant value with corrections that decay logarithmically in the number of spins N. This is consistent with expectations that a finite magnetic field applied to all boundary spins (these in the case of the Cayley tree constitute the finite fraction of all spins) destroys the phase transition. The presence of a gap in the thermodynamic limit allows us to conclude that the quantum adiabatic algorithm can efficiently find the ground state of the 3-XORSAT model on the considered hypergraph.

Due to the degeneracy present in this model, one can arrive at the ground state starting from a different initial state which has values of $O_{3,4,5} = -1$ so that the symmetry-breaking field vanishes. In the initial spin basis, this corresponds to choosing an initial state where the pairs of out-most spins on the boundary triangles have different spin values, i.e., $\sigma_i^x = -1$ for $i = 4, 6, 8$ while $\sigma_i^x = 1$ for all remaining value of i. In this case, however, we encounter a second-order phase transition as a function of parameter s, see Fig. 2.2(e). This result is in agreement with previous findings [NFG+08] of a second-order phase transition at $s_c \approx 0.5733$ which is characterized by a critical correlation length, $\xi = 1/\ln 2$. This peculiar behavior is due to the tree geometry of the lattice, and it is not observed for systems on local lattices, where the correlation length is known to diverge at the critical point.

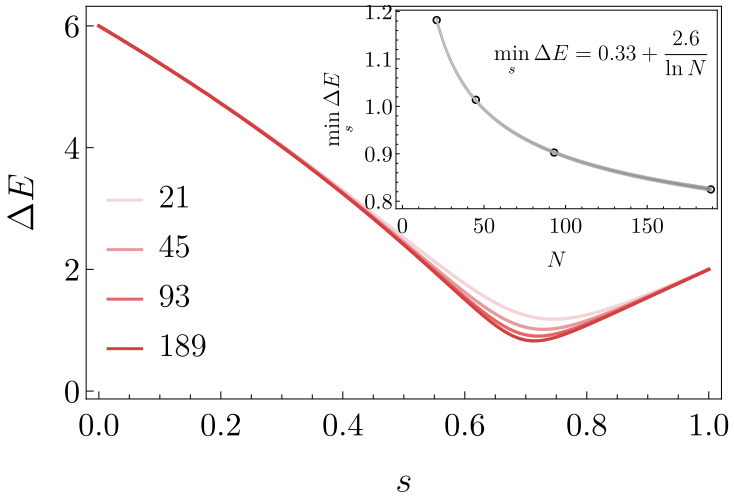

Figure 2.3: The behavior of energy gap as a function of s for open hypergraphs with different numbers of generations in the sector $O_l = 1$ demonstrates that the gap has minimal value around $s \approx 0.7$. The finite size scaling in the inset shows that the gap approaches constant value in the thermodynamic limit with corrections decaying as $1/\ln N$. Data is obtained with DMRG algorithm implemented in iTensor [FWS20] with truncation error 10^{-16}, maximum bond dimension $\chi = 45$, and number of sweeps $n_{\text{sweeps}} = 30$.

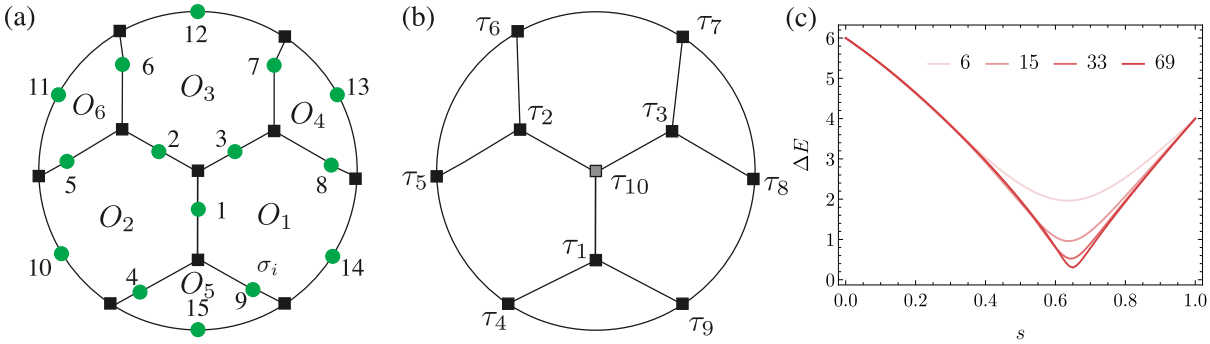

Figure 2.4: (a) Closure of the tree hypergraph at level $g = 2$ removes the boundary and leads to a 3-XORSAT instance where no spins can be decimated by leaf removal algorithm. The conserved charges labeled by $O_{1,\dots 6}$ correspond to internal loops of the lattice. (b) Dual degrees of freedom live on the closure of the tree hypergraph. The central τ-spin shown by the gray square is redundant. (c) The dependence of the minimal gap on the system size is extracted from DRMG algorithm.

2.3.2 Duality for closure of tree hypergraph

We continue the illustration of the duality by applying it to a hypergraph without a boundary shown in Fig. 2.4(a). This hypergraph can be thought of as the closure of the tree hypergraph considered above. It corresponds to an instance of the 3-XORSAT problem that does not admit a solution by the leaf removal algorithm. Indeed, all spins enter into at least two interaction edges, thus the leaf removal algorithm cannot remove any leaves at all. This second example may be considered as an example of the "glassy core" [MRZ02], and the presence of non-trivial loops leads to the appearance of non-local terms in the dual Hamiltonian. Using the duality we will argue that the minimal gap vanishes polynomially in the inverse problem size.

Degeneracies and conserved charges

The closure of the tree hypergraph with g generations has $q = 3 \cdot 2^{g-1}$ independent conserved quantities. The choice of O_l in Fig. 2.4(a) for the graph with $g = 2$ results in six conserved charges that are in one-to-one relation with the spins on the boundary. For example, $O_1 = \sigma_1^x \sigma_3^x \sigma_8^x \sigma_9^x \sigma_{14}^x$ includes only one boundary spin σ_{14}. Given that the total number of spins is $N = 3(3 \cdot 2^{g-1} - 1)$ in the general case, we expect that the dual Hamiltonian has $N_\tau = 3(2^g - 1)$ spins. For the particular instance of the graph in Fig. 2.4(a) this gives $N = 15$ and $N_\tau = 9$.

In comparison with Section 2.3.1 here the structure of the ground state manifold is more complicated. In particular, before we ignored the presence of couplings J_α since their value could be always made positive. In the present case, this is not possible anymore. Instead, we find that for any set of the coupling constants $J_\alpha = \pm 1$ it is possible to relabel operators $\sigma^z \to -\sigma^z$, so that either (i) all couplings $J_\alpha = 1$, or (ii) only one coupling is negative, $J_M = -1$, and all remaining couplings are positive. The relabeling procedure does not influence an overall parity, so option (i) is realized if $\prod_{\alpha=1}^{M} J_\alpha = 1$, while (ii) holds when $\prod_{\alpha=1}^{M} J_\alpha = -1$. Below we focus on case (i), where the system has a ground state with energy $E_0 = -M$, where M is the number of interaction edges, or the classical system of equations has an assignment that satisfies all conditions. On the other hand, in case (ii) the system is UNSAT and the ground state energy is $E_0 = M - 2$. Furthermore, for the UNSAT case, the ground state has an additional M-fold degeneracy compared to the case (i). We reserve consideration of the UNSAT case for future studies.

Dual Hamiltonian

To perform the duality transformation, we associate the τ-spins with interaction edges, see Fig. 2.4(b). However, the number of interaction edges is larger than the number of dual spins: this is related to the fact that each σ spin enters into 2 interaction edges. Thus, the product overall interaction edges, $\prod_{\alpha=1}^{M} \sigma^z_{i_\alpha}\sigma^z_{j_\alpha}\sigma^z_{k_\alpha} = 1$, results in an identity operator. We use the same relation Eq. (2.7) to define τ^x_α operator via the product of σ^z_i spins in the corresponding interaction edge. The presence of a constraint for the product of all interaction edges allows expressing one of the τ spins via the remaining operators,

$$\prod_{\alpha=1}^{M} \tau^x_\alpha = 1, \qquad \tau^x_M = \prod_{\alpha=1}^{M-1} \tau^x_\alpha. \tag{2.12}$$

While there is freedom in choosing the 'redundant' τ-spin, we fix it to be the central spin, see the shaded square in Fig. 2.4(b). In what follows we do not explicitly express τ_M spin via remaining spins to keep the notation compact.

To define τ^z_i operators we use the central site of the dual lattice as a "root". In particular, we define

$$\tau^z_i = \sigma^x_i, \quad \text{for} \quad i = 1, 2, 3. \tag{2.13}$$

Then, the remaining τ^z can be written as the product of $\sigma^x_{s\in\mathcal{P}_i}$, where $\mathcal{P}$ corresponds to a path in the lattice starting from the site $i = 1, 2, 3$. To write the quantum part of Hamiltonian in the dual basis, we express σ^x_i operators via spins τ^z_α. It is straightforward to see that $\sigma^x_i = \tau^z_\alpha\tau^z_\beta$ where edges α and β both share the spin $i = 1, \ldots, 9$ (basically all spins except the outer layer). For the spins at the outer layer of the graph we again rely on the presence of conserved charges, finding that $\sigma^x_i = O_{l_i}\tau^z_\alpha\tau^z_\beta$ where O_{l_i} is the charge that contains spin i. For instance, $\sigma^x_{15} = O_5\tau^z_6\tau^z_7$ in notations of Fig. 2.4.

With the above relations, we can finally write the expression for the dual Hamiltonian

$$\tilde{H}_T(s) = -s\sum_{\alpha=1}^{M} J_\alpha\tau^x_\alpha - (1-s)\sum_{\langle\alpha\beta\rangle}\eta_{\alpha\beta}\tau^z_\alpha\tau^z_\beta - (1-s)\sum_{\alpha=1}^{3}\tau^z_\alpha, \tag{2.14}$$

where effective couplings between dual spins α, β depend on the location of the spin as well as on the value of conserved charges:

$$\eta_{\alpha\beta} = \begin{cases} 1 & \text{if } \{\alpha,\beta\} \notin \partial V, \\ O_{l_i} & \text{if } \{\alpha,\beta\} \in \partial V,\ \tau^z_\alpha \cap \tau^z_\beta = \sigma^x_i. \end{cases}$$

Energy spectrum and minimal gap of dual Hamiltonian

As in the previous case, the value of all conserved charges is fixed by the initial state on the physical basis. The ground state of quantum Hamiltonian leads to all O_l having eigenvalue 1. The dual Hamiltonian in this sector corresponds to Eq. (2.14) with all $\eta_{\alpha\beta} = 1$ supplemented by the expression for τ^x_M via remaining spins, Eq. (2.12). It is interesting to compare Eq. (2.14) with Eq. (2.10). One difference is the appearance of a non-local term in Eq. (2.14) that is implicitly encoded in τ^x_M operator. More importantly, in the case of the tree hypergraph, one could obtain a strong symmetry-breaking magnetic field on the boundary by an appropriate choice of conserved charges. This boundary field allowed to eliminate the second-order phase transition, resulting in a finite value of gap even in the thermodynamic limit. In the case of

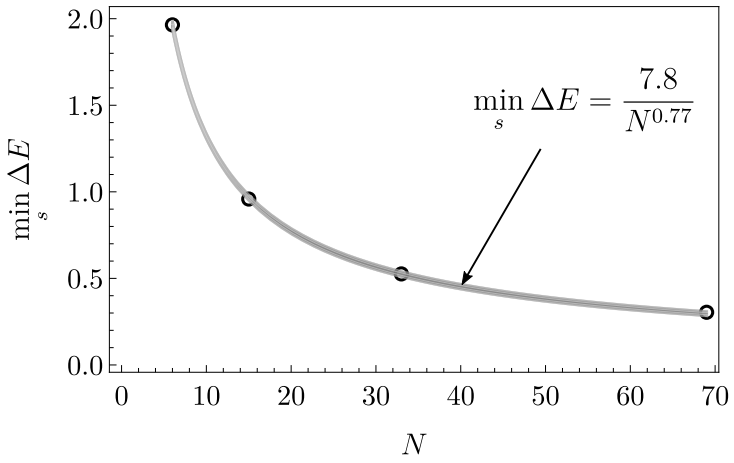

Figure 2.5: The finite size scaling shows that the gap vanishes as a power-law in system size with a coefficient $c = 0.77$. Data is obtained with DMRG implemented in iTensor [FWS20] with truncation error 10^{-16}, maximum bond dimension $\chi = 279$, and number of sweeps $n_{\text{sweeps}} = 40$.

closure of tree hypergraph, the symmetry breaking field is only present for a vanishing fraction of spins (more precisely, three spins in the center for the present gauge choice), resulting in a very different physics as we discuss below.

In Appendix A.2 we demonstrate that Eq. (2.14) with all $\eta_{\alpha\beta} = 1$ is equivalent to the transverse field Ising model on the closed lattice [see Fig. 2.4(b)] in an enlarged Hilbert space that also includes spin τ_M as a physical degree of freedom,

$$\tilde{H}_T(s) = -s \sum_{\alpha=1}^{M} J_\alpha \tau_\alpha^x - (1-s) \sum_{\langle\alpha\beta\rangle} \tau_\alpha^z \tau_\beta^z. \tag{2.15}$$

The behavior of the transverse field Ising model on the closure of the tree hypergraph was not studied before to the best of our knowledge. Due to the presence of loops, the analytical methods applied in the case of the tree hypergraph cannot be used in the present case. Therefore, we resort to numerical simulations, using the same DMRG method as in Sec. 2.3.1.

We compute numerically the ground state energy and the gap to the next excited state as a function of s, see Fig. 2.4(c). Note, that naïvely such gap vanishes in the Hamiltonian (2.15) for values of s close to zero since the model is in symmetry-broken phase. However, as we discuss in Appendix A.2, the success of the quantum adiabatic algorithm depends on the gap restricted to the even Z_2-symmetry sector. The finite-size scaling of the gap performed for systems with up to $N = 69$ spins (corresponding to $M = 45$ dual spins) in Fig. 2.5 shows the gap vanishes as a power-law with system size. This gives strong evidence of a second-order phase transition encountered at $s \approx 0.65$, which can be expected due to the presence of Z_2 symmetry in dual Hamiltonian. We note that the fit of the numerical data to a slow exponential decay is noticeably worse, as reflected by the Bayesian information criterion (BIC), which for the power-law fit equals BIC $= -25.4$ while for the exponential fit equals BIC $= 1.99$.

2.4 Discussion

Motivated by the fact that many interesting classical problems have degeneracy in solution space, in this paper we studied the performance of quantum adiabatic algorithms applied to such problems. To this end, we introduced duality as a generic tool that allows us to efficiently target such problems and formulated it using the language of linear algebra in Appendix A.1. In the main text, we demonstrated the application of duality to two different instances of the 3-XORSAT problem.

First, we applied the general duality to the 3-XORSAT problem on a tree hypergraph, which may be thought of as imitating the structure of the leaves of a generic 3-XORSAT instance. Such an instance of the 3-XORSAT problem can be efficiently solved by a classical algorithm in a polynomial time. In Sec. 2.3.1 we found that the dual Hamiltonian corresponds to the Ising model with longitudinal magnetic fields at the boundary of the graph. Thus, when starting the annealing process from the paramagnet state the gap saturates to a constant value in the thermodynamic limit with corrections decaying as $1/\ln N$. This implies that the application of the quantum adiabatic algorithm could yield a solution in a finite amount of time, even in the thermodynamic limit.

As a more general example, we considered a 3-XORSAT problem on the closure of the tree hypergraph, which may be considered as a cartoon picture of a "glassy core". Despite being non-amenable to the leaf removal algorithm, this instance of the 3-XORSAT problem is still solvable in a polynomial time by a classical algorithm. The presence of non-trivial loops in this geometry translates into the appearance of non-local terms in the dual quantum Hamiltonian. We found that the minimal gap of the annealing Hamiltonian vanishes as a power-law with the problem size, implying the quantum adiabatic algorithm would now require a time that is polynomial in the problem size.

Despite considering only two toy examples of the 3-XORSAT with extensive degeneracy of classical solution space, the application of duality revealed an interesting connection between the behavior of the minimum gap and the structure of the lattice. In particular, we observed that by closing the boundary of the tree hypergraph the minimum gap changes from being constant in the thermodynamic limit to decaying as a power law in system size. This suggests that in the most complex case, a first-order phase transition may emerge, similarly to other instances of 3-XORSAT with unique ground state considered previously [JKSZ10, FGH$^+$12]. In addition, duality may be used to obtain useful analytical results for the entanglement spectrum. In particular, we expect the entanglement spectrum of a given subregion to contain information about conserved charges that are supported within the subregion.

More generally, the two considered examples of the tree hypergraph and its closure can be viewed as a basis of perturbation theory, as more typical hypergraphs can be obtained by "decorating" the tree hypergraph with additional interactions. In particular, changes to the hypergraph geometry that add additional interaction terms typically break the formerly conserved charges. This would correspond to the introduction of additional non-local degrees of freedom into the dual Hamiltonian. Such an approach can be potentially used to target more complex instances of the 3-XORSAT and possibly relate the problem with classical clustering in the ground state manifold to instances of quantum clustering, that was recently considered in the literature [MHS$^+$17]. Additionally, these considerations suggest that frustration that is brought by additional interaction terms naturally corresponds to non-local interactions in the dual language.

Finally, throughout this work, we focused on the ground state and low-lying excitations of

the Hamiltonian used in the quantum adiabatic algorithm to solve the classical 3-XORSAT problem. The study of highly excited states of such Hamiltonians remains an interesting problem, where duality obtained in our work can bring useful insights. In particular, it would be interesting to investigate if these models could allow for a non-ergodic phase similar to the one found in [BLPS17].

CHAPTER 3

Avoiding Barren Plateaus Using Classical Shadows

In this Chapter, we investigate the issue of barren plateaus – flat regions in the cost function landscape – in variational quantum algorithms. We introduce a weaker version of barren plateaus, called *weak barren plateaus* in terms of the entropies of local reduced density matrices. The presence of WBPs can be efficiently quantified using recently introduced shadow tomography of the quantum state with a classical computer. We demonstrate that avoidance of WBPs suffices to ensure sizable gradients in the initialization. In addition, we demonstrate that decreasing the gradient step size, guided by the entropies allows us to avoid WBPs during the optimization process. This Section is based on the paper:

Stefan H. Sack, Raimel A. Medina, Alexios A. Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3:020365, Jun 2022

3.1 Introduction

In recent years the field of quantum computation has seen rapid growth fueled by the arrival of the first generation of quantum computers, dubbed noisy intermediate-scale quantum devices (NISQ) [Pre18]. The NISQ era is characterized by quantum computers with a small number of qubits and limited control. The number of coherent operations that can be performed is small and the implementation of famous algorithms with proven quantum speedups, such as Shor's algorithm [Sho95], remains out of reach. To make use of the current generation of quantum computers, the so-called variational hybrid approach [BCK$^+$21] was proposed. The idea is to use the quantum computer in a feedback loop with a classical computer, where it implements a variational wave function that is measured to compute the value of the so-called cost function. This information is then fed into a classical computer where it is processed and the variational wave function is subsequently updated aiming to find a minimum of the cost function, which provides an (approximate) solution to the computationally hard problem. The variational hybrid approach has seen a wide range of proof of concept applications on NISQ devices ranging from quantum chemistry [KMT$^+$17, Aru20] to quantum optimization [Har21a, LHA$^+$20] and quantum machine learning [HCT$^+$19, JDM$^+$21].

Despite the large number of suggested applications, the variational approach encountered also several obstacles, that have to be overcome for the future success of the method. In

particular, the infamous emergence of *barren plateaus* (BPs) implies that expressive variational ansätze tend to be exponentially hard to optimize [MBS+18]. The main obstacle on the way to optimization lies in the fact that gradients of the cost function are on average zero and deviations vanish exponentially in system size, thus precluding any potential quantum advantage. Moreover, it has been shown that the classical optimization problem is generally NP-hard and plagued with many local minima [BK21].

The problem of BPs attracted significant attention, and numerous approaches were proposed in the literature. In particular, the early research focused on avoidance of BP at the *initialization stage* of variational algorithms [GWOB19, SMM+20, DBW+21, HSCC21, LCS+21]. In a different direction, the relation between the occurrence of BPs and the structure of the cost function was studied [CSV+21, UB20]. Also, notions of so-called entanglement-induced [OKW20] and noise-induced [WFC+20] BPs were introduced. The relation between BPs and entanglement has led to various proposals that suggest controlling entanglement to mitigate BPs [KO21a, KO21b, PNGY21, WZCK21]. However, measuring entanglement is hard, therefore making these approaches impractical on a real quantum device.

In this work, we introduce the notion of *weak barren plateaus* (WBPs), to diagnose and avoid BPs in variational quantum optimization. WBPs emerge when the entanglement of a local subsystem exceeds a certain threshold identified by the entanglement of a fully scrambled state. In contrast to BPs, WBPs can be efficiently diagnosed using the few-body density matrices and we show that their absence is a sufficient condition for avoiding BPs. Based on the notion of WBPs, we propose an algorithm that can be readily implemented on available NISQ devices. The algorithm employs *classical shadow* estimation [HKP20] during the optimization process to efficiently estimate the expectation value of the cost function, its gradients, and the second Rényi entropy of small subsystems. The tracking of the second Rényi entropy enabled by the classical shadows protocol allows for an efficient diagnosis of the WBP both at the initialization step and during the optimization process of variational parameters. If the algorithm encounters a WBP, as witnessed by a certain subregion having a sufficiently large Rényi entropy, the algorithm restarts the optimization process with a decreased value of the update step (controlled by the so-called learning rate). We support the proposed procedure with rigorous results and numerical simulations. The structure of the paper is as follows:

In Sec. 3.2 we introduce the theoretical framework and present our main results. In Sec. 3.2.1 we introduce the framework of variational quantum eigensolvers (VQEs). Sec. 3.2.2 introduces the phenomenon of BPs which dramatically hinders the performance of VQEs. In Sec. 3.2.3 we demonstrate WBPs to be a precursor to BPs. We explain why and how WBPs can be efficiently diagnosed in experiments and contrast this with the much harder task of detecting BPs. Finally, we propose a modification to the VQE algorithms which allows to prevent the occurrence of BPs by avoiding WBPs.

In Sec. 3.3 we present a bound for the expectation value of the second Rényi entropy in quantum circuits drawn from a unitary ensemble forming a 2-design. This bound allows us to use the second Rényi entropy, which is much easier to estimate, instead of the entanglement entropy. In Sec. 3.3.1 we provide a formal definition of WBPs according to the value of the second Rényi entropy of the subsystem and prove that the occurrence of a BP implies the occurrence of a WBP. From this argument, it follows that the absence of a WBP precludes the occurrence of a BP. In addition, we provide an upper bound (whose proof is found in Appendix B.1) for the measurement budget required to estimate a WBP using classical shadows. Finally, in Sec. 3.3.2 we demonstrate numerically how the avoidance of WBPs precludes the presence of a BP using the popular BP-free small-angle initialization [HSCC21, HBK21].

In Sec. 3.4, we explore how BPs/WBPs emerge at different stages in the VQE optimization and perform a systematic performance analysis. Next, in Sec. 3.4.1 we explore the relation of the learning rate and entropy growth for a single update of the VQE algorithm. We analytically and numerically illustrate how a large learning rate leads to an uncontrolled growth in subsystem entropies, essentially driving optimization to a WBP region. In Sec. 3.4.2 we explore the performance of the WBP-free VQE algorithm in different settings for the Heisenberg model on a chain. Finally, in Sec. 3.4.3, we show that our approach allows for the efficient convergence to both, area- and volume-law entangled ground states and compare it to layerwise optimization [SMM+20] which is a popular heuristic for BP avoidance. This is illustrated using the Heisenberg model on a random 3-regular graph, additional results for the Sachdev-Ye-Kitaev (SYK) model can be found in the Appendix B.5 which exhibits a nearly maximally entangled ground state.

Finally, in Sec. 3.5 we summarize our results, discuss their implications, and outline open questions.

3.2 Avoiding barren plateaus in variational quantum optimization

In this section, we first introduce the framework of VQEs, i.e. the unitary ensemble, the cost functions, and the optimization algorithm, and discuss the BP problem. After this, we present our main result – a specific modification of the variational quantum eigensolver (VQE) that avoids the issue of BPs.

3.2.1 Variational quantum eigensolver

The aim of the VQE, initially introduced by [PMS+14], is to approximate the ground state $|GS\rangle$ of a Hamiltonian H with a variational wave function $|\psi(\boldsymbol{\theta})\rangle$. A quantum computer is used to prepare the variational function via the action of a set of unitary gates, $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\,|\psi_0\rangle$, where $|\psi_0\rangle$ is the initial state that is typically assumed to be a product state. The variational parameters are then iteratively updated to minimize the expectation value of the Hamiltonian, also called cost function $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,H\,|\psi(\boldsymbol{\theta})\rangle$.

We consider a unitary circuit $U(\boldsymbol{\theta})$ of the form of the so-called "hardware-efficient" ansatz [KMT+17]

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{p} W_l \left(\prod_{i=1}^{N} R_l^i(\theta_l^i) \right), \tag{3.1}$$

where $\theta_l^i \in [-\pi, \pi]$ are pN variational angles, concisely denoted as $\boldsymbol{\theta}$. In this expression the product goes over spatial dimension $i = 1, \ldots, N$ that labels individual qubits and "time dimension", $l = 1, \ldots, p$ with p specifying the number of layers, see Fig. 4.1 (a). We choose the single qubit gates to be rotations $R_l^i(\theta_l^i) = \exp\left(-\frac{}{2}\theta_l^i G_{l,i}\right)$ with random directions given by $G_{l,i} \in \{\sigma^x, \sigma^y, \sigma^y\}$. W_l is an entangling layer that consists of two-qubit entangling gates represented by nearest-neighbor controlled-Z (CZ) gates with periodic boundary conditions, see Fig. 4.1 (a) for an illustration.

We focus our study on k-local Hamiltonians H, defined as the sum of terms each containing at most k Pauli matrices. We take k to be finite and fixed, while the number of qubits $N \gg k$. A particular example of 2-local Hamiltonian from the many-body physics is provided by the

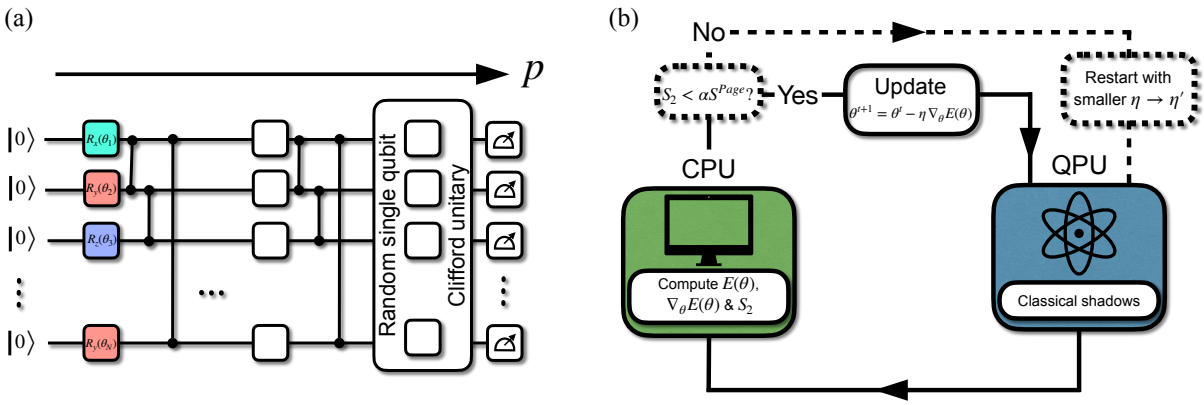

Figure 3.1: (a) Illustration of the variational quantum circuit $U(\boldsymbol{\theta})\,|0\rangle$ that is considered in the main text followed by the shadow tomography scheme [HKP20]. The variational circuit consists of alternating layers of single qubit rotations represented as boxes and entangling CZ gates shown by lines. The measurements at the end are used to estimate values of the cost function, its gradients, and other quantities. (b) The original hybrid variational quantum algorithm shown by solid boxes can be modified without incurring significant overhead as is shown by the dashed lines and boxes. The modified algorithm tracks the entanglement of small subregions and restarts the algorithm if it exceeds the fraction of the Page value that is set by parameter α. The full algorithm is efficient, rigorous sample complexity bounds are provided in Appendix B.1.

Heisenberg (XXX) model subject to a magnetic field

$$H_{XXX} = \sum_{i,j\in V_{\mathcal{G}}} J\left(\sigma_i^z\sigma_j^z + \sigma_i^y\sigma_j^y + \sigma_i^x\sigma_j^x\right) + h_z\sum_{i=1}^{N}\sigma_i^z, \tag{3.2}$$

where $V_{\mathcal{G}}$ refers to the vertex set of the graph $\mathcal{G}$ and, couplings are fixed $J = h_z = 1$. In our simulations, we consider two different graphs: a ring corresponding to a 1D chain with periodic boundary conditions, and a random 3-regular graph. The $U(1)$ symmetry related to the conservation of the z-component of the spin under the action of H, as well as translational invariance present for chains with periodic boundary conditions, can be explored to decrease the space of parameters by using a suitable gate set respecting this symmetry. However, for the sake of generality, we focus on the hardware-efficient unitary ansatz defined in Eq. (3.1).

Obtaining the energy expectation value $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,H\,|\psi(\boldsymbol{\theta})\rangle$ requires measuring a subset or all qubits in the circuit as we schematically show in Fig. 4.1 (a). For our example of 2-local Hamiltonian on the 1D chain, the required measurements include the value of σ^z operator on all sites along with the $\sigma_i^a\sigma_{i+1}^a$ expectation values of all $i = 1,\ldots N$ (periodic boundary condition is assumed, so that bits 1 and $N+1$ are identified) and $a = x, y, z$. Finding the optimal parameters $\boldsymbol{\theta}^\star$ requires minimization of the Hamiltonian expectation value $E(\boldsymbol{\theta}^\star) = \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ performed by a classical computer.

There is a plethora of sophisticated classical optimization algorithms that were applied to this minimization problem [OGB21, SIKC20, KB14, GZCW21]. We use the plain gradient descent (GD) algorithm due to its simplicity which makes analytical considerations easier. A GD update step is given by

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\nabla_{\boldsymbol{\theta}}E(\boldsymbol{\theta}), \tag{3.3}$$

where η is the *learning rate* which controls the update magnitude. This update step is repeated until convergence of $E(\boldsymbol{\theta})$ which results from finding a (local) minimum of $E(\boldsymbol{\theta})$.

The resulting VQE algorithm is shown schematically in Fig. 4.1 (b) by solid lines. Following the initialization of the variational angles $\boldsymbol{\theta}$, which may be chosen to be real random numbers, the quantum computer is used to prepare the variational state and provide quantum measurement results. The classical computer uses the measurements to estimate the value of the cost function and its gradient and performs an update of the variational parameters controlled by the learning rate η.

3.2.2 Barren plateaus and entanglement

Whilst the VQE described above is a promising framework for near-term quantum computing due to its modest hardware requirements, its performance may be ruined by the issue of barren plateaus [MBS$^+$18, CSV$^+$21, HSCC21]. Specifically, the BPs are defined as regions in the space of variational parameters where the variance of the cost function gradient (and consequently its typical value) vanishes exponentially in the number of qubits [MBS$^+$18]:

$$\mathrm{Var}[\partial_{i,l} E(\boldsymbol{\theta})] \sim \mathcal{O}\left(\frac{1}{2^{2N}}\right). \tag{3.4}$$

[MBS$^+$18] were among the first to theoretically investigate BPs. They showed that the appearance of a BP can be related to the circuit matching the Haar random distribution up to the second moment. More precisely, they showed that BPs are a consequence of the unitary ensemble $\mathcal{E} \sim \{U(\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ forming a so-called 2-design [MBS$^+$18] (see Appendix B.2 for details and the definition of a t-design). To understand the different circuit depths at which BPs are encountered, the authors in Ref. [CSV$^+$21] introduced the concept of cost function-dependent BPs. In particular, they argued that the emergence of BP occurs at different circuit depths, depending on the nature of the cost function.

In contrast, for a so-called global cost function, exemplified by the fidelity, Ref. [CSV$^+$21] found that BPs already occur at very modest circuit depths $p \sim \mathcal{O}(1)$. The emergence of BP for fidelity is naturally related to “orthogonality catastrophe" in many-body physics: even a small global unitary rotation applied to the many-body wave function results in it becoming nearly orthogonal to itself. In terms of fidelity, this implies that it vanishes exponentially in the number of qubits. Moreover, most global state features – such as expectation values of general operators, fidelities with general states and global purities – cannot be efficiently accessed on NISQ devices, and are therefore not practical from an algorithmic point of view [FL11, HKP20, HBC$^+$21, CCHL21]. Therefore, in what follows we do not consider the global cost functions and corresponding BPs.

Local cost functions, which are the focus of the present work are characterized by a later onset of BPs. Specifically, for a k-local cost function where k is fixed, the BPs will occur for circuit depth $p \sim \mathcal{O}(poly(N))$ that increases polynomially in system size [MBS$^+$18, CSV$^+$21]. In other words, for a large enough p the VQE algorithm will also suffer from a BP already at the very first step of the GD optimization, provided random choice of variational angles $\boldsymbol{\theta}$. We also note that gradient-free optimization strategies do not circumvent the BP problem since the optimization landscape is inherently flat [ACC$^+$21].

The potential emergence of BPs at the initialization stage of the VQE and other algorithms spurred the investigation of different initialization strategies that avoid BPs. Until now, several BP-free initializations have been considered in the literature. Ref. [GWOB19] suggests initializing the circuit with blocks of identities, Ref. [SMM$^+$20] suggests to optimize the ansatz layer by layer, and Ref. [DBW$^+$21] suggests to use a matrix product state ansatz that is

optimized by a separate algorithm [CPSV20] and map that to a quantum circuit. In this work, we will focus on small single qubit rotation as suggested in Ref. [HSCC21].

More recently, it was observed that the entanglement entropy defined as a trace of the reduced density matrix, $S = -\operatorname{tr}\rho_A \ln \rho_A$ (where $\rho_A = \operatorname{tr}_B \rho$ is the reduced density matrix where A is the subset of qubits that are measured and B is the rest of the system) is another source for the occurrence of BPs [OKW20]. The community has subsequently dubbed this kind of BP, *entanglement induced* BP [OKW20, KO21a, WZCK21, PNGY21]. In this work, we will however show that entanglement-induced BPs and BPs for local cost functions are the same. Avoiding entanglement-induced BPs is equivalent to avoiding BPs for local cost functions, the details are presented in Sec. 3.3.

Experimentally probing a BP is a hard task. The estimation of the exponentially small gradient in Eq. (3.4) requires a number of measurements that is exponential in the number of qubits, and therefore invalidates any potential quantum speedup. Moreover, small values of gradient encountered in BP have to be distinguished from the case when the gradient vanishes due to convergence to a local minimum. Experimentally diagnosing BPs via entanglement is also impractical. For example, quantum circuits that implement 2-design and thus lead to BPs for local cost functions are characterized by typical volume-law entanglement that approaches nearly maximal values. Checking volume-law entanglement scaling on any device is a formidable challenge.

In the process of variational quantum optimization, the majority of approaches to mitigate BPs apply to the initialization stage [GWOB19, VBM$^+$19, VC21] and not during the optimization. In Sec. 3.4, we illustrate the importance of BP mitigation during the optimization. This motivates the need to devise a BP mitigation strategy for the initialization and during the optimization procedure that is efficient. This procedure will be discussed in the algorithm proposed below.

3.2.3 Weak barren plateaus and improved algorithm

To devise an efficient algorithm for BP-free initialization and optimization of the VQE we introduce the notion of WBPs. Specifically, for a Hamiltonian that is k-local, we define the WBP as the point where the second Rényi entropy $S_2 = -\ln \operatorname{tr}\rho_A^2$ of any subregion of k-qubits satisfies $S_2 \geq \alpha S^{\text{Page}}(k, N)$, where the Page entropy in the limit $k \ll N$ corresponds to the (nearly) maximal possible entanglement of subregion A,

$$S^{\text{Page}}(k, N) \simeq k \ln 2 - \frac{1}{2^{N-2k+1}}, \tag{3.5}$$

where we explicitly used the Hilbert space dimensions of regions A is 2^k and its complement B has Hilbert space dimension 2^{N-k}. The naive choice for the parameter α is $\alpha = 1$. Given some a priori knowledge of the entanglement structure of the target state $|GS\rangle$, the choice can, however, be more informed to help avoid large entanglement local minima, see Sec. 3.3.

The notion of WBP is practical since it is defined by k-body density matrices, being much easier to access on a real NISQ device. The fact that the prevention of a WBP is sufficient for avoiding the BP may be understood by the intuition from quantum many-body dynamics and the process of thermalization or scrambling of quantum information. In the thermalization process, the small subsystems are first to become strongly entangled, as is witnessed by the proximity of their density matrix to the infinite temperature density matrix. This intuition suggests that it is enough to keep in check the density matrices of small subsets of qubits. If

their entanglement or other properties are far away from thermal, the system overall is still far away from the BP.

Practically, the WBP can be diagnosed using the shadow tomography scheme [HKP20]. This scheme enables an efficient way of representing a classical snapshot of a quantum wave function on a classical computer. In essence, shadow tomography replaces the measurements performed in the computational basis with more general measurements, that turn out to be sufficient for reconstructing linear and non-linear functions of the state, such as expectation values of few-body observables and second Rényi entropy of few-body reduced density matrices respectively.

Relying on the shadow tomography, we propose the following modification of the VQE shown by dashed lines in Figure 4.1 (b). In essence, we suggest using the tomography to *simultaneously* measure the cost function value and the k-body second Rényi entropy. For the derivative we require an additional $2pN$ tomographies (two for each parameter) to compute the gradient exactly using the parameter shift rule [MNKF18, SBG$^+$19], a detailed derivation of the computational cost for each operation is presented in Appendix B.1. Access to the second Rényi entropy allows to prevent the appearance of WBPs not only at the initialization step but throughout the optimization cycle. The explicit algorithm works as follows:

Algorithm 1 `WBP free optimization with shadows`

Choose α, default is $\alpha = 1$ ▷ see Sec. 3.3.1 for details
Choose $\boldsymbol{\theta}$ such that $S_2 < \alpha S^{\text{Page}}(k, N)$
Choose learning rate η
repeat ▷ see Appendix B.1 for details
    Obtain classical shadows $\hat{\rho}^{(t)}(\boldsymbol{\theta})$
    Use them to compute $E(\boldsymbol{\theta})$, $\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ and $S_2(\boldsymbol{\theta})$
    if $S_2 < \alpha S^{\text{Page}}(k, N)$ **then**
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$
    else
        Start again with smaller $\eta \leftarrow \eta'$
    end if
until convergence of $E(\boldsymbol{\theta})$

If a WBP is diagnosed at the initialization, one may have to take a different initial value of the variational angles or change the initialization ensemble. These aspects are discussed in detail in Sec. 3.3. In addition, the WBP can occur in the optimization loop. This can be mitigated by keeping track of the second Rényi entropies in the optimization process. If the WBP condition is fulfilled, one must restart the algorithm with a smaller learning rate. In the Section 3.4 we discuss the optimization process in greater detail. In particular, we will show how the learning rate is related to the potential change in entanglement entropy which implies that a smaller learning rate is generally better at avoiding WBPs.

3.3 Weak barren plateaus and initialization of VQE

3.3.1 Definition and relation to barren plateaus

As mentioned above, BPs for local cost functions are a consequence of the unitary ensemble $\mathcal{E} \sim \{U(\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ forming a 2-design [MBS$^+$18, CSV$^+$21] which leads to an exponentially vanishing

gradient variance, i.e. a BP. What is important to note is that the exponential decay is simply a witness of the emergence of a 2-design. Another equivalent witness is the second Rényi entropy, where we have:

Theorem 1. *(2-design and Rényi entropy) If the unitary ensemble $\mathcal{E} \sim \{U(\boldsymbol{\theta})\}$ forms a 2-design, then for typical instances the second Rényi of the state ρ_A concentrates around the Page value*

$$S^{Page}(k,N) - \frac{1}{2^{N-2k+1}} \leq \mathbb{E}_{\mathcal{E}}\Big[S_2(\rho_A)\Big] \leq S^{Page}(k,N),$$

for all subregions A of size $k \ll N$.

These results are known in the literature, and in the context of random quantum circuits, can be found in Refs. [PSW06, ODP07, DOP07]. However, for completeness, we also provide proof in Appendix B.3.

The Theorem above implies that a large amount of entanglement naturally follows from the similarity between the considered circuit and a random unitary (2-design). Such similarity also gives rise to the vanishing variance of local cost function gradients that define BPs. Therefore, so-called entanglement-induced BPs [OKW20] and BPs for local cost functions are the same. Entanglement provides an intuitive picture of the emergence of BPs and their circuit depth dependence. Every entangling layer in the circuit typically increases entanglement of the resulting wave function, until it saturates to its maximal value for any subregion of k-qubits at a circuit depth $p \sim \mathcal{O}(\text{poly}(N))$. If the second Rényi entropy for half of the subsystem $k = N/2$ has saturated, it has saturated for all smaller subsystem sizes and is thus a sufficient check for a BP. Computing the second Rényi is however typically exponentially hard in subsystem size on NISQ devices (for single-copy access this was recently proven in Ref. [CCHL21, HBC$^+$21]). It is therefore only practical to check a small subregion where k is small and independent of system size.

The above considerations naturally lead us to introduce the notion of WBPs as a modification of the BP that is computationally efficient to diagnose on NISQ devices. More formally we have that:

Definition 2. *(Weak barren plateaus) Let H be an N-qubit Hamiltonian, and A is a region containing k qubits. We define a weak barren plateau by the second Rényi entropy of the reduced density matrix ρ_A satisfying $S_2 \geq \alpha S^{Page}(k,N)$ with $\alpha \in [0,1)$.*

This definition works for any k, however, it is reasonable to use k that corresponds to the number of spins involved in interaction terms in the Hamiltonian H since it provides a natural length scale. Moreover, in such a case, the reduced density matrix of the subregion with k spins contains all the necessary information needed to extract the expectation values of Hamiltonian terms localized inside this region.

While a WBP is a necessary condition for a BP, it is however not sufficient (which motivates the term *weak*). From a practical perspective, we are only interested in avoiding a BP. For this, WBPs provide a powerful tool, since:

Corollary 2.1. *If we find a particular subregion A such that ρ_A does not satisfy the weak barren plateau condition, i.e. Definition 2, it is on average also not in a barren plateau where the variance is exponentially small.*

Proof. This assertion immediately follows from negating Theorem 1. □

The Corollary above formalizes the intuition behind the dynamics of entanglement in a circuit: If the state restricted to the smaller subsystem has not scrambled, then neither has the state restricted to a larger subregion. In practice, using classical shadows we can efficiently check one subregion of size k with a total measurement budget

$$T \geq \frac{4^{k+1} \operatorname{tr} \rho_A^2}{\epsilon^2 \delta}, \tag{3.6}$$

where ϵ is the desired accuracy and δ is a failure probability (over the randomized measurement process). Parameters ϵ and δ do not depend on the number of qubits, whereas the factor $\operatorname{tr} \rho_A^2$ is upper bounded by one for weakly entangled states and can be as small as 2^{-k} when entanglement is large. Moreover, checking all size k subregions incurs an additional overhead of only $k \ln N$. A derivation of this result is presented in Appendix B.1, see Eq. (B.7). Provided that k is small and does not scale with system size, N, this can be efficiently implemented on NISQ devices.

If any of these subregions avoids the WBP condition, we are guaranteed to also avoid an actual BP. For simplicity, in the numerical results below we check for the WBP condition for a particular region containing the first k qubits, i.e. $A = \{1, \cdots, k\}$.

This argument is also intuitive to see by considering a causal cone (blue region) that indicates the extent of the so-called scrambled region (i.e. extent of a subregion with entropy close to the maximal value) in the circuit, see Fig. 4.2 (a). Such scrambled region grows with every consecutive entangling layer W_l (see Eq. (3.1)). When this region extends beyond k qubits, the WBP is reached (left orange dashed line). Later, when the "scrambling lightcone" has extended to the full system, the BP is reached (right orange dashed line). Once the BP is reached all smaller regions are also fully entangled and will satisfy the WBP condition on average.

Fig. 4.2 provides a numerical illustration for the Corollary 2.1 stated above. We use the hardware-efficient circuit, presented in Eq. (3.1), and compute the gradient variance and second Rényi entropy as a function of circuit depth p for different system sizes N. We fix $|\psi_0\rangle = |0\rangle$ as the initial state, which is simply all qubits in the zero state. Panel (b) shows the exponential decay of the gradient variance that is usually used to diagnose a BP. Panel (c) shows the corresponding bipartite second Rényi entropy. We see that it indeed approaches the Page value (gray dashed line). The Page value is not fully reached since we are considering the second Rényi instead of the von Neumann entanglement entropy, this difference however becomes negligible once the subsystem size is decreased. This numerically illustrates that when the 2-design is reached both the gradient variance and bipartite second Rényi entropy have converged. In panel (d) we consider a smaller region of two qubits and see that the second Rényi for this region saturates to its maximal value at a significantly lower circuit depth. This illustrates the emergence of the WBP that precedes the onset of the BP after a few more entangling layers. Before the WBP is reached, gradients are well-behaved and do not decrease exponentially with the system size.

Finally, we address the effects of the control parameter α, that enters in Definition 2 of the WBP. The naive choice is $\alpha = 1$ which means that a WBP is reached if the subregion is maximally entangled with the rest of the system. However, in the case when some a priori knowledge about the entanglement properties of the target state $|GS\rangle$ is available, it can be used to set a smaller value of α. If, for instance, the ground state is only weakly entangled,

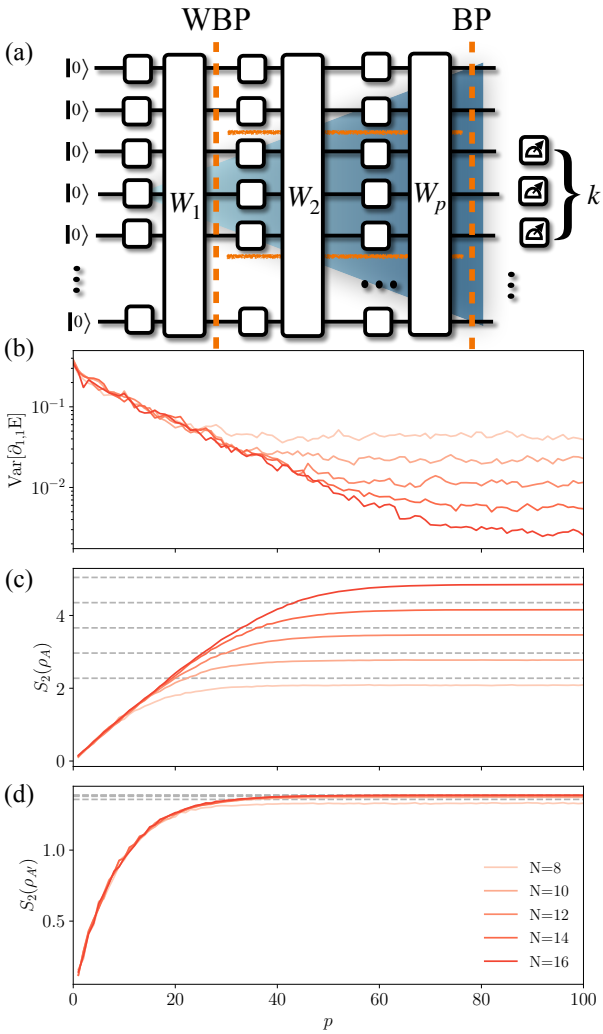

Figure 3.2: (a) Sketch of the circuit, where the blue color shows the scrambling lightcone. The lightcone first extends over k qubits, where the WBP occurs, and for larger circuit depths extends to the full system size where the BP occurs. (b) The saturation of the gradient variance $\mathrm{Var}[\partial_{1,1}E]$ and (c) saturation of the bipartite second Rényi entropy $S_2(\rho_A)$ of the region A consisting of qubits $1,\ldots,N/2$ nearly to the Page value happen at the similar circuit depths p, that increases with the system size N. (d) In contrast, the saturation of the second Rényi for two qubits ($A' = \{1,2\}$) is system size-independent, illustrating that WBP precedes the onset of a BP. Data was averaged over 100 random initializations. Gradient variance is computed for the local term $\sigma_1^z\sigma_2^z$, typically used in BP illustrations. Gradient variance for the full Heisenberg Hamiltonian, Eq. (3.2), looks similar.

a choice of $\alpha \ll 1$ may be appropriate. In this way Algorithm 1 in Sec. 3.2.3 can also help in avoiding convergence to highly entangled local minima. We discuss this in more detail in Sec. 3.4.2.

3.3.2 Illustration of WBP-free initialization

In order to illustrate the notion of WBP in a more specific setting we apply it to the initialization process of the VQE. Specifically, we focus on the family of initializations that was proposed earlier in order to avoid the issue of BPs [HSCC21, HBK21]. The one-parametric family of initializations restricts the single qubit rotation angles from ansatz Eq. (3.1) as $\theta_l^i \in \epsilon_\theta[-\pi,\pi)$,

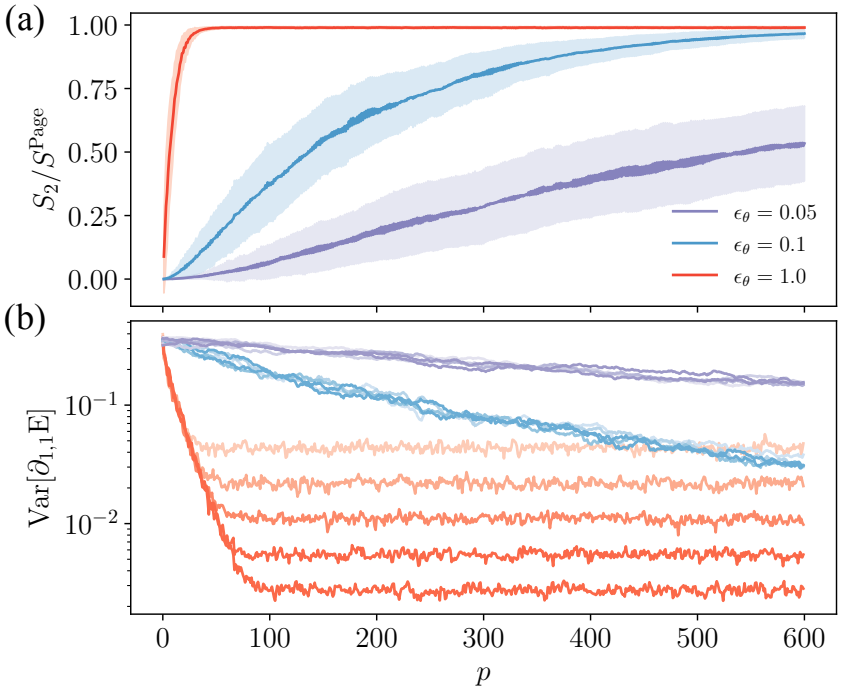

Figure 3.3: (a) Decreasing parameter ϵ_θ from 1 slows down the growth of the second Rényi entropy with the circuit depth p. The chosen region contains two qubits. (b) The encounter of BP in the variance of the gradient of the cost function is visible only for the case $\epsilon_\theta = 1$, and it is preceded by the onset of a WBP. We use a system size of $N = 16$ for (a) and $N = 8, \cdots, 16$ for (b), color intensity corresponds to system size, same as in Fig. 4.2. Data is averaged over 100 random instances, variance is for the local term $\sigma_1^z \sigma_2^z$.

where $\epsilon_\theta \in [0, 1)$ is the control parameter. This strategy allows to delay of the onset of BP to arbitrary circuit depths by tuning ϵ_θ accordingly.

Similarly, it allows for a delay in the onset of WBPs. Depending on the parameter ϵ_θ one can afford a deeper circuit without encountering a WPB in the initialization when compared to the full parameter range ($\epsilon_\theta = 1$). It is straightforward to see that for $\epsilon_\theta = 0$, the ansatz is WBP-free for all circuit depths. Indeed, in the absence of the single qubit rotations, the entangling gates in W_l do not create any entanglement (since the CZ gates used in Eq. (3.1) are diagonal in the computational basis), leaving $|0\rangle$ invariant. Note that, for example, the *identity block* initialization, proposed by [GWOB19] works similarly in that the unitary is constructed such that it also implements the identity and one is equally left with the zero states.

In Fig. 4.3 we numerically illustrate the influence of ϵ_θ on the growth of entanglement and its relation to the gradient variance. Panel (a) illustrates the growth of the second Rényi entropy in the circuit for three different small angle parameters ϵ_θ and panel (b) shows the corresponding gradient variance. Outside of the WBP the gradient variance vanishes at most polynomially in system size N. This illustrates that the avoidance of a WBP is sufficient for avoiding a BP and thus allows for a simple strategy for constructing BP-free initializations.

3.4 Entanglement control during optimization

3.4.1 Bounding entanglement increase at a single optimization step

In Sec. 3.2 we presented how the general VQE can be extended with minimal overhead to avoid WBPs in the optimization procedure. The learning rate, as presented in Algorithm 1, hereby plays a crucial role. A smaller learning rate, as observed in Fig. 4.1 (c)-(e) is more likely to avoid a WBP. To understand this phenomenological observation on more rigorous grounds, let us consider a sufficiently deep circuit (with a polynomial number of layers in system size), so that the optimization landscape is dominated by WBPs. Careful selection of the parameters allows for an initialization outside of a WBP. However, to remain in the WBP-free region, the optimization has to be performed in a controlled manner, such that the optimizer does not leave the region of low entanglement due to large learning rate and does not end in a WBP.

Since WBPs are defined in terms of the second Rényi entropy S_2, we need to bound the change in S_2 between iteration steps t and $t+1$. For practical purposes, we instead use the purity ($\operatorname{tr}\rho_A^2 = e^{-S_2}$). The change in purity is upper bounded by [CMNF16]

$$\left|\operatorname{tr}\rho_A^2(t+1) - \operatorname{tr}\rho_A^2(t)\right| \leq 1 - (1 - T_A(t))^2 - \frac{T_A^2(t)}{2^k - 1}, \tag{3.7}$$

where $T_A(t) \equiv T(\rho_A(t), \rho_A(t+1))$ is the trace distance between the reduced density matrices at iteration steps t and $t+1$, and we assume that region A has k qubits.

Assuming that the states at consecutive update steps of gradient descent are perturbatively close (see Appendix B.4 for details), as measured by the trace distance, one can show that

$$T(\rho_A(t+1), \rho_A(t)) \lesssim \sqrt{\frac{\eta^2}{4}(\nabla_{\boldsymbol{\theta}} E)^T \mathcal{F}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} E}, \tag{3.8}$$

where $\mathcal{F}_{i,j}(\boldsymbol{\theta}) = 4\,\mathrm{Re}[\langle\partial_i\psi|\partial_j\psi\rangle - \langle\partial_i\psi|\psi\rangle\langle\psi|\partial_j\psi\rangle]$ is the quantum Fisher information matrix (QFIM) [Mey21] and η is the learning rate. Inequalities (3.7)-(3.8) imply that the learning rate η can be used to limit the maximal possible change of the purity.[1] Provided that the change in purity is sufficiently small, the Taylor expansion can be used to argue that the corresponding change in the second Rényi entropy S_2, related to the purity as $e^{-S_2} = \operatorname{tr}\rho_A^2$, also remains controlled. Therefore, the choice of an appropriately small learning rate can guarantee the avoidance of a WBP at $t+1$, provided the absence of one at t.

To illustrate the bound numerically, we prepare an initialization outside of the WBP using a small angle parameter $\epsilon_{\boldsymbol{\theta}}$ and compute the change in the purity $\operatorname{tr}\rho_A^2$ after one GD update step for different learning rates η. The results of this procedure for four different learning rates are shown in Fig. 4.4. We see that larger learning rates correspond to a bigger change in purity and are thus more prone to encounter a WBP. At the same time, all data points are below the theoretical bound. While up to the best of our knowledge the bound Eq. (3.7) is not proven to be tight, we observe that points corresponding to the extreme learning rates closely approach the theoretical line.

Using Eq. (3.8), the bound can be efficiently approximated on NISQ hardware: the QFIM can be estimated efficiently on a quantum device using techniques suggested in Ref. [GZCW21] or Ref. [RBMV21] using classical shadows. For the computation of the gradient, one can use the

[1] A similar continuity bound that does not require the QFIM can be found in terms of the maximum operator norm of the gate generators. We acknowledge Johannes Jakob Meyer for this remark.

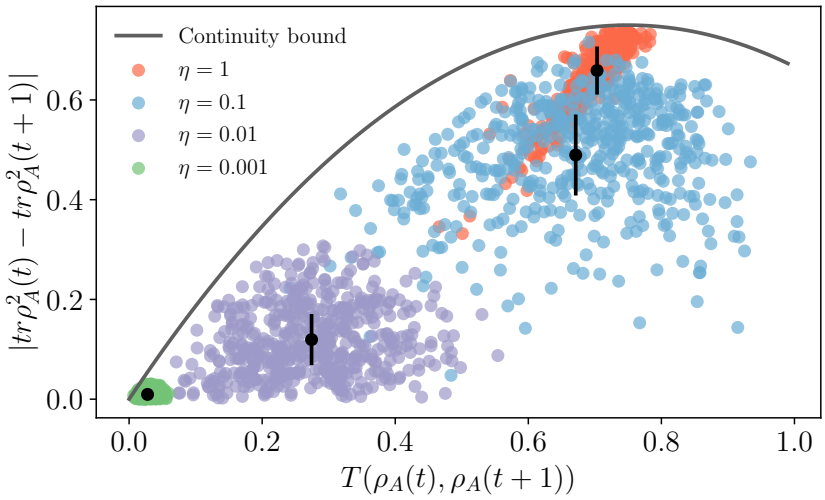

Figure 3.4: We numerically illustrate the continuity bound Eq. (3.7) and its relation to the learning rate η for $t = 0$, i.e. at the beginning of the optimization schedule. This shows that one should be careful with the choice of the learning rate since a large learning rate leads to a big change in the trace distance and a change in purity. We use a system size of $N = 10$ and a random circuit with circuit depth $p = 100$ and small qubit rotations ($\epsilon_\theta = 0.05$) to generate a BP-free initialization. Data was averaged over 500 random instances.

parameter shift rule [MNKF18, SBG+19] also with shadow tomography. The expression can thus be efficiently evaluated on a real device and used together with the continuity bound to estimate a suitable learning rate η. However, in practice, this might not be needed and simply following Algorithm 1 could be more efficient and easier to implement.

3.4.2 Optimization performance with learning rate

Finally, we illustrate Algorithm 1 in practice. To this end, we first prepare a WBP-free initial state using small qubit rotation angles and compare the performance of GD optimization with different learning rates. If we start with a large learning rate, $\eta = 1$, corresponding to red lines in Fig. 3.5 (a)-(c), we see that the energy expectation value in Fig. 3.5 (a) rapidly (within one or two update steps) converges to a value far away from the target ground state energy E_{GS}. At the same time, panel (b) reveals that this can be attributed to an onset of a WBP, as the second Rényi entropy spikes up to the Page value. Finally, panel (c) shows that the gradient norm also is convergent, though at a non-zero value. We attribute this to the fact that the system gets trapped in the WBP region.

As suggested by Algorithm 1, we thus decrease the learning rate to $\eta = 0.1$ and start again. This time a WBP is avoided, the algorithm however gets stuck in a local minimum with large entanglement entropy. In this instance a choice of parameter α that defines an onset of a WBP in Def. 2 being smaller than one may be beneficial. For instance, setting $\alpha = 0.5$ could help avoid the suboptimal local minima characterized by large entanglement, see the grey dashed line in Fig. 3.5 (b). Note that the large gradient persistent after many iterations for the blue line in Fig. 3.5 (c) may also indicate that the learning rate is chosen too large for the width of the local minima.

Provided that our algorithm uses $\alpha = 0.5$, the system would satisfy a WBP condition even for learning rate $\eta = 0.1$, forcing us to restart the algorithm with an even smaller learning rate. Setting $\eta = 0.01$, we see that the algorithm is now able to converge very close to the true

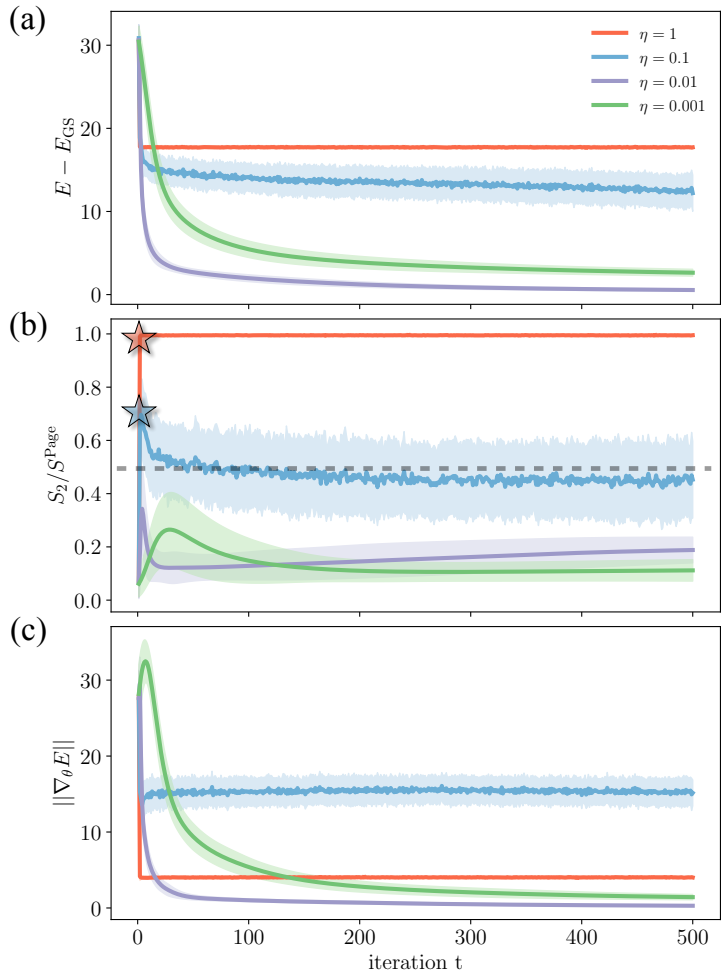

Figure 3.5: (a-c) The application of the proposed Algorithm to the problem of finding the ground state of the Heisenberg model. For large learning rates $\eta = 1$ and 0.1 (red and blue lines) the optimization gets into a large entanglement region as is shown in panel (b), indicated by colored stars, forcing the restart of the optimization with a smaller value of η. For $\eta = 0.01$ the algorithm avoids large entanglement regions and gets a good approximation for the ground state. Finally, setting even smaller learning rate (green lines) degrades the performance. The normalized second Rényi entropy of the true ground state is $S_2/S^{\text{Page}}(k, N) \approx 0.246$. (c) Shows the corresponding gradient norm. A small gradient norm equally corresponds to the BP and the good local minima found with $\eta = 0.01$ and 0.001. We use a system size of $N = 10$, subsystem size $k = 2$ and a random circuit (see Eq. (3.1)) with circuit depth $p = 100$ and small qubit rotations ($\epsilon_\theta = 0.05$) to generate a BP-free initialization. Here we choose $\alpha = 0.5$ indicated by the grey dashed line, see the last paragraph of Sec. 3.3.1 for a discussion on the choice of α. Data was averaged over 100 random instances.

ground state energy (violet line in Fig. 3.5 (a)-(c)). In particular, the norm of the gradient assumes the smallest value among all learning rates. We note, that the further decrease of the learning rate (i.e. to $\eta = 0.001$) degrades the performance of GD. While WBPs are not encountered during the optimization process, the GD optimization converges slower and within the given iterations to a larger energy expectation value. This highlights the fact that it is best to choose the highest possible learning rate, that still avoids a WBP. We speculate, that an optimization strategy that adapts the learning rate at each optimization step would give the best performance, though testing this assumption is beyond the scope of the present work.

3.4.3 Classical simulatability and performance comparison

Now that we have illustrated the procedure outlined in Algorithm 1 in detail, let us comment on the restrictions that our Algorithm imposes, its relation to classical simulatability and finally compare our method with other common means for mitigating BPs.

To avoid WBPs and thus BPs we require that the second Rényi entropy of a small subregion is less than a fraction α of the Page value, where $\alpha \in (0, 1]$ and the default choice is $\alpha = 1$. This definition does restrict the amount of entanglement generated by the circuit and thus does imply the classical simulatability of the circuit. Indeed, it is the scaling of the entanglement entropy with system size that is important for classical simulatability of a quantum system. Only in the special case when the entanglement entropy of the quantum state scales poly-logarithmically with the number of qubits, we can simulate the states on a classical computer in polynomial time [Vid03, VdNDVB07, BH13]. In contrast, the criteria for WBP, Def. 2 is generally consistent with volume-law entanglement as we illustrate below, thus allowing our algorithm to be applied to systems that cannot be efficiently simulated on a classical computer.

Here we focus on two types of systems: namely systems where the ground state satisfies area-law, which implies that the entanglement entropy of an arbitary bipartition of the state scales with the size of the boundary $S(\rho_A) \sim |\partial A|$, as well as volume-law, which implies that it scales with the volume, $S(\rho_A) \sim |A|$ (see Ref. [ECP10] for a review on these concepts). For area-law states in 1D, the entanglement entropy is constant and therefore allows for an efficient classical representation using techniques such as matrix product states [Sch11]. The 1D Heisenberg model, considered in the previous subsection, is an example of such a system.

The Heisenberg model, however, can be made hard to simulate classically by considering a random graph geometry illustrated in Fig. 3.6 (a), instead of a 1D chain. This leads to non-local interactions and a volume-law entanglement scaling for a typical bipartite cut. Due to the non-local nature of the model we choose $\alpha = 1$ since we have no prior knowledge on the entanglement properties of the ground state. We again use the small-angle initialization [HSCC21, HBK21] to generate a BP-free initial state. We compare this with layerwise optimization [SMM$^+$20] which is another common heuristic for avoiding BPs. There the circuit is initialized with a single layer which is optimized, the circuit is then grown by one layer at a time and optimized while keeping the parameters in the previous layers constant.

Fig. 3.6 (b)-(c) reveals that for the Heisenberg model on a graph, layerwise optimization ends up in a WBP during the optimization for both learning rates that we considered. The small-angle initialization successfully avoids the WBP for both learning rates, however, good convergence is only achieved with $\eta = 0.01$. This is similar to the situation encountered in the Heisenberg model in 1D, see Fig. 3.6, where a too-large learning rate prevents convergence to the basin of attraction of the local minimum. Likewise, to the case of 1D Heisenberg model, the fact that the learning rate $\eta = 0.1$ does not lead to convergence to a minimum can be revealed through the norm of the gradient which stays large even after 500 iterations.

In addition to the Heisenberg model on the random graph, we also considered the Sachdev–Ye–Kitaev (SYK) model [Kit15] that features a volume-law entangled ground state [HG19]. In Appendix B.5 we illustrate that our method is also successful in preventing the BP occurrence and results in finding the SYK ground state.

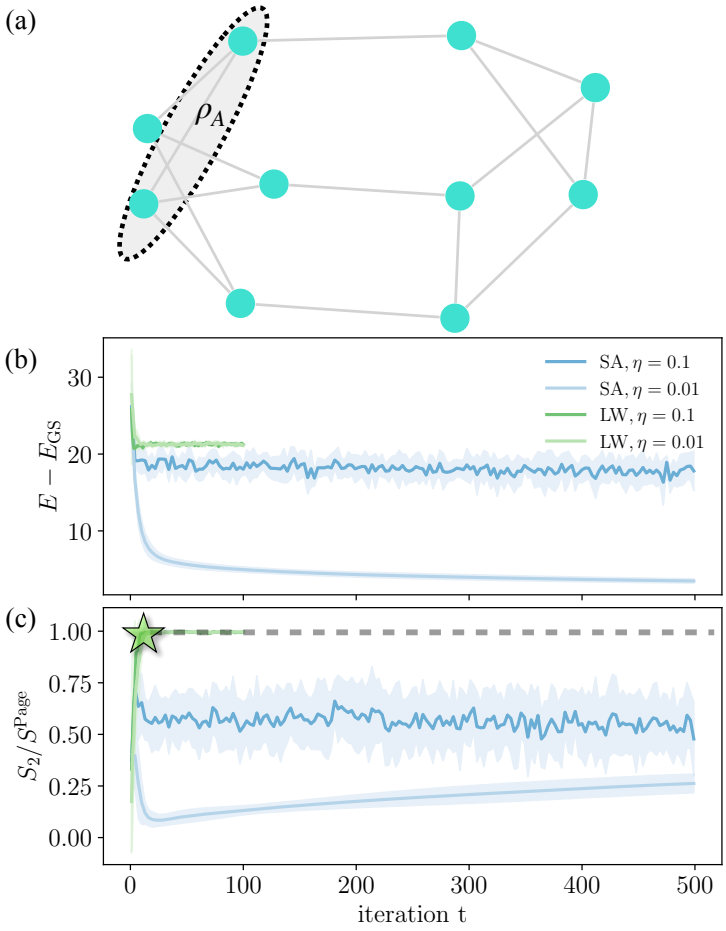

Figure 3.6: The application of our Algorithm to the problem of finding the ground state for the Heisenberg model on a 3-regular random graph depicted in (a). Panel (b) shows the energy as a function of GD iterations t and panel (c) illustrates the second Rényi entropy of two-spin region A with $k = 2$ shown in panel (a). Since the interactions are now non-local and we do not have any prior knowledge on the entanglement properties of the target state we set $\alpha = 1$ (gray dashed line). For the initialization, we use the small-angle initialization (SA) with $\epsilon_\theta = 0.1$ and compare it to layerwise optimization (LW). LW encounters a WBP for both learning rates that we considered (green star). In contrast, SA avoids the WBP for both learning rates. Good performance and further convergence in the local minimum is only achieved trough a smaller learning rate of $\eta = 0.01$. We use a system size of $N = 10$ and a random circuit from Eq. (3.1) with circuit depth $p = 100$. Data is averaged over 100 random instances.

3.5 Summary and Discussion

The main result of this work is the introduction of the concept of WBPs, which in essence provide an efficiently detectable version of BPs. In particular, we propose to use the classical shadows protocol to estimate the second Rényi entropy of a small subregions that are independent of system size. If these subregions avoid nearly maximal entanglement – a condition sufficient for avoiding WBPs – the system also avoids conventional BPs. Building on this definition of the WBP, we proposed an algorithm that is capable of avoiding BPs on NISQ devices without requiring a computational overhead that scales exponentially in system size.

To illustrate the notion of WBPs and the proposed algorithm, we studied a particular BP-free initialization of the variational quantum eigensolver. Furthermore, we considered an

optimization procedure that uses gradient descent. Phenomenologically, we observed that the encounter of a BP during the optimization crucially depends on the learning rate, which controls the parameter update magnitude between consecutive optimization steps. A smaller learning rate is less likely to lead to the encounter of a BP during the optimization. However, choosing the learning rate to be very small degrades the performance of GD. These results support the feasibility of the proposed algorithm for efficiently avoiding BPs on NISQ devices. While our results and numerical simulations are focused on variational quantum eigensolvers (VQE), they readily extend to other variational hybrid algorithms, such as quantum machine learning [BLSF19a, HCT$^+$19, SBSW20], quantum optimization [FGG14, SS21b, Har21a] or variational time evolution [BVC21, LDG$^+$21].

Although the issue of avoiding BPs at the circuit initialization is a subject of active research [GWOB19, DBW$^+$21, SMM$^+$20, HSCC21, LCS$^+$21], the influence and role of BPs in the optimization process has received much less attention [LJGM$^+$21]. Our results indicate that entanglement, in addition to playing a crucial role in circumventing BPs at the launch of the VQE, is also important for achieving a good optimization performance. In addition, our heuristic results in Sec. 3.4 suggest that post-selection based on the entanglement of small subregions may help to avoid low-quality local minima that are characterized by higher entanglement. Algorithm 1 allows for such post-selection by appropriately tuning the value of α. Doing so, however, requires some prior knowledge about the entanglement structure of the target state. This may be inferred from the structure of the Hamiltonian (for instance, for a Hamiltonian that is diagonal in the computational basis, the eigenstates are product states with no entanglement), or by targeting small instances of the computational problem using exact diagonalization.

Beyond that, one could imagine an algorithm where the learning rate is not only adapted when a WBP is encountered but dynamically adjusted at every step of the optimization process. This may allow for efficiently maneuvering complicated optimization landscapes by staying clear of highly entangled local minima. VQE, for instance, is known to have many local minima [BK21], but a systematic study of their entanglement structure, required for devising such dynamic entanglement post-selection procedure, has yet to be done.

Another important question concerns the effect of noise, which has been suggested to be an additional source for the emergence of BPs [WFC$^+$20]. Noise cannot be avoided on NISQ machines and has a profound impact on any near-term quantum algorithm which is difficult to analyze analytically. Fortunately, none of the tools we propose are especially susceptible to noise corruption. In fact, both the classical shadow protocol and the estimation of observables and purities are stable with respect to the addition of a small but finite amount of noise, and there have even been some proposals for noise mitigation techniques [CYZF21, EG20].

Finally, we comment on the possibility of testing Algorithm 1 on a real NISQ device. While the shadows protocol can readily be implemented on near-term devices to diagnose WBPs, whether a variational circuit with enough entangling layers that lead to a BP can be realized on an NISQ device is not entirely clear at this stage. Nevertheless recent results of Ref. [Mi 21] observed convergence of the out-of-time correlators to zero, indicating that a 2-design might already have been reached. This implies that large entanglement, as present in a BP, could be realizable on available NISQ devices, and opens the door to experimental studies of the effect of entanglement on the optimization performance of current NISQ machines using the proposed shadows protocol.

CHAPTER 4

Recursive greedy initialization of the QAOA with guaranteed improvement

In this chapter, we investigate the quantum approximate optimization algorithm (QAOA), focusing on how the choice of parameter initialization and optimization strategy affects the algorithm's performance. We begin by analytically constructing index-1 saddle points at circuit depth $p+1$, starting from a local minimum at circuit depth p. These index-1 saddle points maintain the same energy as the initial local minimum. Moreover, the presence of a unique local descent direction in parameter space ensures a decrease in the cost function, thereby providing a guaranteed improvement at each circuit depth. This section is based on the paper:

Stefan H. Sack, Raimel A. Medina, Richard Kueng, and Maksym Serbyn. Recursive greedy initialization of the quantum approximate optimization algorithm with guaranteed improvement. *Phys. Rev. A*, 107:062404, Jun 2023

4.1 Introduction

The Quantum Approximate Optimization Algorithm (QAOA) [FGG14] is a prospective near-term quantum algorithm for solving hard combinatorial optimization problems on Noisy Intermediate-Scale Quantum (NISQ) [Pre18] devices. In this algorithm, the quantum computer is used to prepare a variational wave function that is updated in an iterative feedback loop with a classical computer to minimize a cost function (the energy expectation value), which encodes the computational problem. A common bottleneck of the QAOA is the convergence of the optimization procedure to one of the many low-quality local minima, whose number increases exponentially with the QAOA circuit depth p [ZWC$^+$20, SS21a].

Much effort has been devoted to finding good initialization strategies to prevent convergence to such low-quality local minima. Researchers have proposed to: first solve a relaxed classical optimization problem and to use that as an initial guess [EMW21], to use machine learning to infer patterns in the optimal parameters [JCKK21], interpolating optimal parameters between different circuit depths [ZWC$^+$20], or to use the parallels between the QAOA and quantum annealing [SS21a]. Recently the success of the interpolation strategies that appeal to annealing was attributed to the ability of the QAOA to effectively speed up adiabatic evolution via the so-called counterdiabatic mechanism [WL22]. This result was used to explain cost

function concentration for typical instances [BBF+18] and parameter concentration [ARCB21] of optimal, typically smoothly varying, parameters.

Despite this progress, all proposed initialization strategies remain heuristic or physically motivated at best, and our understanding of the QAOA optimization remains limited. One of the main puzzles is the exponential improvement of the QAOA performance with circuit depth p, observed numerically [ZWC+20, Cro18]. Here we propose an analytic approach that relates QAOA properties at circuit depths p and $p+1$. The recursive application of our result leads to a QAOA initialization scheme that guarantees improvement of performance with p.

Our analytic approach relies on the consideration of stationary points of QAOA cost function beyond local minima. Inspired by the theory of energy landscapes [Wal04], we focus on stationary configurations with a unique unstable direction, known as *transition states* (TS). We show that $2p+1$ distinct TS can be constructed *analytically* for a QAOA at circuit depth $p+1$ (denoted as QAOA$_{p+1}$) from minima at circuit depth p. All these TS for QAOA$_{p+1}$ exhibit the same energy as the QAOA$_p$-minimum from which they are constructed, thus providing a good initialization for QAOA$_{p+1}$. Descending in the negative curvature direction connects each of the $2p+1$ TS to two local minima of QAOA$_{p+1}$, which are thus guaranteed to exhibit lower energy than the initial minima of QAOA$_p$. Iterating this procedure leads to an exponentially increasing (in p) number of local minima which are guaranteed to have a lower energy at circuit depth $p+1$ than at p [Not]. We visualize this hierarchy of minima and their connections in a graph and propose a Greedy approach to explore its structure. We numerically show that optimal parameters at every circuit depth p are smooth (i.e. the variational parameters change only slowly between circuit layers) and directly connect to a smooth parameter solution at $p+1$ through the TS. Our results explain existing QAOA initializations and establish a recursive analytic approach to study QAOA.

The rest of the paper is organized as follows. In Section 4.2 we review the QAOA, present newly found symmetries, and introduce the analytical construction of TS. In Section 4.3 we show how TS can be used as an initialization to systematically explore the QAOA optimization landscape. From this, we introduce a new heuristic method, dubbed Greedy for exploring the landscape and provide a comparison to popular optimization strategies. Finally, in Section 4.4 we discuss our results and potential future extensions of our work. Appendices C.1-C.6 present detailed proofs of our analytical results, as well as supporting numerical simulations.

4.2 QAOA optimization landscape

4.2.1 MaxCut problem on random regular graphs

The QAOA was originally proposed for a graph partitioning problem, known as finding the maximal cut (MaxCut) [FGG14] and has also been applied to a variety of other optimization problems [SYS+21, FGG20a, MH21]. MaxCut seeks for a partition of the given undirected graph $\mathcal{G}$ into two groups such that the number of edges E that connect vertices from different groups are maximized. Finding the solution of MaxCut for a graph with n vertices is equivalent to finding a ground state for the n-qubit classical Hamiltonian $H_C = \sum_{\langle i,j\rangle \epsilon E} \sigma_i^z \sigma_j^z$, with the sum running over a set of graph edges E and σ_i^z being the Pauli-Z matrix acting on the i-th qubit.

The depth-p QAOA algorithm [FGG14] minimizes the expectation value of the classical Hamiltonian over the variational state $|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle$ with angles $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

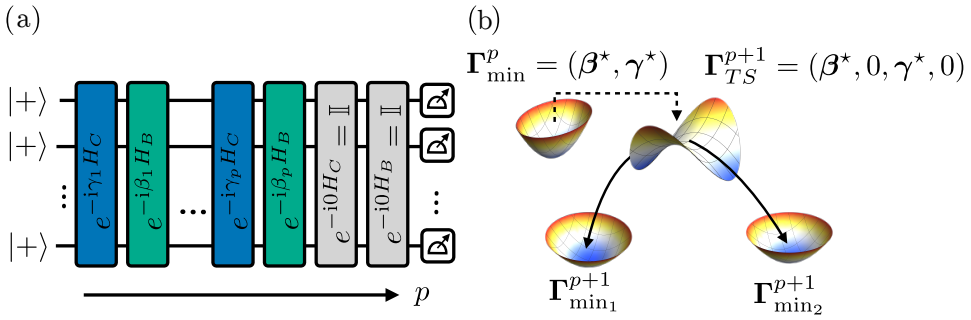

Figure 4.1: (a) Circuit diagram that implements the QAOA ansatz state with circuit depth p, see Eq. (5.2). Gray boxes indicate the identity gates that are inserted when constructing a TS, as indicated in Theorem 3. (b) Local minima $\mathbf{\Gamma}^{p}_{\min}$ of QAOA$_p$ generate a TS $\mathbf{\Gamma}^{p+1}_{\text{TS}}$ for QAOA$_{p+1}$ that connects to two *new local minima*, $\mathbf{\Gamma}^{p+1}_{\min_{1,2}}$ with lower energy.

shown in Fig. 4.1(a):

$$|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle = \prod_{i=1}^{p} e^{-\beta_i H_B} e^{-\gamma_i H_C} |+\rangle^{\otimes n} . \tag{4.1}$$

Here $H_B = -\sum_i^n \sigma_i^x$ is the mixing Hamiltonian and the circuit depth p controls the number of applications of the classical and mixing Hamiltonian. The initial product state $|+\rangle^{\otimes n}$, where all qubits point in the x-direction is an equal superposition of all possible graph partitions which is also the ground state of H_B.

Finding the minimum of $E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle \boldsymbol{\beta}, \boldsymbol{\gamma}| H_C |\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle$ over angles $(\beta_1, \ldots, \beta_p)$ and $(\gamma_1, \ldots, \gamma_p)$ that form a set of $2p$ variational parameters, $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, yields a desired approximation to the ground state of H_C, equivalent to an approximate a solution of MaxCut. The scalar function $E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ thus defines a $2p$-dimensional energy landscape where the QAOA seeks to find the best minimum. The performance of the QAOA is typically reported in terms of the approximation ratio $r_{\beta,\gamma} = E(\boldsymbol{\beta}, \boldsymbol{\gamma})/C_{\min}$, where $C_{\min}$ is the cost function value for the MaxCut. Symmetries of the QAOA ansatz when restricted to graphs with only odd connectivity, such as random 3-regular graphs (RRG3) used in this work, restrict the parameter range to the following fundamental region:

$$\beta_i \in [-\frac{\pi}{4}, \frac{\pi}{4}]; \;\; \gamma_1 \in (0, \frac{\pi}{4}), \;\; \gamma_j \in [-\frac{\pi}{4}, \frac{\pi}{4}], \tag{4.2}$$

with $i \in [1, p]$ and $j \in [2, p]$. Note that the fundamental region presented above is smaller than what has been previously reported [ZWC+20, WHJR18], see the Appendix C.1 for details.

4.2.2 Energy minima and transition states

Previous studies of the QAOA landscape were restricted to local minima of the cost function $E(\boldsymbol{\beta}, \boldsymbol{\gamma})$, since they can be directly obtained using standard gradient-based or gradient-free optimization routines. Local minima are stationary points of the energy landscape (defined as $\partial_i E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ for derivative running over all $i = 1, \ldots, 2p$ variational angles), where all eigenvalues of the Hessian matrix $H_{ij} = \partial_i \partial_j E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are positive, that is the Hessian at the local minimum is positive-definite. However, the study of energy landscapes [Wal04] of chemical reactions and molecular dynamics has shown that TS, which corresponds to stationary points with a single negative eigenvalue of the Hessian matrix (index-1), also plays an important role [1]. There, TS are particularly relevant as they correspond to the

[1] Note, that on physical grounds we do not consider singular Hessians that have one or more vanishing eigenvalues, see Appendix C.2.

highest-energy configurations along a reaction pathway. They often serve as bottlenecks in the reaction process and thus are crucial for understanding reaction rates, designing catalysts, and predicting chemical behavior. By studying the role of transition states in the QAOA landscape, we aim to uncover insights that could lead to improved optimization strategies or better convergence properties of the algorithm. This motivates the construction of TS achieved below.

4.2.3 Analytic construction of transition states

The structure of the QAOA variational ansatz allows us to analytically construct the TS of QAOA$_{p+1}$ using any local minima of QAOA$_p$:

Theorem 3 (TS construction, simplified version)**.** *Assume that we found a local minimum of QAOA$_p$ denoted as $\mathbf{\Gamma}^p_{\min} = (\boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star) = (\beta^\star_1, \ldots, \beta^\star_p, \gamma^\star_1, \ldots, \gamma^\star_p)$. Padding the vector of variational angles with zeros at positions i and j, results in*

$$\begin{aligned}\mathbf{\Gamma}^{p+1}_{TS}(i,j) = (\beta^\star_1, ..., \beta^\star_{j-1}, 0, \beta^\star_j, ..., \beta^\star_p,\\ \gamma^\star_1, ..., \gamma^\star_{i-1}, 0, \gamma^\star_i, ..., \gamma^\star_p)\end{aligned} \tag{4.3}$$

being a TS for QAOA$_{p+1}$ when $j = i$ or $j = i+1$ and $\forall i \in [1, p]$, and also for $i = j = p+1$.

Proof. The argument consists of two steps. First, by relating the first derivative over newly introduced parameters to derivatives over existing angles we show that Eq. (4.3) is a stationary point of QAOA$_{p+1}$. More specifically, we observe that the gradient components where the zero insertion is made satisfy the following relations

$$\begin{aligned}\partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l)} &= \partial_{\beta_{l-1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}},\\ \partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l+1)} &= \partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}},\\ \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l)} &= \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}},\\ \partial_{\gamma_{l+1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l+1)} &= \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\Big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}}.\end{aligned} \tag{4.4}$$

Since $\nabla E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}^p_{\mathsf{min}}} = 0$, it directly follows that the TS constructed using Theorem 3 are also stationary points. In the second step, we show that the Hessian at the TS has a single negative eigenvalue. To this end in the Appendix C.2 we show that we can always write the Hessian at the TS in the following form

$$H(\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,k)) = \begin{pmatrix} H(\mathbf{\Gamma}^p_{\mathsf{min}}) & v(l,k) \\ v^T(l,k) & h(l,k)\end{pmatrix}, \tag{4.5}$$

where $H(\mathbf{\Gamma}^p_{\mathsf{min}}) \in \mathbb{R}^{2p\times 2p}$, $v(l,k) \in \mathbb{R}^{2p\times 2}$ and, $h(l,k) \in \mathbb{R}^{2\times 2}$. Here, the largest block $H(\mathbf{\Gamma}^p_{\mathsf{min}})$ corresponds to the old Hessian at the stationary point. The matrix $h(l,k)$ corresponds to the second derivatives of the energy with respect to new parameters that are initially set to zero, whereas matrix $v(l,k)$ represents the "mixing" terms, with one derivative taken over the old parameters and the second derivative corresponds to one of the new parameters, which are initialized at zero. By employing this representation of the Hessian at the TS, we utilize the Eigenvalue Interlacing theorem ([Ref. [Bel97], Theorem 4 on page 117] summarized in

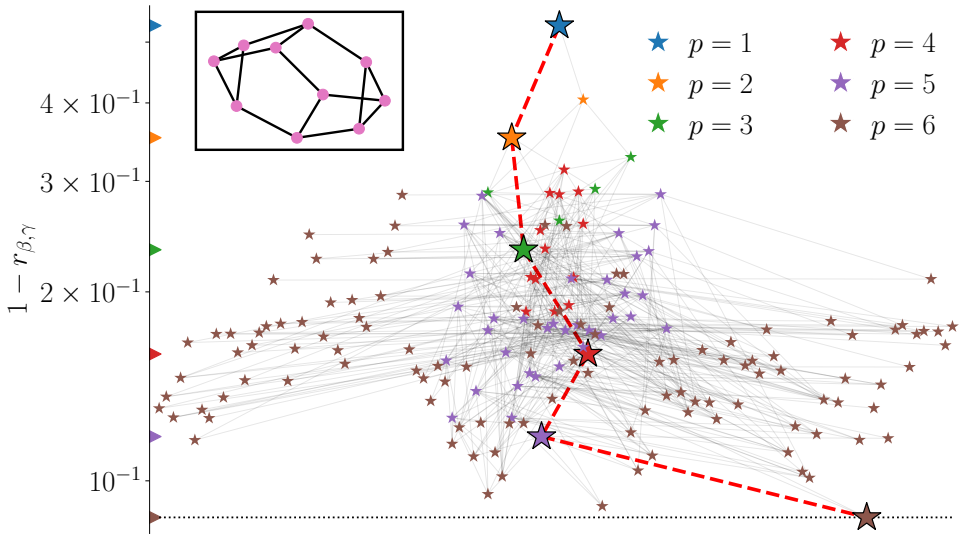

Figure 4.2: Initialization graph for the QAOA for MaxCut problem on a particular instance of RRG3 with $n = 10$ vertices (inset). For each local minima of QAOA$_p$ we generate $p+1$ TS for QAOA$_{p+1}$, find corresponding minima as in Fig. 4.1(b), and show them on the plot connected by an edge to the original minima of QAOA$_{p+1}$. Position along the vertical axis quantifies the performance of QAOA via the approximation ratio, points are displaced on the horizontal axis for clarity. Color encodes the depth of the QAOA circuit, and large symbols along with the red dashed line indicate the path that is taken by the Greedy procedure that keeps the best minima for any given p resulting in an exponential improvement of the performance with p. The Greedy minimum coincides with an estimate of the global minimum for $p = 6$ (dashed line) obtained by choosing the best minima from 2^p initializations on a regular grid.

Theorem 8) to establish that $H(\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,k))$ has at most two negative eigenvalues. Subsequently, we prove that the determinant of $H(\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,k))$ is negative for each of the $2p+1$ transition states, which implies the presence of only one negative eigenvalue (i.e., the index-1 direction). It is important to note that this result is independent of the choice of classical Hamiltonian, which is fixed to encode MaxCut in this work. □

The simplified theorem above ignores the possibility of vanishing eigenvalues of the Hessian, which can be ruled out only on physical grounds. This issue and complete proof of the theorem are discussed in Appendix C.2.

4.3 From transition states to QAOA intialization

4.3.1 Initialization graph

For each local minimum of QAOA$_p$, Theorem 3 provides $p+1$ symmetric TS where zeros are padded at the same position, $i = j$, like in Fig. 4.1(a), and additionally p non-symmetric TS with $j = i+1$, where zeros are padded in adjacent layers of the QAOA circuit. Fig. 4.1(b) shows how one can descend from a given TS along the positive and negative index-1 direction, finding two new local minima of QAOA$_{p+1}$ with lower energy. Thus Theorem 3 provides us with a powerful tool to systematically explore the local minima in the QAOA in a recursive fashion.

Such exploration of the QAOA initializations for a particular graph with $n = 10$ vertices is summarized in Fig. 4.2. We find a unique minimum for QAOA$_1$ using grid search (see Appendix C.5) in the fundamental region defined in Eq. C.52 from which we construct two symmetric TS according to Eq. (4.3), descend from these TS in index-1 directions with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) [Bro70, Fle70, Gol70, Sha70] algorithm, finding two new local minima of QAOA$_2$. These minima are connected to the minima of QAOA$_1$, since it was used to construct a TS. Repeating this procedure recursively for each of the $p + 1$ symmetric TS [2] we obtain the tree in Fig. 4.2. Assuming that all minima found in this way from symmetric TS are unique, their number would increase as $2^{p-1}p!$. Numerically, we observe that the number of unique minima is much smaller compared to the naïve counting, increasing approximately exponentially with p.

4.3.2 Greedy maneuvering through the graph

The exponential growth of the number of minima in QAOA depth p makes the naïve construction and exploration of the full graph a challenging task. To deal with the rapidly growing number of minima we introduce:

Corollary 3.1 (Greedy recursive strategy)**.** *Using the lowest energy minimum that is found for QAOA depth p, we generate $2p + 1$ transition states (TS) for QAOA$_{p+1}$. Each transition state corresponds to the same state in the Hilbert space as the initial local minimum, so the energy of all the transition states is the same and equal to the energy of the initial local minimum. We then optimize the QAOA parameters starting from each of these transition states and select the best new local minimum of QAOA$_{p+1}$ to iterate this procedure. This* Greedy *recursive strategy is guaranteed to lower energy at every step.*

Proof. Let the initial local minimum at QAOA depth p have energy E_p. Since all the $2p + 1$ transition states are generated from this minimum and have the same energy E_p, when we optimize the QAOA parameters for QAOA$_{p+1}$ starting from these transition states, all the converged local minima will have energy less than or equal to E_p. As a result, the energy can either decrease or stay the same (provided that curvature vanishes, which we do not expect on physical grounds, see Appendix C.2), but it cannot increase. Therefore, the Greedy recursive strategy is guaranteed to lower or maintain the energy at every step. □

The Greedy path that is taken by this strategy in the initialization graph is shown in Fig. 4.2 as a red dashed line. We can see that this heuristic allows to very effectively maneuver the increasingly complex graph with its numerous local minima and find the global minimum for circuit depths up to $p = 7$. A detailed description of the algorithm is presented in Appendix C.5.

To systematically explore how Greedy maneuvers the initialization graph, we compare it to two initialization strategies proposed in the literature: The so-called Interp approach [ZWC+20] interpolates the optimal parameters found for circuit depth p to $p + 1$ and uses it as a subsequent initialization. This procedure creates a *smooth parameter pattern* that mimics an annealing schedule. Numerical studies demonstrated that Interp has the same performance as the best out of 2^p random initializations. The second method that we use for comparison is the Trotterized Quantum Annealing (TQA) method [SS21a], that initializes QAOA$_p$ using

[2] Note, that we restrict only to symmetric TS since we numerically find no performance gain from including the non-symmetric TS in the initialization procedure.

$\gamma_j = (1 - \frac{j}{p})\Delta t$ and $\beta_j = \frac{j}{p}\Delta t$. The step size Δt is a free parameter determined in a pre-optimization step. The TQA has similar performance to INTERP at moderate circuit depths, notably having lower computational cost. Obtaining an initialization for QAOA$_p$ within the INTERP framework requires running the optimization for all $p' = 1, \ldots, p-1$, while in the TQA the search for an optimal Δt is performed directly for a given p.

Fig. 4.3 reveals that the GREEDY approach yields similar performance to existing methods. Moreover, the performance of TQA slightly degrades at higher p, however, GREEDY is fully on par with INTERP initialization. The comparable performance between GREEDY and earlier heuristic approaches is surprising. Indeed, the GREEDY method for QAOA$_p$ explores $p+1$ symmetric TSs and chooses the best out of the resulting up to $2(p+1)$ minima (if none are equivalent), in contrast to INTERP, which uses a single smooth initialization pattern at every p and thus at a smaller computational cost.

4.3.3 Smooth pattern of variational angles and heuristic initializations

We find that having a smooth dependence of the variational angles on p (referred to as a "smooth pattern") is an important characteristic for efficiently maneuvering the initialization graph. A smooth pattern means that the variational angles change gradually and continuously as the QAOA depth p increases, without abrupt jumps or discontinuities. This smoothness property can be visually inspected by plotting the variational angles as a function of p and observing whether the curve appears continuous and smooth. Assuming we found a smooth pattern of QAOA$_p$, Theorem 3 produces a TS of QAOA$_{p+1}$ by padding it with zeros, effectively introducing a discontinuity (bump). Optimization from the TS with such a bump can proceed by rolling down either side of the saddle, see Fig. 4.4(a), finding two new minima. Remarkably, the eigenvector corresponding to the index-1 direction of the Hessian has dominant weight

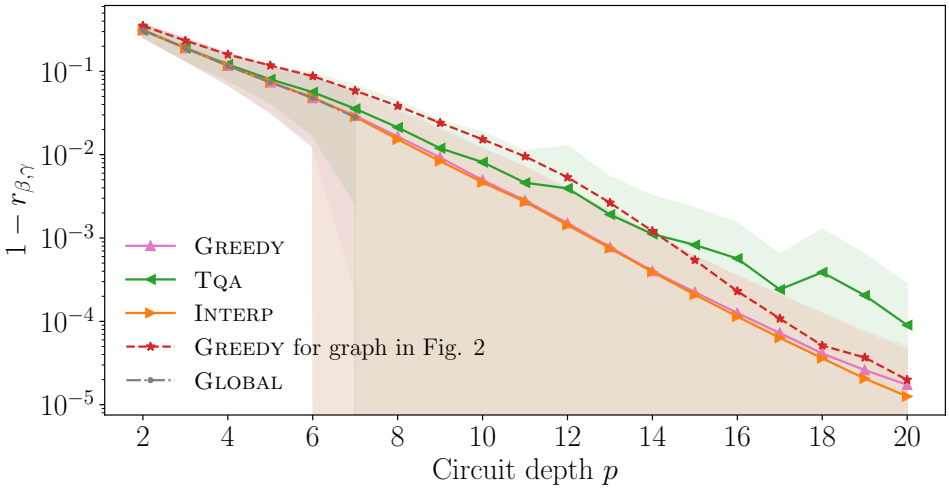

Figure 4.3: Performance comparison between different QAOA initialization strategies used for avoiding low-quality local minima. GREEDY approach proposed in this work yields the same performance as INTERP [ZWC+20] and slightly outperforms TQA [SS21a] at large p. GLOBAL refers to the best minima found out of 2^p initializations on a regular grid. Data is averaged over 19 non-isomorphic RRG3 with $n = 10$, shading indicates standard deviation. System size scaling for up to $n = 16$ and performance comparison for different graph ensembles can be found in the Appendix C.6.

on the variational angles with initially zero value, see C.4 for details. Thus descending along the index-1 direction, we can either enhance or heal the resulting discontinuity in the pattern of variational angles. As a result, among two new local minima of QAOA$_{p+1}$ one typically exhibits a smooth parameter pattern where the bump was removed, while the other minimum has an enhanced discontinuity, see Fig. 4.4(b) for an example. Utilizing these observations in a numerical study, we find that minima exhibiting a non-smooth parameter pattern exhibit usually a worse or the same performance as smooth minima. In fact, in the Greedy procedure we find that in most cases, in particular at the beginning of the protocol, smooth minima are selected. However, there are cases where a non-smooth minimum is selected if it exhibits the same energy as the smooth one. Greedy then branches off in the optimization graph into a sub-graph involving only non-smooth minima. Usually, this process of branching off is followed by a smaller gain in performance from increasing p.

The preferred smoothness of QAOA optimization parameters has been explored in the literature [ZWC$^+$20, MBS$^+$22, WL22] and is believed to be linked to quantum annealing [BBB$^+$21] (QA). In QA the ground state of the Hamiltonian H_C is obtained by preparing the ground state of H_B and smoothly evolving the system to H_C such that the system remains in the ground state during the evolution. A fast change, as generated by a bump in the protocol, leads to leakage into excited energy levels and thus decreased overlap with the target ground state of H_C. Since the QAOA can be understood as a Trotterized version of QA [FGG14, SS21a, ZWC$^+$20], for large p, we believe that a similar process is present in the QAOA and thus makes a smooth parameter pattern preferable.

We find that smooth Greedy minima coincide with Interp minima as shown in Fig. 4.4(b). The Interp naturally creates a smooth parameter pattern since the minima found at p is interpolated to a QAOA$_{p+1}$ initialization. The optimizer only slightly alters the parameters from its initial value, as can be seen in Fig. 4.4(b). Geometrically, the Interp initialization can be obtained from the symmetric TS constructed by Theorem 3 as $\mathbf{\Gamma}^{p+1}_{\text{Interp}} = \frac{1}{p}\sum_{i=1}^{p+1} \mathbf{\Gamma}^{p+1}_{\text{TS}}(i,i)$. In other words, $\mathbf{\Gamma}^{p+1}_{\text{Interp}}$ is the *rescaled center of mass point* of all symmetric TS, with the rescaling factor $(p+1)/p$ being physically motivated. Considering the center of mass of all TS smoothens out discontinuities present in individual TS. The re-scaling is related to the notion of "total time" of the QAOA, given by the sum of all variational angles, $T = \sum_j |\gamma_j| + |\beta_j|$ [ZWC$^+$20, LLL20], that resembles the total annealing time in the limit $p \to \infty$. This parameter has been shown to scale as $T \sim p$ [SS21a], naturally explaining the role of factor $(p+1)/p$ in yielding the correct increased total time of QAOA$_{p+1}$. In other words, the Interp strategy seems to essentially execute a Greedy search without optimizing in the index-1 direction from the TS. This insight lends credence to the success of Interp. However, only Greedy offers a guarantee for performance improvement with increasing p, while for Interp this behavior is supported only by numerical simulations.

4.4 Discussion

In this work, we analytically demonstrated that minima of QAOA$_p$ can be used to obtain transition states (TS) for QAOA$_{p+1}$ which are stationary points with a unique negative eigenvalue in the Hessian. These TS provide an excellent initialization for QAOA$_{p+1}$, because they connect to two new local minima with lower energy. This construction allows us to visualize how local minima emerge at different energies for increasing circuit depth using an initialization graph. Categorizing the local minima on this graph by their smooth (discontinuous) patterns of variational parameters, we find that the smooth minima achieve the best performance.

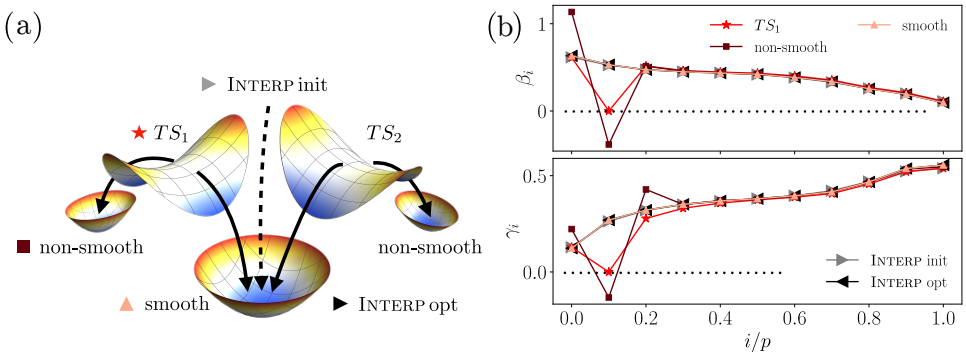

Figure 4.4: (a) Cartoon of descent from two different TS at of QAOA$_{p+1}$ generated from a QAOA$_p$ minimum with a smooth pattern leads to the same new smooth pattern minima of QAOA$_{p+1}$, also reached from the INTERP [ZWC+20] initialization. Two additional non-smooth local minima typically have higher energy. (b) shows the corresponding initial and convergent parameter patterns for the RRG3 graph shown in Fig. 4.2 for $p = 10$.

Incorporating the smooth nature of minima allows us to establish a relation between the GREEDY approach for the exploration of the initialization graph and the best available initialization strategy [ZWC+20].

The use of TS and their analytic construction for the study of QAOA provide the first steps towards an in-depth understanding of the full optimization landscape of the QAOA. The constructed TS are guaranteed to provide an initialization that improves the QAOA performance, suggesting that our construction may be useful for establishing analytic QAOA performance guarantees [FGG14, WL21, FGG20a] for large p in a recursive fashion. Of particular interest is here an analytical understanding of the numerically observed exponential performance improvement with circuit depth. On a practical side, the established relation between heuristic initializations [ZWC+20] and GREEDY exploration of TS suggests that our construction of TS may be useful as a starting point for constructing simple initialization strategies in a broader class of quantum variational algorithms, such as the variational quantum eigensolver [KMT+17, PMS+14] and quantum machine learning [BLSF19b].

In addition, our results invite a more complete characterization of the QAOA landscape using the energy landscapes perspective [Wal04]. What fraction of minima does our procedure find out of the complete set of QAOA local minima? Are there more TS and are our analytically constructed TS typical? How is the Hessian spectrum distributed at these minima and TS? How do these properties depend on the choice of the QAOA classical Hamiltonian, in particular for classical problems with intrinsically hard landscapes [CLSS21]? Answering these and related questions will most likely lead to practical ways of further speeding up the QAOA by reducing the overhead of the classical optimization [WVG+22].

CHAPTER 5

A Recursive Lower Bound on the Energy Improvement of the Quantum Approximate Optimization Algorithm

In this Chapter, building on previous results [SMKS23], we provide a lower bound on the energy improvement of the quantum approximate optimization algorithm (QAOA) between consecutive circuit depths p and $p+1$. We first discuss the construction of transition states which are stationary points of the QAOA with a unique negative curvature direction. We then construct an analytic estimate of the negative Hessian eigenvalue and corresponding eigenvector at each transition state, which enables us to obtain an analytical lower bound on the improvement of the cost function, and to reduce the cost of optimization by bypassing the need to construct and diagonalize the Hessian matrix. Finally, we numerically verify the accuracy of our estimates. Although the obtained energy lower bound underestimates the improvement of the cost function, we find it shows an exponential decrease with the number of layers p. This section is based on the preprint:

Raimel A. Medina and Maksym Serbyn. A Recursive Lower Bound on the Energy Improvement of the Quantum Approximate Optimization Algorithm. *arXiv*, 2405.10125, May 2024

5.1 Introduction

Variational quantum algorithms [CAB+21, BCLK+22] have emerged as a promising approach to leveraging the capabilities of noisy intermediate-scale quantum (NISQ) devices [Pre18]. Among these algorithms, the Quantum Approximate Optimization Algorithm (QAOA) [FGG14] and the Variational Quantum Eigensolver (VQE) [PMS+14] stand out due to their potential for solving optimization problems and quantum chemistry simulations, respectively. The idea is to use the quantum computer in a feedback loop with a classical computer, where it implements a variational wave function that is measured to compute the value of the so-called cost function. This information is then fed into a classical computer where it is processed and the variational wave function is subsequently updated aiming to find a minimum of the cost function, which provides an (approximate) solution to the computationally hard problem.

In the QAOA, the state is prepared by a p-level circuit specified by $2p$ variational parameters. It was shown that even at the lowest circuit depth $p = 1$, QAOA has non-trivial provable

performance guarantees [FGG14, FGG15]. The existence of known analytical performance guarantees makes the QAOA — in contrast to the VQE — a reference algorithm to explore quantum speedups on NISQ devices. In particular, the QAOA has been the subject of both analytical studies [FGG20b, BM21, BM22, BBF$^+$18, WL21, ZBM24, BGMZ22, BFM$^+$22, SHS$^+$24, YBL20, ZTB$^+$22] and practical implementations [Har21b, WVG$^+$22, WSW24, E$^+$22] for small values of the circuit depth, p. These studies suggest that significant gains can be expected as p increases, particularly when $p \geq \ln N$, with N representing the number of qubits involved. However, the behavior of QAOA in this high-depth limit remains largely unexplored. Heuristic numerical studies indicate that while the optimization landscape of QAOA becomes increasingly complex [Cro18, ZWC$^+$20], a robust initialization strategy can lead to rapid convergence. Unfortunately, most existing strategies for initialization rely on heuristic approaches [SS21a, ZWC$^+$20, JCKK21], lacking a rigorous analytical foundation.

Addressing this gap, the recent work [SMKS23] by present authors and collaborators introduced a *recursive* QAOA initialization strategy based on the concept of transition states. Assuming convergence of QAOA at depth p to a local minimum, Ref. [SMKS23] analytically constructed $2p+1$ transition states for QAOA at depth $p+1$. These transition states, characterized by a single negative curvature direction, ensure a reduction in the cost function value when used as initialization points. Drawing upon this analytical foundation, Ref. [SMKS23] proposed a Greedy strategy for sequential QAOA initializations that systematically improve the cost function value with increasing p. Such a recursive approach is practical even in the limit of large p, providing an analytical basis for the QAOA initialization. While the Greedy strategy comes with guarantees of improvement, it requires computing and diagonalizing the Hessian of the cost function at each transition state, thus increasing the cost of optimization.

In this work, we focus on obtaining analytical insights into deep QAOA using transition states. To this end, we construct an analytical estimate of the minimum Hessian eigenvalue and corresponding eigenvector at each transition state. These results simplify the Greedy initialization strategy [SMKS23] by effectively eliminating the need to construct or estimate the Hessian of the cost function. Furthermore, we provide a physical intuition behind the expression for the minimal Hessian eigenvalue at the transition state and relate it to the energy variance of the state prepared by the QAOA circuit.

The analytical approximation of the Hessian eigenvalue and eigenvector, enables us to expand the QAOA cost function to the fourth order around the transition state. A similar expansion was formulated in Ref. [DBW$^+$19], where it was performed to the third order and applied to the optimal quantum control problem. Our expansion results in a non-trivial local energy minimum located in the vicinity of the transition state, thereby giving a *recursive* lower bound on the cost function improvement achievable through optimization. We check our approximations and bound using numerical simulations of the QAOA on instances of 3-regular unweighted/weighted MaxCut instances with $N = 10$ to 22 vertices, and find the analytic estimates of the Hessian properties to be accurate within a percent. The analytically obtained lower bound on the cost function improvement scales correctly with the number of qubits N. However, our bound decays exponentially with p at a faster rate compared to the numerically obtained cost function improvement.

Our results suggest that although the immediate vicinity of analytic transition states does not contain the true cost function minimum of QAOA at depth $p+1$, it qualitatively captures the "flattening" of the QAOA landscape with p. We speculate that our analytic results on the energy expansion around the transition states may be expanded into an analytic performance guarantee for the QAOA performance at large circuit depths. Specifically, our work establishes

a lower bound on the energy gain expressed via the fidelity of the prepared QAOA state and the true ground state of the cost Hamiltonian. Provided that one manages to bound the increase in fidelity, or potentially, the decrease in the energy variance, this may lead to the desired performance guarantee.

The rest of the paper is structured as follows. In Sec. 5.2, we review the QAOA and the transition states construction introduced in previous work. In Sec. 5.3 we present and verify our estimates for the minimum negative Hessian eigenvalue and the corresponding eigenvector. Furthermore, we discuss numerical results that show a connection between the minimum Hessian eigenvalue and the energy dispersion of the QAOA state. Next, in Sec. 5.4 we present a lower bound on the energy improvement between local minima of the QAOA at circuit depths p and $p+1$. We also test the tightness of the presented bound and discuss its wide implications for the performance of the QAOA. Finally, in Sec. 5.5 we discuss our results and potential future extensions of our work. Appendices D.1-D.3 present detailed proofs of our analytical results, as well as supporting numerical simulations.

5.2 QAOA and transition states

In this section, we start with defining the QAOA algorithm as applied to the MaxCut problem and review the analytic construction of transition states from Ref. [SMKS23].

5.2.1 QAOA and the MaxCut

The QAOA [FGG14] was first introduced as a near-term algorithm for approximately solving classical combinatorial optimization problems. Here, we focus on the particular case of the maximum cut MaxCut problem. MaxCut seeks to partition a given (un)weighted graph $\mathcal{G}$ with $n_{\mathcal{E}}(\mathcal{G})$ edges into two groups such that the number of edges (or the sum of their weights, for weighted problems) that connect vertices from different groups are maximized. Finding the MaxCut for a graph with N vertices is equivalent to finding a ground state for the N-qubit classical Hamiltonian

$$H_C = \sum_{\langle i,j\rangle \in \mathcal{E}} J_{ij}\sigma_i^z\sigma_j^z, \tag{5.1}$$

with the sum running over a set of graph edges $\mathcal{E}$ with weights J_{ij} and σ_i^z being the Pauli-z matrix acting on the i-th qubit. We assume that this problem has a unique ground state, denoted as $|E_0\rangle$ (this state is unique in the proper sector of global Z_2 symmetry, which we use to improve the efficiency of the numerical simulations). The full spectrum of H_C consists of all product states ordered according to their energies and will be used in what follows as a complete basis, $|E_0\rangle, |E_1\rangle, \ldots, |E_{2^N-1}\rangle$.

The depth-p QAOA algorithm [FGG14], denoted in what follows as QAOA$_p$, minimizes the expectation value of the classical Hamiltonian over the variational state $|\Gamma^p\rangle$ where $\Gamma^p = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ encodes variational angles $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ shown in Fig. 5.1(a):

$$|\Gamma^p\rangle = U(\Gamma^p)|+\rangle = \prod_{k=1}^{p} e^{-\beta_k H_B} e^{-\gamma_k H_C} |+\rangle . \tag{5.2}$$

Here

$$H_B = -\sum_{i=1}^{N} \sigma_i^x, \tag{5.3}$$

is the mixing Hamiltonian and the circuit depth p controls the number of applications of the classical and mixing Hamiltonian. The initial product state $|+\rangle = \otimes_{i=1}^{N} |+\rangle_i$, where all qubits point in the x-direction is an equal superposition of all possible graph partitions which is also the ground state of H_B. Finding the minimum of

$$E(\Gamma^p) = \langle \Gamma^p | H_C | \Gamma^p \rangle \tag{5.4}$$

over angles $(\beta_1, \ldots, \beta_p)$ and $(\gamma_1, \ldots, \gamma_p)$ that form a set of $2p$ variational parameters, $\Gamma^p = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, yields a desired approximation to the ground state of H_C, equivalent to an approximate a solution of MaxCut. The scalar function $E(\Gamma^p)$ thus defines a $2p$-dimensional energy landscape where the global minimum yields the best set of QAOA parameters. The performance of the QAOA is typically reported in terms of how close is the approximation ratio to one,

$$1 - r(\Gamma^p) = \frac{E_0 - E(\Gamma^p)}{E_0}, \tag{5.5}$$

where E_0 is the ground state of the classical Hamiltonian (5.1). From here we see that a decrease in $1 - r$ implies that the expectation value of the cost function is approaching the ground state energy of the classical Hamiltonian.

Here, we restrict our attention to MaxCut on 3-regular graphs, where every vertex is connected to exactly 3 other vertices. In the main text, we focus on unweighted 3-regular graphs, while we delegate the results for weighted 3-regular graphs, with weights J_{ij} chosen uniformly at random from the interval $[0, 1]$, for the Appendix D.1. It is important to note that the results presented here are fully general (up to algebraic details), meaning that they hold for generic non-commuting Hamiltonians, H_C and H_B, provided that the initial state $|\psi_0\rangle$ (or $|+\rangle$ in this work) is an eigenstate of H_B.

5.2.2 Analytical transition states

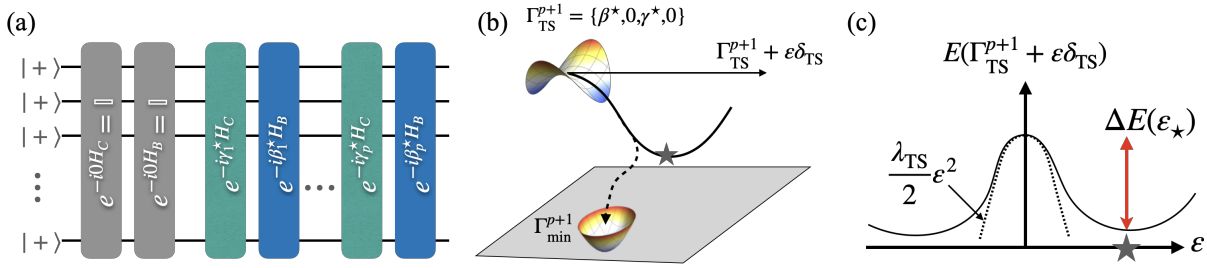

Figure 5.1: (a) Analytic construction of the particular transition state obtained from inserting two identity gates into QAOA$_p$ circuit. (b) We inspect the energy alongside the unique descent direction associated with each of the transition states. The minimum along the unique descent direction (gray star marker) does not correspond to a stationary state of the energy. However, it lower bounds the energy of the minimum obtained by running optimization. (c) Sketch of the projected dependence of the cost function, with $\Delta E(\varepsilon_*)$ putting a rigorous lower bound on the energy improvement at this iteration.

Most studies of the QAOA optimization landscape to date were restricted to local minima of the cost function $E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ since they can be directly obtained using standard gradient-based or some gradient-free optimization routines. Local minima are stationary points of the energy landscape, defined as $\partial_i E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$, with the index i ranging over all $2p$ variational

parameters, where all eigenvalues of the Hessian matrix $H_{ij} = \partial_i \partial_j E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are positive, that is the Hessian at the local minimum is positive-definite. The other stationary points with $0 < k < 2p$ negative Hessian eigenvalues are known as index-k saddle points.

Work [SMKS23] analytically constructed index-1 saddle points dubbed transition states (TS) hereafter, of the QAOA$_{p+1}$ using a given local minimum of the QAOA$_p$, $\Gamma^p_{\min}$. This construction is illustrated in Fig. 5.1(a), and it consists of the insertion of a pair of identity gates, viewed as additional variational parameters initialized at a value equal to zero. Such insertion is allowed at $2p+1$ possible positions, giving rise to $2p+1$ distinct stationary points of the QAOA$_{p+1}$, $\Gamma^{p+1}_{\mathsf{TS}}$ with a *unique* negative eigenvalue of the Hessian. We note that while showing that $\Gamma^{p+1}_{\mathsf{TS}}$ constructed as above are stationary points is relatively straightforward, demonstrating that their Hessian has a single negative eigenvalue is less trivial, with a detailed proof available in Ref. [SMKS23].

Given that, by construction, all the $2p+1$ TS have the same energy as the initial local minimum $\Gamma^p_{\min}$, one can use the direction associated with the negative eigenvalue of the Hessian (hereafter referred to as index-1 direction) to further decrease the energy, see Fig. 5.1(b) for an example. In this way, the TS construction can be used as an initialization scheme that guarantees *improvement* of the QAOA performance with the circuit depth p [SMKS23]. In this work, using the properties of the QAOA energy landscape in the vicinity of the transition states, we quantify the performance improvement of the QAOA by providing a lower bound for the energy improvement after an iteration of the QAOA.

5.3 Curvature of energy landscape near transition state

In this section, we explore the curvature which is the first nontrivial local property of the QAOA energy landscape around the TS constructed out of a local minima $\Gamma^p_{\min}$ of the QAOA$_p$. Using the structure of the Hessian at the TS, we develop an approximation to its unique negative eigenvalue and its corresponding eigenvector. We also uncover connections between the negative curvature of the Hessian at the TS and the excited state population of the prepared QAOA state as a function of the circuit depth p.

5.3.1 Minimum Hessian eigenvalue and eigenvector

In this work, our analysis is specifically tailored to the scenario where additional identity gates are incorporated at the initial layer of the pre-existing QAOA$_p$ circuit, as illustrated in Fig. 5.1(a). This particular scenario corresponds to the transition state denoted as $\Gamma^{p+1}_{\mathsf{TS}}(1,1)$ in Ref. [SMKS23]. To maintain clarity and avoid unnecessary complexity in notation, we refer to this state more simply throughout our discussion when the context permits. We denote the transition state configuration as:

$$\Gamma^{p+1}_{\mathsf{TS}} = (0, \beta^\star_1, \ldots, \beta^\star_p; 0, \gamma^\star_1, \ldots, \gamma^\star_p). \tag{5.6}$$

The primary reason for focusing on this specific transition state is that it significantly simplifies the analytical manipulations required for deriving the worst-case energy improvement achievable through optimizing QAOA$_{p+1}$, starting from a local minimum obtained by QAOA$_p$. Additionally, Appendix D.1 presents numerical analyses that benchmark the effectiveness of this approach against the GREEDY optimization strategy introduced on [SMKS23], which exploits all $2p+1$ transition states derived from $\Gamma^p_{\min}$.

In Appendix D.2, we first establish rigorous lower and upper bounds for the minimum Hessian eigenvalue. However, these bounds do not include an estimate for the corresponding Hessian eigenvector. To obtain this estimate, we construct a similarity transformation that transforms the Hessian at the transition state Γ_{TS}^{p+1} into an almost block-diagonal form. By taking advantage of the Hessian structure, we derive a vector that refines the previously introduced upper bound for the minimum Hessian eigenvalue, which we then use as our estimate

$$\boldsymbol{\delta}_{\text{TS}} = \left(-\frac{1}{2}, \frac{1}{2}, \underbrace{0, \dots}_{p-1 \text{ zeros}} ; -\frac{\text{sign}(b)}{\sqrt{2}}, \underbrace{0, \dots}_{p \text{ zeros}} \right), \tag{5.7}$$

where the parameter b is the second derivative of the cost function $b = \partial_{\gamma_1}\partial_{\beta_1} E(\Gamma_{\text{TS}}^{p+1})$, which can be expressed as a nested commutator of the following three operators:

$$b = \langle + | [H_C, [H_B, U^\dagger(\Gamma_{\min}^p) H_C U(\Gamma_{\min}^p)]] | + \rangle. \tag{5.8}$$

The approximate form of the eigenvector (5.7) shows that when initialized from the former minima of QAOA$_p$, the classical optimization procedure changes values of angles β_1, γ_1 that were initialized at zero initially, as well as the value of $\beta_2 = \beta_1^\star$ initialized at the value set by the local minimum at depth p. All remaining parameters are left intact at the start of gradient descent. The expression for b above can be further simplified relying on the specific expressions for H_C and H_B, using Eq. (5.1) and Eq. (5.3) respectively

$$b = 8 \langle + | H_C U^\dagger(\Gamma_{\min}^p) H_C U(\Gamma_{\min}^p) | + \rangle. \tag{5.9}$$

Finally, we approximate the minimum Hessian eigenvalue λ_{TS} by the expectation value of the Hessian on the approximate eigenvector $\boldsymbol{\delta}_{\text{TS}}$, obtaining that it is proportional to the second derivative $b = \partial_{\gamma_1}\partial_{\beta_1} E(\Gamma_{\text{TS}}^{p+1})$ defined above:

$$\lambda_{\text{TS}} = -\frac{|b|}{\sqrt{2}} = -4\sqrt{2} |\langle + | H_C U^\dagger(\Gamma_{\min}^p) H_C U(\Gamma_{\min}^p) | + \rangle|. \tag{5.10}$$

We refer the discussion of the physical intuition behind this expression to the Sec. 5.3.3, where we show that λ_{TS} vanishes in the case when QAOA unitary circuit rotates the $|+\rangle$ state into an exact eigenstate of H_C. It is critical to note that our estimation of the minimum Hessian eigenvalue is potentially computable on a NISQ device. From the form of the transition state specified by Eq. (5.6), we deduce that the value of b in Eq. (5.9) can be estimated on a NISQ device using additionally $\mathcal{O}(n_{\mathcal{E}}(\mathcal{G}))$ more circuit executions. Here, $n_{\mathcal{E}}(\mathcal{G})$ denotes the number of edges (interaction terms) in the problem graph $\mathcal{G}$ that determines the cost Hamiltonian Eq. (5.1).

Although in this Section we focused on the transition state obtained by padding with zeros the first layer of the QAOA, in the Appendix D.2, we show that a similar approach allows us to obtain estimates of eigenvectors and eigenvalues for all the $2p+1$ TS. For a generic transition state $\Gamma_{\text{TS}}^{p+1}(i,j)$, the approximate Hessian eigenvector has non-zero components corresponding to adjacent gates, that is $\beta_i, \beta_{i+1}, \gamma_j$, and γ_{j-1}. Moreover, the approximate eigenvalue is given by a particular matrix element of the Hessian when the zeros insertion is at the first or last layer, or by a difference of two particular matrix elements in the Hessian matrix for all remaining transition states. In all the cases, our approximation to the eigenvalue is a measurable quantity, which can be estimated analogously to the procedure described after Eq. (5.10).

5.3.2 Quality of the curvature approximation

To assess the accuracy of our estimates for the true minimum Hessian eigenvalue and its corresponding eigenvector, we examine a collection of non-isomorphic random instances of 3-regular unweighted graphs with $N = 10, \ldots, 16$ vertices—18, 34, 55, and 40 instances respectively—executing the QAOA for circuit depths within the range $p \in [1, 30]$. For each local minimum obtained at a given circuit depth p, we construct the $2p + 1$ transition states and compute their exact numerical Hessians. After determining the Hessian spectrum, we calculate the relative error of our minimum eigenvalue estimate and the deviation of the absolute overlap between the approximate and exact Hessian eigenvector from 1.

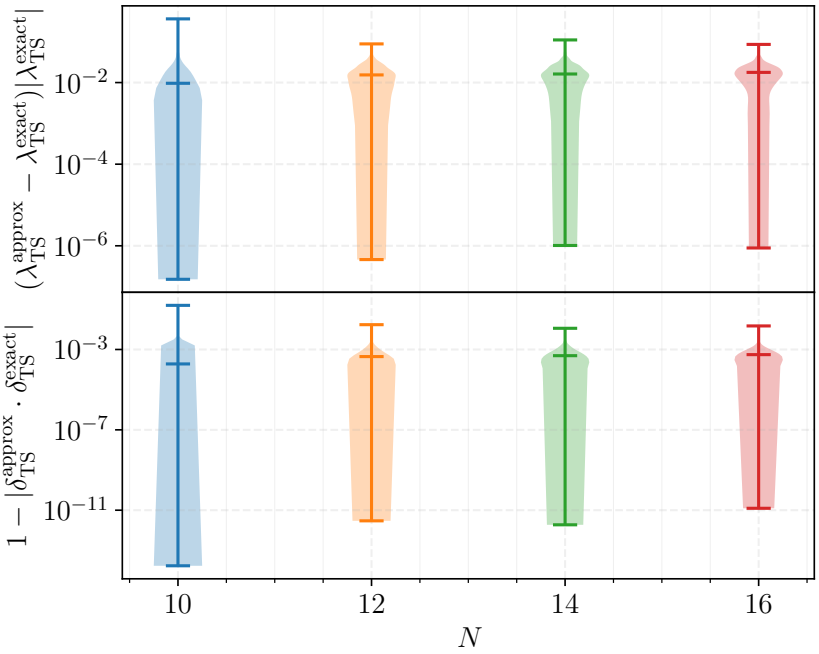

Figure 5.2: Accuracy of curvature and descent direction estimates shown by violin plots for QAOA transition states across graph instances with 10 to 16 vertices and circuit depths ranging from 1 to 30. (*Top*) Relative error in the negative Hessian eigenvalue estimation; the median error is indicated by the horizontal line. (*Bottom*) Deviation from unity in the absolute overlap between the estimated and exact eigenvectors associated with the negative eigenvalue. The shaded regions capture the probability density of the data, reflecting that the accuracy of our eigenvector estimate is consistent across different system sizes.

We consolidate our findings in Fig. 5.2, which is composed of "violin plots" that illustrate the distribution of the obtained numerical data. The width of each violin indicates the frequency of data points at different error or overlap values, providing insight into the variability of the measures. Typically, the median of the data—indicated by the horizontal line within each violin—reveals that the relative error in the minimum Hessian eigenvalue is on the order of $O(10^{-2})$, while the deviation from 1 of the absolute value of the overlap between the exact and approximate eigenstates is on the order of $O(10^{-3})$. More critically, the shape and extent of the violins suggest that the accuracy of our estimates remains consistent across all system sizes and instances examined. This visual analysis underlines the reliability of the estimates provided by our method, with the precision of our estimate appearing stable upon increasing the number of qubits N.

5.3.3 Evolution of the curvature with the depth of QAOA

In this section, we focus on the behavior of the curvature with the circuit depth p. First, we demonstrate that $\lambda_{\rm TS}$ is vanishing when QAOA$_p$ prepares an eigenstate of the cost Hamiltonian H_C, thereby being proportional to the square root of the *infidelity* of the eigenstate preparation. Moreover, we build a physical intuition for the value of curvature by relating it to the action of QAOA circuit on the excited states of the mixing Hamiltonian. Next, we suggest the parallel in the behavior of the curvature and the energy variance of the state prepared in the QAOA circuit with respect to the cost Hamiltonian. Finally, we test our arguments numerically, demonstrating that similarly to the $1-r$, where r is the QAOA approximation ratio, vanishing exponentially with the circuit depth p, the curvature and energy variance also decrease exponentially.

To provide the physical intuition for the value of $\lambda_{\rm TS}$ in Eq. (5.10), we write quantum states $U(\Gamma^p_{\rm min})|+\rangle$ and $U(\Gamma^p_{\rm min})H_C|+\rangle$ in the following form

$$\begin{aligned} U(\Gamma^p_{\rm min})|+\rangle &= \alpha^0|E_0\rangle + \alpha^0_\perp|\psi_0\rangle, \\ U(\Gamma^p_{\rm min})H_C|+\rangle &= n_C\kappa^0|E_0\rangle + n_C\kappa^0_\perp|\phi_0\rangle, \end{aligned} \tag{5.11}$$

where we have selected out the ground state of the classical Hamiltonian, $|E_{l=0}\rangle$ (this can be any eigenstate of H_C, not necessarily the ground state, as we will discuss later), and remaining states $|\psi_0\rangle$ and $|\phi_0\rangle$ in this expansion are normalized superposition of all other eigenstates of H_C, thus being orthogonal to $|E_0\rangle$ by construction[1]. The constant $n_C = \sum_{\langle ij\rangle} J_{ij}^2$ comes from the norm of $H_C|+\rangle$ and it is added such that $|\kappa^0|^2 + |\kappa^0_\perp|^2 = 1$.

Notations introduced in Eq. (5.11) allow us to rewrite the expression for the curvature, Eq. (5.10), as

$$|\lambda_{\rm TS}| = 4\sqrt{2}n_C\alpha^0_\perp\kappa^0_\perp|\langle\phi_0|H_C - E_0|\psi_0\rangle|, \tag{5.12}$$

where without loss of generality we assume that factors in this expression, $\alpha^0_\perp$ and $\kappa^0_\perp$, are real positive numbers. In notations of Eq. (5.11) α^0 corresponds to the square root of the fidelity of the QAOA prepared state $U(\Gamma^p_{\rm min})|+\rangle$ to the ground state $|E_0\rangle$ of H_C with eigenvalue E_0, and $\alpha^0_\perp = \sqrt{1-|\alpha^0|^2}$ is the infidelity. Crucially, the second line of Eq. (5.11) defines κ^0 as square root of the fidelity between states $|E_0\rangle$ and $U(\Gamma^p_{\rm min})H_C|+\rangle$, where the latter state physically corresponds to the QAOA circuit applied to an *excited eigenstate* of the mixing Hamiltonian with eigenvalue $-N+4$ (for more general forms of H_C and H_B, we expect this state to be a combination of low-lying excited eigenstates of H_B). Since the curvature $\lambda_{\rm TS}$ is proportional to the product of $\alpha^0_\perp$ and $\kappa^0_\perp$, it implies that it is sensitive not only to the infidelity resulting from the QAOA circuit preparing the desired ground state of the cost Hamiltonian, but also to the behavior of the low-lying excited state of the mixing Hamiltonian as it is acted upon by the QAOA circuit.

From expression (5.12), we realize that a non-zero curvature around the transition state $\Gamma^{p+1}_{\rm TS}$ comes from the fact that the QAOA circuit prepares a superposition of energy eigenstates as quantified by $\alpha^0_\perp \neq 0$ and $\kappa^0_\perp \neq 0$. Furthermore, to determine the curvature we need information on what superposition of eigenstates of the classical Hamiltonian the QAOA unitary creates when acting on the superposition of states $\sigma^z_i\sigma^z_j|+\rangle$, with $\langle i,j\rangle \in \mathcal{E}_G$ corresponding to an edge in the problem graph G, that are the superposition of second excited states (two spin flips in x basis) of the mixing Hamiltonian.

[1]The terms in Eq. (5.11) are not completely independent. Using the fact that states $H_C|+\rangle$ and $|+\rangle$ are orthogonal, we can show that following relation holds $\alpha^0(\kappa^0)^* + \alpha^0_\perp\kappa^0_\perp\langle\phi_0|\psi_0\rangle = 0$.

Finally, we did not discuss the role of the expectation value, $\langle\phi_0|H_C - E_0|\psi_0\rangle$ in Eq. (5.12). On the one hand, this matrix element is expected to be extensive, i.e. increasing proportionally to number of degrees of freedom, N, as is also confirmed by our numerical simulations, see Figs. 5.3 and D.4. On the other hand, estimating the scaling of this matrix element with p remains an open challenge. The contributions to this expectation value primarily arise from eigenstates of H_C where both $|\phi_0\rangle$ and $|\psi_0\rangle$ have significant weight. Developing a framework to accurately assess these contributions and their scaling with p based on physically motivated assumptions remains an intriguing open problem.

Another physical quantity that quantifies the deviation of the state $|\Gamma^p_{\text{min}}\rangle$ from an eigenstate of the cost Hamiltonian H_C is the energy variance. Using the notation of Eq. (5.11) we express the energy variance as

$$\begin{aligned}\text{var}_{|\Gamma^p_{\text{min}}\rangle}[H_C] &= \langle\Gamma^p_{\text{min}}|H_C^2|\Gamma^p_{\text{min}}\rangle - \langle\Gamma^p_{\text{min}}|H_C|\Gamma^p_{\text{min}}\rangle^2, \\ &= |\alpha^0_\perp|^2\text{var}_{|\psi_0\rangle}[H_C] + |\alpha^0_\perp|^2|\alpha^0|^2(\langle\psi_0|H_C|\psi_0\rangle - E_0)^2. \end{aligned} \tag{5.13}$$

From this expression, it is apparent that the energy variance is proportional to the same infidelity of the state $U(\Gamma^p_{\text{min}})|+\rangle$ to the ground state.

Comparing expressions (5.12)-(5.13), we expect both $|\lambda_{\text{TS}}|$ and energy variance to display similar qualitative behavior with the circuit depth p. Thus, we establish a connection between two seemingly unrelated properties: the energy variance of the quantum state prepared by the QAOA$_p$ circuit and the curvature of the cost function of the QAOA$_{p+1}$ at the transition state. Since so far we focused on the ground state $|E_0\rangle$ of H_C, and previous literature [ZWC$^+$20, Cro18] heuristically demonstrated that for the MaxCut problem on unweighted 3-regular graphs, the approximation ratio $r(\Gamma)$ tends towards 1 exponentially with increasing circuit depth p, we anticipate that infidelity $\alpha^0_\perp$ also tends to zero exponentially, thus leading to an exponential decrease in both curvature and energy variance to zero.

To validate our expectations, we first reproduced the numerically observed exponential convergence of QAOA. We then calculated the average absolute value of λ_{TS} across various unweighted MaxCut instances with N ranging from 12 to 22 vertices. The obtained results are displayed in Fig. 5.3 together with the average of the approximation ratio for circuit depths ranging in the interval $p \in [1, 30]$. The top panel of this figure reproduces the exponential decrease of $1 - r$ with circuit depth p [Cro18, ZWC$^+$20]. The bottom panel shows the surprisingly close quantitative agreement between the averaged energy variance and the absolute value of the negative curvature $|\lambda_{\text{TS}}|$. This signals that these quantities can be related in a tighter way than what we discussed above, in particular, the constants $\alpha^0_\perp$ and $\kappa^0_\perp$ may be proportional to each other.

Finally, we want to highlight some possible implications of the above observations for the performance of the QAOA at finite but deep circuit depth. First, our results above imply that if the QAOA prepares an eigenstate $|E_l\rangle$ that is different from the ground state, $l > 0$, of the cost Hamiltonian H_C the optimization strategy that uses transition states as initialization will halt since there are no descent directions around any of the transition states. Assuming that the ground state is challenging to prepare for whatever reason, it may be feasible for the QAOA to instead prepare a low-lying eigenstate $|E_l\rangle$ with $l > 0$ but not too large, with high fidelity ($|\alpha^l| \sim 1$). As a consequence, the local negative curvature around each TS will become small and as a result, we expect the optimization to slow down. In Appendix D.1, we illustrate this scenario in an instance of a weighted 3-regular graph that was originally studied in [ZWC$^+$20].

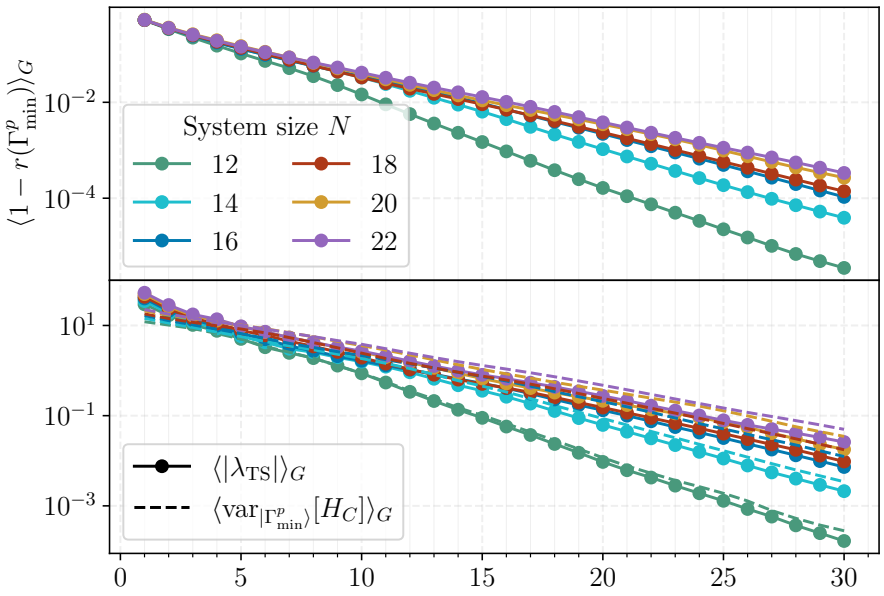

Figure 5.3: (*Top*) Circuit depth dependence of the approximation ratio $r(\Gamma^p_{\text{min}})$, which approach zero exponentially with p. These results were initially observed in [Cro18, ZWC+20]. (*Bottom*) Relation between the magnitude of the negative curvature around the transition state Γ^{p+1}_{TS}, and the energy variance $\text{var}_{\Gamma^p_{\text{min}}}[H_C]$ as functions of the circuit depth p. The numerical data reveals a notable quantitative alignment between the curvature and the energy variance for varying system sizes N.

Thus, our results that relate the landscape curvature at the TS, the energy variance of the QAOA state, and infidelity, suggest that using the Greedy strategy (or similarly using the TS $\Gamma^{p+1}_{\text{TS}}(1,1)$) the QAOA may effectively converge to a (low) energy manifold of the cost Hamiltonian H_C in the regime of deep circuit. Quantifying how this convergence happens remains an open problem, but in the next section we make the first steps in this direction by expanding the cost function to higher order around the transition state.

5.4 Higher order expansion of energy along index-1 direction

In this section we use the approximate index-1 direction given by Eq. (5.7) to expand the QAOA$_{p+1}$ cost function up to the fourth order along the descent direction. This results in lower bound on the improvement of the QAOA cost function resulting from increasing the number of parameters from $2p$ to $2p+2$.

5.4.1 Taylor expansion

Using the explicit knowledge of the index-1 descent direction, we estimate how much the energy can be improved using the Taylor series expansion around the point Γ^{p+1}_{TS}. Specifically, Appendix D.3 details our computation of the cost function's expansion, $E(\Gamma^{p+1}_{\text{TS}} + \varepsilon\boldsymbol{\delta}_{\text{TS}})$, to the *fourth order* in ε. In what follows, we neglect the cubic term in the energy expansion, delegating the details of such step to the Appendix D.3, and obtain a simple expression:

$$E(\Gamma^{p+1}_{\text{TS}} + \varepsilon\boldsymbol{\delta}_{\text{TS}}) \approx E(\Gamma^{p}_{\text{min}}) + \frac{\lambda_{\text{TS}}}{2}\varepsilon^2 + \partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})\varepsilon^4, \tag{5.14}$$

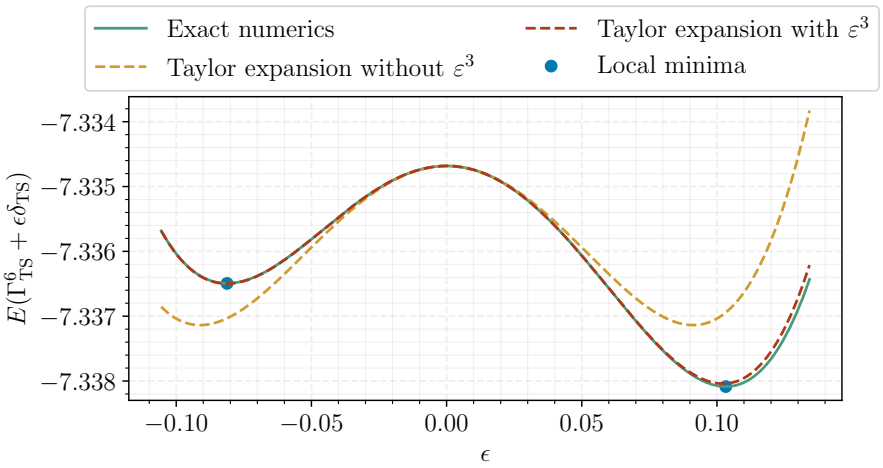

Figure 5.4: Taylor approximation of the energy at a transition state obtained from a local minima of QAOA_5 when perturbed in the index-1 direction. We inspect the impact of the cubic term in the perturbation parameter ε in the energy expansion around the index-1 direction. The instance studied corresponds to that of Appendix D.1.

where $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\mathrm{TS}})$ is expressed as a combination of two expectation values:

$$\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\mathrm{TS}}) = 2\langle +|H_C U(\Gamma^p_{\mathrm{min}})^\dagger H_C U(\Gamma^p_{\mathrm{min}}) H_C|+\rangle - 2\,\mathrm{Re}\Big\{\langle +|U(\Gamma^p_{\mathrm{min}})^\dagger H_C U(\Gamma^p_{\mathrm{min}}) H_C^2|+\rangle\Big\}. \tag{5.15}$$

The fact that the fourth order expansion term of energy is proportional to the second derivative, $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\mathrm{TS}})$ can be understood from the specific form of the descent direction vector, Eq. (5.6). Using the explicit form of the descent vector, in Appendix D.3 we show that the first non-trivial expansion term in ε of the state $U(\Gamma^{p+1}_{\mathrm{TS}} + \varepsilon\boldsymbol{\delta}_{\mathrm{TS}})\,|+\rangle$ is proportional to $\varepsilon^2 H_C\,|+\rangle$. Combining two of such terms, we precisely get the contribution in the first line of Eq. (5.15) above, which is thus coming with an order of ε^4.

In Fig. 5.4, we assess the Taylor approximation's accuracy (which incorporates a cubic term in ε) against exact numerical data obtained by computing the Hessian at $\Gamma^{p+1}_{\mathrm{TS}}$ and determining the energy along the exact index-1 direction. Additionally, we examine the impact of omitting the cubic term from the expansion. While the omission of the cubic term leads to underestimation of the energy improvement, it significantly streamlines the energy expansion analysis and still faithfully replicates the qualitative behavior of exact energy dependence on the slice.

5.4.2 Comparing estimated and true energy gains

Using the expression for the energy Eq. (5.14) along the index-1 direction we compute the value of the perturbation parameter ε that minimizes the energy in this univariate optimization problem. The solution then reads

$$\Delta E(\varepsilon_*) = E(\Gamma^{p+1}_{\mathrm{TS}} + \varepsilon_*\delta_{\mathrm{TS}}) - E(\Gamma^p_{\mathrm{min}}) = -\frac{\lambda^2_{\mathrm{TS}}}{16\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\mathrm{TS}})}, \tag{5.16}$$

with the distance from the transition state to the local minimum on the slice corresponding to $\varepsilon^2_* = -\lambda_{\mathrm{TS}}/4\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\mathrm{TS}})$. We note that although from the expansion of the cost function, we get the local energy minimum, this is the artifact of considering only one out of $2p+2$ directions in the energy landscape. When viewed without projection, we do not expect the

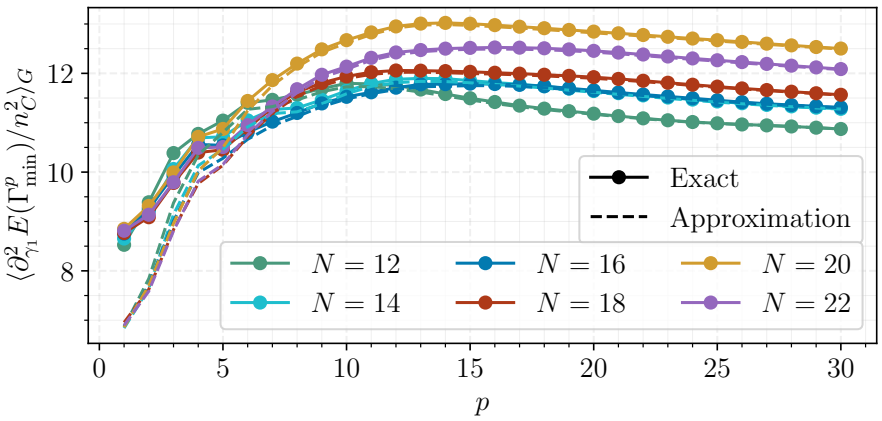

Figure 5.5: Averaged circuit depth behavior of $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$ and its approximation Eq. (5.17) for different system sizes agree for $p \geq 5$.

point $\Gamma^{p+1}_{\text{TS}} + \varepsilon_* \delta_{\text{TS}}$ to be a local minimum or even a saddle point of the cost function, see Fig. 5.1(b).

Using the expression for λ_{TS} obtained in the previous section in Eq. (5.12) we see that the numerator in Eq. (5.16) is proportional to the square root of the infidelity $\alpha^0_\perp$ to the ground state $|E_0\rangle$. Furthermore, we expect that $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$ in the denominator remains finite and extensive in the deep QAOA limit. In particular, in Appendix D.3 we discuss that $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$ can be approximated as follows:

$$\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}}) \approx 2n_C^2 \langle + | \frac{H_C}{n_C} U^\dagger(\Gamma^p_{\min}) H_C U(\Gamma^p_{\min}) \frac{H_C}{n_C} | + \rangle - 2n_C^2 E(\Gamma^p_{\min}). \tag{5.17}$$

where for clarity we explicitly singled out the common factor of n_C that highlights the scaling of the quartic expansion coefficient. From Eq. (5.17) the quartic coefficient in the expansion of energy, $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$, can be understood as the energy difference between states $U(\Gamma^p_{\min}) H_C |+\rangle$ and $U(\Gamma^p_{\min})|+\rangle$, multiplied by the extensive constant $n_C^2 \sim N$. It is natural to expect that QAOA circuit, when applied to the excited eigenstate of the mixing Hamiltonian, $H_C|+\rangle$, yields the final state that has higher energy compared to the state $U(\Gamma^p_{\min})|+\rangle$, that QAOA circuit by design aims to rotate into the ground state of classical Hamiltonian. This physical reasoning implies that the quartic expansion coefficient, $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$, is positive, as is also confirmed in numerical simulations. As the algorithm converges at large enough circuit depths p, we expect $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$ to plateau at an extensive value. Finally, it is important to note that the n_C^2 factor in the denominator of Eq. (5.16) cancels out with the same factor coming from $|\lambda_{\text{TS}}| \propto n_C$, see Eq. (5.12).

We use numerical simulations to verify the validity of the approximation given by Eq. (5.17), on random instances of unweighted 3-regular graphs with $N = [12, 22]$ vertices. From Fig. 5.5, we observe a clear extensive behavior of $\partial^2_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}})$. Interestingly, we see that even though the data for different system sizes does not perfectly collapse onto a single curve, its dependence on the system size at all circuit depths is relatively weak. For example, at $p = 30$ the values for different systems sizes lie in the interval $[11, 13]$.

Using the intuition that the quartic term that represents the denominator in the expression for energy gain, Eq. (5.16) is saturating to the finite value for large p, we conclude that the lower bound on the energy gain is proportional to the curvature around the transition state, and thus is expected to decrease exponentially with the circuit depth p, as supported by the analysis

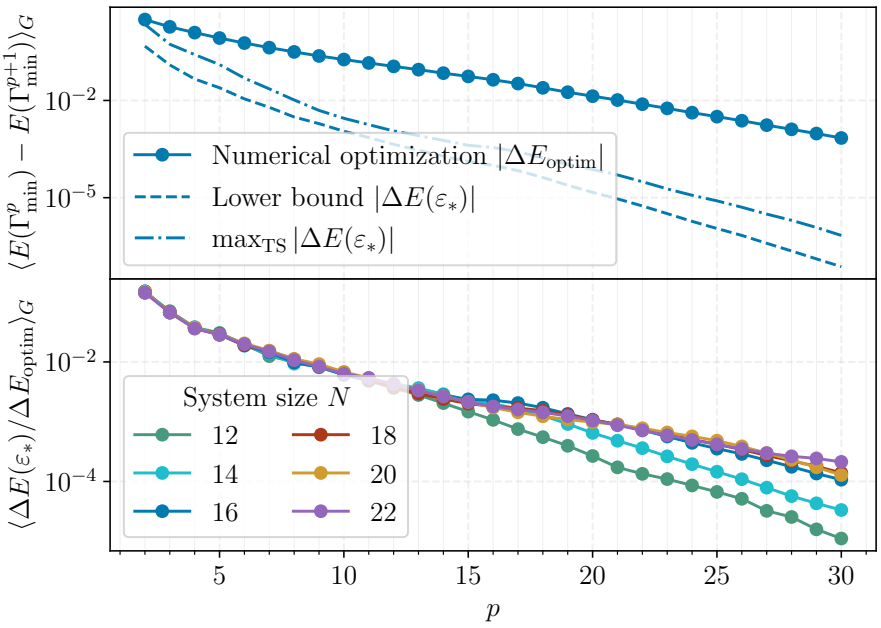

Figure 5.6: (*Top*) Average energy improvement between local minima of QAOA$_p$ and QAOA$_{p+1}$ as a function of the circuit depth p for an unweighted 3-regular graph with $N = 16$ vertices. The lower bound Eq. (5.16), which relies on local information about the cost function landscape around index-1 saddle points overestimates the results obtained by numerically optimizing using the GREEDY strategy of [SMKS23]. (*Bottom*) Averaged quality of the lower bound on the energy improvement, as given by $\Delta E(\varepsilon_*)/\Delta E_{\text{optim}}$, for systems sizes ranging from 12 to 22 vertices.

in the previous section. We numerically check the tightness of the lower bound provided by Eq. (5.16).

To this end, in Fig. 5.6 we first take random instances of unweighted 3-regular graphs with $N = 16$ vertices and compare Eq. (5.16) to the energy improvement coming from performing numerical optimization, using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [Bro70, Fle70, Gol70, Sha70], following the GREEDY strategy introduced in [SMKS23]. We also show the best improvement selected from improvements obtained from moving along the index-1 direction of $2p+1$ distinct TS obtained from the initial local minima $\Gamma^p_{\min}$, labeled as $\max_{\text{TS}}[\Delta E(\varepsilon_*)] = \max_{(i,j)} \left[E(\Gamma^p_{\min}) - E(\Gamma^{p+1}_{\text{TS}}(i,j) + \varepsilon\delta_{\text{TS}})\right]$.

Figure 5.6 reveals that the true energy improvement, and our lower bound Eq. (5.16) both decrease exponentially with p, although with different slopes. In particular, the energy improvement from moving alongside the index-1 direction underestimates the improvement obtained from numerical optimization. This allows us to conclude that although the BFGS optimization algorithm using a large number of iterations can find better local minima by moving far away from the transition state, the entire cost function landscape is getting more "flat" with p. Thus, while our lower bound, which is conceptually similar to one step of local optimization, is underestimating the magnitude of the energy improvement, it has the same functional dependence on p. It remains to be understood if one can establish a (heuristic) relation between our bound and the true energy decrease, thus allowing us to predict the QAOA performance quantitatively from Eq. (5.16).

Finally, we discuss the scaling of energy improvement with system size, as shown in the bottom panel of Fig. 5.6. Similar scaling is also evident in instances of MAXCUT on 3-regular

weighted graphs, as illustrated in Fig. D.5 in Appendix D.1. The numerical results show that at a fixed circuit depth our bound on energy improvement is proportional to the system size N. Indeed, we show that the ratio between the numerical energy improvement and our estimate tends to be a constant value with increasing system size, and numerical energy improvement is known to be proportional to N. Analytically, this behavior arises from the expectation value $|\langle\phi_0|H_C - E_0|\psi_0\rangle|$ in the expression for the approximate negative Hessian eigenvalue in Eq. (5.12) which is extensive in N. In summary, the scaling of our bound on energy improvement with N implies that the improvement in approximation ratio does not scale with N, which is consistent with the gains from numerical optimization.

5.5 Discussion

In this work, we perform an analytic study of transition states of the QAOA cost function, that were constructed in Ref. [SMKS23]. These transition states are characterized by the vanishing gradient of the cost function and a unique negative eigenvalue of Hessian. In the present work, we provide an accurate *analytic* estimate of the minimum eigenvalue of the Hessian and its corresponding eigenvector for each of the $2p + 1$ TS. Moreover, we relate the curvature in the vicinity of transition states to physical observables such as the infidelity of the ground state preparation of the QAOA circuit, and construct the higher-order expansion of the QAOA cost function along the negative curvature direction, which allows us to put a lower bound on the QAOA cost function improvement.

Crucially, the results obtained in this paper are recursive. Assuming that QAOA at depth p found a local minimum, we provide a lower bound on the cost function improvement of the QAOA at depth $p + 1$. Thus, our approach is applicable to QAOA in the regime of large p and we envision that it may be potentially used to obtain a QAOA performance guarantee [FGG14, WL21, FGG20b]. Indeed, the only missing link in such performance guarantee remains to be the bound on the improvement in infidelity, which determines the landscape curvature and lower-bounds energy improvement as shown by our work.

Beyond being a potential step towards performance guarantee, the significance of these estimates is twofold: first, they substantially reduce the computational effort required to implement the Greedy optimization strategy outlined by [SMKS23], as they circumvent the need to construct and diagonalize the Hessian of the cost function at each TS. Second, our results establish the analytical framework that reveals properties of the QAOA cost function at arbitrarily large p.

In particular, we show that for unweighted 3-regular graphs, the negative curvature of the landscape at the TS is intimately connected to the energy variance of the QAOA state. This not only offers a physical interpretation of the negative curvature at the TS but also raises questions about the QAOA performance in the limit of large circuit depth. While it is anticipated that the QAOA will prepare the ground state as $p \to \infty$ [FGG14], our findings suggest that the QAOA may effectively converge to a (low) energy manifold of the cost Hamiltonian H_C in the deep circuit regime.

Furthermore, our numerical analyses indicate that the lower bound on the energy improvement has the same qualitative dependence on the QAOA depth as the true energy improvement. At the same time, our lower bound parametrically underestimates the actual improvements achieved through numerical optimization. This observation suggests that the local vicinity of the transition state that we can explore analytically may be non-trivially related to the global

properties of the QAOA cost function. Establishing such a relation even heuristically may be useful for accurately forecasting the performance of the QAOA, and may provide a useful step towards a more complete understanding of the QAOA.

Finally, given the broad applicability of the transition states-based approach, it becomes intriguing to consider its extension to problems beyond the MaxCut. Exploring the impact of different cost and mixer Hamiltonians on the QAOA cost function landscape presents a promising avenue for future research. Additionally, applying the transition state (TS) strategy to other variational algorithms offers an exciting opportunity. By leveraging the unique characteristics of both the circuit and the problem structure, it may be possible to provide similar initialization strategies and devise analytic estimates for the improvement of the cost function from the optimization process.

APPENDIX A

Appendices to Chapter 2

A.1 Generic formulation of duality

In the main text, we describe the duality procedure for two particular instances of the 3-XORSAT problem. However, it is desirable to formulate the general procedure of deriving the dual Hamiltonian for general (possibly random) instances of classical 3-XORSAT. In this section, we introduce a general description of the duality transformation that uses the language of linear algebra.

A.1.1 Algorithmic description of duality

The matrix A from Eq. (2.1) is the starting point of our procedure. This formulation, can be seen as an extension of the duality mapping used in [FGH$^+$12] for non-invertible A matrices. Since $\{\sigma_i^a\}$ and $\{\tau_j^b\}$ operators, with $a, b \in \{x, y, z\}$ and $i, j \in [1, N]$, belong to different Hilbert spaces, in what follows we will use the symbol " $\equiv$ " to refer to equivalences between then.

Introducing linear algebra notations

In contrast to the particular case of duality in [FGH$^+$12], which required matrix A to be invertible, here we generally deal with the matrix A that is not square and thus is not a full-rank matrix. First, let us denote by r the rank (mod 2) of the matrix A, $\mathrm{rank}_2(A) = r$. We further define matrices S_A and S'_A, which will be used to find τ^x operators. The matrix S_A contains all linearly independent rows of A,

$$S_A = (v_1, \ldots, v_r)^T. \tag{A.1}$$

The matrix S'_A contains the remaining rows of A which by construction can be obtained from those in S_A. Hence, this matrix can be written as a linear superposition of the vectors $v_j \in S_A$,

$$S'_A = F S_A, \tag{A.2}$$

encoded by the $(M - r) \times r$ matrix F.

In order to find the τ^z operators we use a matrix Z

$$Z = (z_i, \ldots, z_r)^T, \tag{A.3}$$

that contains an orthonormal set of vectors z_i, such that $z_j \cdot v_i = \delta_{ij}$. In practice these vectors can be obtained by finding the left-inverse of transposed matrix S_A from Eq. (A.1), $Z \cdot S_A^T = \mathbb{I}_{r\times r}$.

Finally, the conserved charges are associated with the vectors spanning the kernel (mod 2) of A. Since the basis of any linear space is not uniquely defined, we use the following choice of these kernel basis vectors

$$\mathcal{O} = ((S_A^T \cdot Z)_{r+1} + \hat{e}_{r+1}, \dots (S_A^T \cdot Z)_N + \hat{e}_N)^T, \tag{A.4}$$

where $\hat{e}_i$ is the unit vector of length N in the i-th direction. This choice leads to a particularly simple expression for the dual version of quantum terms σ_i^x.

Finding τ^x operators

To construct the τ_α^x operators we use a set of linearly independent rows of A matrix contained in matrix S_A, see Eq. (A.1). Each row of A and S_A contains exactly three entries that are equal to one since we are dealing with the 3-XORSAT problem. Therefore, we identify

$$\tau_\alpha^x \equiv \bigotimes_{l=1}^{N} (\sigma_l^z)^{(S_A)_{\alpha,l}} = \sigma_{i_\alpha}^z \sigma_{j_\alpha}^z \sigma_{k_\alpha}^z, \quad \forall \alpha \in [1, r], \tag{A.5}$$

where $(i_\alpha, j_\alpha, k_\alpha)$ are indices of non-zero entries of row α of matrix S_A. The remaining $M - r$ rows are then expressed as a linear combination of the vectors in S_A as in Eq. (A.2). This implies that a product of σ^z operators encoded by those vectors can be obtained from τ^x operators defined above. Specifically, the product of σ^z's corresponding to a given row $(S'_A)_l$, where $(S'_A)_l = \sum_{k=1}^r F_{l,k}(S_A)_k$, reads:

$$\prod_{k=1}^{r} (\tau_k^x)^{F_{\alpha,k}} \equiv \sigma_{i_\alpha}^z \sigma_{j_\alpha}^z \sigma_{k_\alpha}^z, \tag{A.6}$$

where we imply that $(\tau_k^x)^{F_{\alpha,k}} = \tau_k^x$ if $F_{\alpha,k} = 1$ and $(\tau_k^x)^{F_{\alpha,k}} = 1$ if $F_{\alpha,k} = 0$.

Finally, using equations (A.5)-(A.6) we can express classical Hamiltonian H_C via dual operators as:

$$\tilde{H}_C = \sum_{\alpha=1}^{r} J_\alpha \tau_\alpha^x + \sum_{\alpha=r+1}^{M} J_\alpha \prod_{\beta=1}^{r} (\tau_\beta^x)^{F_{\alpha,\beta}}. \tag{A.7}$$

Finding τ^z operators

Operators τ_β^z can be constructed using matrix Z defined in Eq. (A.3) in a way similar to how operators τ^x were constructed above. Specifically, we set

$$\tau_\alpha^z \equiv \bigotimes_{l=1}^{N} (\sigma_l^x)^{Z_{\alpha,l}}, \quad \forall \alpha \in [1, r]. \tag{A.8}$$

The important difference is that vectors z_α contained in matrix Z may contain a different number of non-zero entries. The commutation and anti-commutation properties of the $\{\tau_\alpha^z, \tau_\beta^x\}$ operators follows directly from the orthogonality properties between z_α and v_β vectors

$$z_\alpha \cdot v_\beta = \delta_{\alpha\beta} \Rightarrow \begin{cases} \{\tau_\alpha^z, \tau_\alpha^x\} = 0, \\ [\tau_\alpha^z, \tau_\beta^x] = 0 \text{ for } \alpha \neq \beta \in [1, r]. \end{cases}$$

To find the dual operator of $H_X = \sum_i \sigma_i^x$ we have to invert Eq. (A.8) and find an expression for σ^x operators via τ^z. This inversion procedure is straightforward for first r spins that correspond to the invertible submatrix of S_A. Thus operators σ_i^x for $i \in [1, r]$ read:

$$\sigma_i^x \equiv \prod_{l=1}^{N} (\tau_l^z)^{(S_A)_{l,i}}, \quad \forall i \in [1, r]. \tag{A.9}$$

To obtain an expression for remaining σ_{r+i}^x with $i \in [1, N-r]$ we use the knowledge of conserved charges from Eq. (A.4) and find that

$$\sigma_{r+i}^x \equiv O_i \prod_{l=1}^{N} (\tau_l^z)^{(S_A)_{l,r+i}}, \tag{A.10}$$

where the particular choice of conserved charges is used as dictated by definition of O matrix in Eq. (A.4):

$$O_l = \prod_i (\sigma_i^x)^{\mathcal{O}_{l,i}}. \tag{A.11}$$

Dual Hamiltonian

Finally, joining Eq. (A.7), (A.9) and (A.10), we obtain the expression for the dual Hamiltonian

$$\tilde{H}_T(s) = -s\Big(\sum_{\alpha=1}^{r} J_\alpha \tau_\alpha^x + \sum_{\alpha=r+1}^{M} J_\alpha \prod_{\beta=1}^{r} (\tau_\beta^x)^{F_{\alpha,\beta}} \Big) - (1-s)\Big(\sum_{i=1}^{r} \prod_{l=1}^{N} (\tau_l^z)^{(S_A)_{l,i}} + \sum_{i=1}^{N-r} O_i \prod_{l=1}^{N} (\tau_l^z)^{(S_A)_{l,r+i}} \Big). \tag{A.12}$$

A.1.2 Example

Let us now illustrate the abstract procedure defined above using a specific example. We start from the matrix A corresponding to an instance of the 2-regular 3-XORSAT model with $N = 6$ and $M = 4$. The example considered here is a particular instance of the closure of the tree hypergraph Fig. 2.4(a) with $g = 1$, corresponding to the following A matrix:

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Similar to the main text we restrict to the case with all couplings $J_\alpha = 1$.

For this particular case, it is easy to check that the rank mod 2 of A is $r = 3$. To see this, for example, we could realize that the first row is the sum (mod 2) of all the other rows. We then pick a submatrix of A containing all linearly independent rows as:

$$S_A = (v_2, v_3, v_4)^T = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Using Eq. (A.5) we then obtain:

$$\tau_1^x \equiv \sigma_1^z \sigma_4^z \sigma_6^z, \quad \tau_2^x \equiv \sigma_2^z \sigma_4^z \sigma_5^z, \quad \tau_3^x \equiv \sigma_3^z \sigma_5^z \sigma_6^z.$$

The F matrix in this case corresponds to a row vector with all r components being equal to one, $F = (1, 1, 1)$. Using Eq. (A.6) we obtain

$$\prod_{i=1}^{3} \tau_i^x = \sigma_1^z \sigma_2^z \sigma_3^z.$$

Hence, we can write the dual form of the classical Hamiltonian H_C which reads

$$\tilde{H}_X = \tau_1^x + \tau_2^x + \tau_2^x + \tau_1^x \tau_2^x \tau_3^x.$$

We now focus on defining the τ_α^z operators. For this particular case, it is easy to check that

$$Z = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Using Eq. (A.8), we get

$$\tau_\alpha^z \equiv \sigma_\alpha^x, \quad \forall \alpha = 1, 2, 3.$$

From the above point, we can already read the expression for the σ_i^x operators in terms of the τ_i^z operators for $i = 1, 2, 3$. Furthermore, to find the expression of the remaining σ_i^x operators ($i = 4, 5, 6$) we need to find the conserved charges of the theory.

Computing the kernel (mod 2) of S_A we obtain:

$$\mathcal{O} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix},$$

which in the spin language from Eq. (A.11) corresponds to

$$O_1 = \sigma_1^x \sigma_2^x \sigma_4^x, \quad O_2 = \sigma_2^x \sigma_3^x \sigma_5^x, \quad O_3 = \sigma_1^x \sigma_3^x \sigma_6^x. \tag{A.13}$$

Thus, it only remains to find the set of clauses in which the spins $i = 4, 5, 6$ participate. From that and using Eq. (A.13), we find the dual expressions for the remaining σ^x operators:

$$\sigma_4^x \equiv O_4 \tau_1^z \tau_2^z, \quad \sigma_5^x \equiv O_5 \tau_2^z \tau_3^z, \quad \sigma_6^x \equiv O_6 \tau_1^z \tau_3^z.$$

The dual Hamiltonian follows directly from all the above results

$$H_T(s) = -s\Big(\sum_{\alpha=1}^{3} \tau_\alpha^x + \tau_1^x \tau_2^x \tau_3^x\Big) - (1-s)\Big(\sum_{\alpha=1}^{3} \tau_\alpha^z + O_4 \tau_1^z \tau_2^z + O_5 \tau_2^z \tau_3^z + O_6 \tau_1^z \tau_3^z\Big). \tag{A.14}$$

A.2 Ising on the closure of the tree hypergraph

In this appendix, we provide details on the procedure that allows removing the non-local term τ_M^x in the dual Hamiltonian (2.14). The approach we present here is inspired by the one carried out in Ref. [SJ16]. We engineer a Hamiltonian $\tilde{K}_T$ with an Abelian Z_2 symmetry, which is equivalent to $\tilde{H}_T$ in a given symmetry sector (which we denote as physical subspace).

In addition, we request that in $\tilde{K}_T$ the non-local term becomes equivalent to a single spin operator. For this, we define the following projector $P_0 = \sum_{v\in\{0,1\}^{M-1}} |\, v_+\rangle\langle v_+\,|$, where

$$|\, v_+\rangle = \frac{1}{\sqrt{2}}\Big(|\, v,0\rangle + |\, v,1\rangle\Big).$$

As a result, it is easy to check the following relations holds:

$$\begin{aligned} &P_0\Big(\tau^x_\alpha \otimes 1\Big)P_0 = X_\alpha,\ \alpha \in [1, M-1], \\ &P_0\Big(\prod_{\alpha=1}^{M-1} \tau^x_\alpha \otimes 1\Big)P_0 = X_M. \end{aligned} \tag{A.15}$$

Eq. (A.15) indicates that when restricted to the physical subspace the action of the X_i operators for $i \in [1, M-1]$ is identical to that of the τ^x_i operators. On the other hand, the non-local term $\prod_{i=1}^{M-1}\tau^x_i$ is now encoded in the X_M operator associated with a new degree of freedom. We then have all the information needed to construct the Hamiltonian $\tilde{K}_T$.

We note that Eq. (A.15) directly implies that $\prod_{i=1}^{M} X_i = 1$, which completely specifies the physical subspace. Furthermore, the form that the remaining terms of $\tilde{H}_T$ take can be obtained from their (anti)commutation relations with the non-local term $\prod_{i=1}^{M-1}\tau^x_i$. More specifically, we note that for operators O_c commuting with the non-local operator the following holds

$$P_0\Big([O_c, \prod_{i=1}^{M-1}\tau^x_i] \otimes 1\Big)P_0 = [P_0\Big(O_c \otimes 1\Big)P_0, X_M] = 0. \tag{A.16}$$

Hence, it implies that $P_0\Big(O_c \otimes 1\Big)P_0$ contains either the X_M operator or acts as the identity on the spin M. However, using the definition of P_0 we see that only the identity on spin M is permitted. In the same spirit, we see that for operators O_{ac} anticommuting with the non-local term it holds that

$$P_0\Big(\{O_{ac}, \prod_{i=1}^{M-1}\tau^x_i\} \otimes 1\Big)P_0 = \{P_0\Big(O_{ac} \otimes 1\Big)P_0, X_M\} = 0, \tag{A.17}$$

which in turns implies that $P_0\Big(O_{ac} \otimes 1\Big)P_0$ has to contain either the Z_M operator or the Y_M operator. Using again the definition of P_0 we see that only the Z_M operator is permitted. In this way, $\tilde{K}_T$ takes a form of transverse-field Ising model on the closed lattice, Fig. 2.4:

$$\tilde{K}_T(s) = -s\sum_{\alpha=1}^{M} J_\alpha X_\alpha - (1-s)\sum_{\langle\alpha,\beta\rangle} Z_\alpha Z_\beta. \tag{A.18}$$

As a consistency check, we note that the subspace specified by the constraint $\prod_{\alpha=1}^{M} X_\alpha = 1$ corresponds to the positive parity sector of $\tilde{K}_T(s)$ Hamiltonian with respect to the Z_2 symmetry implemented by the operator $\prod_\alpha X_\alpha$.

APPENDIX B

Appendices to Chapter 2

B.1 Classical shadows and implementation details

Shadow tomography attempts to directly estimate interesting properties of an unknown state without performing full state tomography as an intermediate step. [Aar17] and [AR19] showcased that such a direct estimation protocol can be exponentially more efficient, both in terms of Hilbert space dimension (2^N in our case) and in the number of target properties (we will use L to denote this cardinality). These techniques do, however, require to store copies of the underlying quantum state in parallel within a quantum memory and performing highly entangled gates on all copies simultaneously. This is too demanding for current and near-term quantum devices.

[HKP20] developed a more near-term friendly variant of this general idea known as prediction with *classical shadows*. Similar ideas have been independently proposed by [PK19] and [MD19], respectively. As explained in detail below, the key idea is to sequentially generate state copies and perform randomly selected single-qubit Pauli measurements. Such measurements can be routinely implemented in current quantum hardware and enable the prediction of many (linear and polynomial) properties of the underlying quantum state. Importantly, the measurement budget (number of required measurements) still scales logarithmically in the number of target properties L, but it may scale exponentially in the support size k of these properties. This is not a problem for local features, like subsystem purities or terms in a quantum many-body Hamiltonian, but does prevent us from directly estimating global state features like fidelity estimation.

The general measurement budget that is required to simultaneously estimate L local observables using classical shadows, necessary for the energy expectation value estimation, is provided in Theorem 4. Typically the estimation of L observables would scale linearly in L (essentially every term is estimated individually). This is traded with a $\ln L$ dependence instead and an exponential dependence on the support k of the operators. The cost for estimating the subsystem purities and thus second Rényi entanglement entropies is provided in Eq. (B.7) and is exponential in k (this dependence was recently proven to be unavoidable [CCHL21]). However since for the WBP check outlined in the main text k is small, this is generally an efficient operation. Lastly, the cost for estimating the gradients is given in Eq. (B.9). The efficiency of using classical shadows to estimate the energy expectation value and gradients is system dependent (see Ref. [HKP20] for the application of classical shadow tomography to the

lattice Schwinger model). For the estimation of the purities, the shadow protocol, however, generally provides the most efficient technique currently available [EKH$^+$20a]. One possibility to circumvent these restrictions is to use a hybrid scheme where the energy and gradients are estimated with either classical shadows or the usual approach dependent on the structure of the Hamiltonian while the second Rényi entropies for the WBP check are always estimated using classical shadows.

B.1.1 Data acquisition via classical shadows

We use randomized single-qubit measurements to extract information about a variational N-qubit state represented by a density matrix

$$\rho(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})| \quad \text{with } \boldsymbol{\theta} \in \mathbb{R}^m.$$

To this end, we repeat the following procedure a total of T times. For $1 \leq t \leq T$ we

1. Prepare quantum state $\rho(\boldsymbol{\theta})$ on the NISQ device.
2. Select N single-qubit Pauli observables independently and uniformly at random.
3. Perform the associated N-qubit Pauli measurement (single-shot) to obtain N classical bits (0 if we measure 'spin down' and 1 if we measure 'spin up').
4. Store N single-qubit 'post-measurement' states, $|s_i^{(t)}\rangle$, where an i-th qubit measurement outcome, s_i, can take six possible values denoted as $|0\rangle$, $|1\rangle$ if qubit was measublue in z-basis, $|+\rangle$ and $|-\rangle$ for x-basis, and, finally, $|+\mathrm{i}\rangle$ and $|-\mathrm{i}\rangle$ for y-basis. Here, $|\pm\rangle = (|0\rangle \pm |1\rangle)/\sqrt{2}$ denote Pauli-x matrix eigenstates and $|\pm\mathrm{i}\rangle = (|0\rangle \pm \mathrm{i}|1\rangle)/\sqrt{2}$ are two Pauli-y eigenstates. In practice, this is achieved by applying random single qubit Clifford gates that effectively implement a change of basis such that the usual z-basis measurement can be used, see Fig. 4.1 (a) for a visualization.
5. (Implicitly) construct the N-qubit *classical shadow*

$$\hat{\rho}^{(t)}(\boldsymbol{\theta}) = \bigotimes_{i=1}^{N} \left(3|s_i^{(t)}\rangle\langle s_i^{(t)}| - \mathbb{I}\right). \tag{B.1}$$

Repeating this procedure a total of T times provides us with T classical shadows $\rho^{(1)}(\boldsymbol{\theta}), \ldots, \rho^{(T)}(\boldsymbol{\theta})$. These are random matrices that are statistically independent (because they are constructed from independent quantum measurements). By construction, each classical shadow reproduces the true underlying state in expectation (over both the choice of Pauli observable and the observed spin direction):

$$\mathbb{E}\left[\hat{\rho}^{(t)}(\boldsymbol{\theta})\right] = \rho(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|, \tag{B.2}$$

see e.g. Ref. [HKP20, Proposition S.2]. We can now approximate this ideal expectation value by empirical averaging over all samples:

$$\rho(\boldsymbol{\theta}) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{\rho}^{(t)}(\boldsymbol{\theta}).$$

This approximation becomes exact in the limit $T \to \infty$ of infinitely many measurement repetitions. But the main results in Refs. [HKP20, PK19] highlight that convergence actually happens much more rapidly.

This is, in particular, true for subsystem density matrices. The tensor product structure of classical shadows (B.1) plays nicely with taking partial traces. Let $A \subseteq \{1, \ldots, N\}$ be a collection of $|A| = k$ qubits. Then,

$$\hat{\rho}_A^{(t)}(\boldsymbol{\theta}) = \mathrm{tr}_{\neg A}\left(\hat{\rho}_A^{(t)}\right) \tag{B.3}$$

is a k-qubit shadow that can be used to approximate the associated subsystem density matrix. More precisely, Eq. (B.2) asserts

$$\mathbb{E}\left[\rho_A^{(t)}(\boldsymbol{\theta})\right] = \mathrm{tr}_{\neg A}\left(\mathbb{E}\left[\hat{\rho}^{(t)}(\boldsymbol{\theta})\right]\right) = \mathrm{tr}_{\neg A}(\rho(\boldsymbol{\theta})) = \rho_A(\boldsymbol{\theta}) \tag{B.4}$$

which can (and should) form the basis of empirical averaging directly for the subsystem in question. Here is a mathematically rigorous result in this direction. In what follows, the range (or weight) of an observable is the number of qubits on which it acts nontrivially. E.g. coupling terms in the Heisenberg Hamiltonian (3.2) have range $k = 2$, while the external field terms have range $k = 1$.

Theorem 4. *Fix a collection of L range-k observables O_l, as well as parameters $\epsilon, \delta > 0$. Then, with probability (at least) $1 - \delta$, classical shadows of size*

$$T \geq \frac{4^{k+1} \ln(2L/\delta)}{\epsilon^2}$$

suffice to jointly estimate all *L expectation values up to additive accuracy ϵ. I.e.*

$$\hat{\rho}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\rho}^{(t)}(\boldsymbol{\theta}) \text{ obeys } \left|\mathrm{tr}\left(O_l \hat{\rho}(\boldsymbol{\theta})\right) - \mathrm{tr}\left(O_l \rho(\boldsymbol{\theta})\right)\right| \leq \epsilon,$$

for all $1 \leq l \leq L$.

We emphasize that it is not necessary to form global shadow approximations. If O_l only acts non-trivially on subsystem $A_l \subseteq \{1, \ldots, N\}$ ($O_l = \tilde{O}_l \otimes \mathbb{I}_{\neg A_l}$), then $\mathrm{tr}\left(O_l \hat{\rho}(\boldsymbol{\theta})\right) = \mathrm{tr}\left(\hat{O}_l \hat{\rho}_{A_l}\right)$. Theorem 4 is slightly stronger than a related result in [**?**] (it does not require median-of-means estimation). Conceptually similar results have been established in Refs. [HKT$^+$21] and [EHF19, HKP21]. Notably, the authors of Ref. [ASS21] pointed out to us that they provided a similar statement as in Theorem 4 in their work. We present a formal proof in Appendix B.1.5 below.

B.1.2 Estimating subsystem purities

Suppose we are interested of estimating a collection of multiple subsystem purities

$$p_A(\boldsymbol{\theta}) = \mathrm{tr}\left(\rho_A(\boldsymbol{\theta})^2\right) = \mathrm{tr}\left(\rho_A(\boldsymbol{\theta})\rho_A(\boldsymbol{\theta})\right), \tag{B.5}$$

where $A \subseteq \{1, \ldots, N\}$ labels different subsystems of size $|A| = k$ each. Then, we can use the corresponding subsystem shadows (B.3) to approximate each p_A by empirical averaging:

$$\hat{p}_A(\boldsymbol{\theta}) = \frac{1}{T(T-1)} \sum_{t \neq t'} \mathrm{tr}\left(\hat{\rho}_A^t \hat{\rho}_A^{t'}\right). \tag{B.6}$$

It is important that we restrict our averaging operation to distinct pairs of classical shadows ($t \neq t'$). This guarantees that the expectation values factorize, i.e.

$$\mathbb{E}\left[\hat{\rho}_A^t \hat{\rho}_A^{t'}\right] = \mathbb{E}\left[\hat{\rho}_A^t\right] \mathbb{E}\left[\hat{\rho}_A^{t'}\right] = \rho_A^2,$$

where the last equality is due to Eq. (B.3). Formula (B.6) is an empirical average over all distinct shadow pairs contained in the data set. It converges to the true average $p_A(\boldsymbol{\theta}) = \mathbb{E}\left[\hat{p}_A(\boldsymbol{\theta})\right]$, and the speed of convergence is governed by the variance. As data size T increases, this variance decays as

$$\mathrm{Var}\left[\hat{p}_A(\boldsymbol{\theta})\right] \leq \frac{2}{T}\left(2 \times 4^k p_2(\boldsymbol{\theta}) + \frac{1}{T-1} 2^{4k}\right),$$

see e.g. Ref. [NCV+21, SM Eq. (12)]. In the large-T limit, this expression is dominated by the first term in parentheses, $4 \times 2^k p_2(\boldsymbol{\theta})/T$, and Chebyshev's inequality allows us to bound the probability of a large approximation error. For $\epsilon > 0$,

$$\Pr\left[\left|\hat{p}_A(\boldsymbol{\theta}) - \mathrm{tr}\left(\rho_A(\boldsymbol{\theta})^2\right)\right| \geq \epsilon\right] \lesssim \frac{4^{k+1}\mathrm{tr}\left(\rho_A^2\right)}{T\epsilon^2},$$

provided that the total number of measurements T is large enough to suppress the higher-order contribution in the variance bound (this is why we write $\lesssim$). In this regime, a measurement budget that scales as

$$T \geq \frac{4^{k+1}\mathrm{tr}\left(\rho_A^2\right)}{\epsilon^2 \delta} \tag{B.7}$$

suppresses the probability of a sizable approximation error ($\geq \epsilon$) below δ. It is worthwhile to point out that this bound depends on the subsystem purity under consideration. Smaller purities are cheaper to estimate than large ones. It is also important to note that the accuracy parameter ϵ has to be small enough in order to accurately capture the purity in the WBP regime, which decays exponentially fast, but only with the subsystem size k.

The δ-dependence in Eq. (B.7) can be further improved to $\ln(1/\delta)$ by replacing simple empirical averaging in Eq. (B.6) by median-of-means estimation [**?**]. Doing so would allow us to estimate all possible $L = \binom{N}{k} \leq N^k$ size-k subsystem purities with only a $k \ln N$-overhead. Median-of-means estimation does, however, worsen the dependence on ϵ by a constant amount. Empirical studies conducted in Ref. [EKH+20b] showcase that such a tradeoff only becomes viable if one wishes to approximate polynomially many subsystem purities.

B.1.3 Estimating gradients

To perform the GD update step suggested in Algorithm 1 we require the knowledge of gradient $\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ which consists of pN derivatives $\partial_{i,l} E(\boldsymbol{\theta})$. The derivative can naively be approximated using finite difference, though for variational single qubit rotation gates, as used in the main text [see Eq. (3.1)], we can use the parameter-shift rule to compute the gradients exactly (up to finite sampling errors) [MNKF18, SBG+19]. The parameter-shift rule is given by

$$\partial_{i,l} E\left(\boldsymbol{\theta}\right) = \frac{1}{2}\left(E\left(\boldsymbol{\theta} + (\pi/2)\boldsymbol{e}_{i,l}\right) - E\left(\boldsymbol{\theta} - (\pi/2)\boldsymbol{e}_{i,l}\right)\right),$$

where i labels the qubits and l cycles through all circuit layers, and $\boldsymbol{e}_{i,l}$ is the unit vector. In order to approximate a single gradient, we need to estimate the difference of two energy expectation values $E(\boldsymbol{\theta}_+) = \langle\psi(\boldsymbol{\theta}_+)|H|\psi(\boldsymbol{\theta}_+)\rangle$ with $\boldsymbol{\theta}_+ = \boldsymbol{\theta} + (\pi/2)\boldsymbol{e}_{i,l}$ and $E(\boldsymbol{\theta}_-) = \langle\psi(\boldsymbol{\theta}_-)|H|\psi(\boldsymbol{\theta}_-)\rangle$ with $\boldsymbol{\theta}_- = \boldsymbol{\theta} - (\pi/2)\boldsymbol{e}_{i,l}$ (we suppress i and l indices in $\boldsymbol{\theta}_\pm$ for the sake of brevity). Typically,

the Hamiltonian itself can be decomposed into a sum of L 'simple' terms: $H = \sum_{l=1}^{L} h_l$, where often L can be proportional to the number of qubits, N. This allows to express the gradient as a linear combination of $2L$ expectation values,

$$\partial_{i,l} E(\boldsymbol{\theta}) = \frac{1}{2} \sum_{l=1}^{L} \left(\langle \psi(\boldsymbol{\theta}_+) | h_l | \psi(\boldsymbol{\theta}_+) \rangle - \langle \psi(\boldsymbol{\theta}_-) | h_l | \psi(\boldsymbol{\theta}_-) \rangle \right), \tag{B.8}$$

each of which can be estimated by performing a collection of single-qubit Pauli measurements. If each term h_l is supported on (at most) k-qubits, then Theorem 4 applies. Performing $T \approx 4^k \ln(L/\delta)/\epsilon^2$ randomized Pauli measurements on state $\rho(\boldsymbol{\theta}_+)$ and $\rho(\boldsymbol{\theta}_-)$ each allows us to ϵ-approximate all $2L$ simple terms in Eq. (B.8).

Unfortunately, approximation errors may accumulate when taking the sum over all $2L$ terms. Suppose that we obtain ϵ-accurate estimators $\hat{E}_l(\boldsymbol{\theta}_\pm)$ of contribution of the local Hamiltonian term to the energy $E_l(\boldsymbol{\theta}_\pm) = \langle \psi(\boldsymbol{\theta}_\pm) | h_l | \psi(\boldsymbol{\theta}_\pm) \rangle$. A triangle inequality over all approximation errors then only produces

$$\begin{aligned}
&\left| \partial_{i,l} E(\boldsymbol{\theta}) - \hat{\partial}_{i,l} E(\boldsymbol{\theta}) \right| \\
=&\frac{1}{2} \left| \sum_{l=1}^{L} \left(\hat{E}_l(\boldsymbol{\theta}_+) - E_l(\boldsymbol{\theta}_+) - \hat{E}_l(\boldsymbol{\theta}_-) + E_l(\boldsymbol{\theta}_-) \right) \right| \\
\leq&\frac{1}{2} \sum_{l=1}^{L} \left| \hat{E}_l(\boldsymbol{\theta}_+) - E_l(\boldsymbol{\theta}_+) \right| + \frac{1}{2} \sum_{l=1}^{L} \left| \hat{E}_l(\boldsymbol{\theta}_-) - E_l(\boldsymbol{\theta}_-) \right| = L\epsilon.
\end{aligned}$$

This upper bound only equals ϵ if we rescale the accuracy of original approximation to ϵ/L. Inserting this rescaled accuracy into Theorem 4 produces an overall measurement cost of

$$T \geq \frac{4^{k+1} L^2 \ln(2L/\delta)}{\epsilon^2}. \tag{B.9}$$

The number L of terms in the Hamiltonian typically scales (at least) linearly in the number of qubits N. This implies that the measurement budget (B.9) requiblue to (conservatively) estimate gradients scales quadratically in the system size and thus is parametrically larger than the (conservative) cost of estimating purities of size-k subsystems (B.7). To obtain the full gradient $\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ the procedure has to be repeated pN times since the parameters-shift rule has to implemented for every variational parameter. It should be noted though, that in principle this can be computed in parallel, provided large enough (quantum) computational resources. For example, different NISQ computers could be used to estimate different gradient components at the same time.

B.1.4 Example of error accumulation in an Ising model

The extra scaling with L^2 in Eq. (B.9) is a consequence of error accumulation. If we use the same measurement data to jointly estimate many Hamiltonian terms, then all these estimators become highly correlated. And the effect of outlier corruption – which occurs naturally in empirical estimation – becomes amplified.

Here, we illustrate this subtlety by means of a simple example. Let $H = -J \sum_{i=1}^{N-1} \sigma_i^z \sigma_{i+1}^z$ be the Ising Hamiltonian on a 1-D chain comprised of N qubits ($L = N - 1$). Let us also assume that N is even. This Hamiltonian is diagonal in the Z-basis $|i_1, \ldots, i_N\rangle = |i_1\rangle \otimes \cdots \otimes |i_N\rangle$

with $i_1, \ldots, i_N \in \{0, 1\}$. So, in order to estimate H, it suffices to perform measurements solely in this basis. Born's rule asserts, that we observe bitstring $\hat{s}_1, \ldots, \hat{s}_N$ with probability

$$\Pr[\hat{s}_1, \ldots, \hat{s}_N] = \langle \hat{s}_1, \ldots, \hat{s}_N | \rho | \hat{s}_1, \ldots, \hat{s}_N \rangle,$$

where ρ denotes the underlying N-qubit state. And, we can use these outcomes to directly estimate the total energy. It is easy to check that

$$\begin{aligned}\hat{E} =& \langle \hat{s}_1, \ldots, \hat{s}_N | H | \hat{s}_1, \ldots, \hat{s}_N \rangle \\ =& -J \sum_{i=1}^{N} \langle \hat{s}_i | \sigma_i^z | \hat{s}_i \rangle \langle \hat{s}_{i+1} | \sigma_{i+1}^z | \hat{s}_{i+1} \rangle\end{aligned}$$

obeys $\mathbb{E}\left[\hat{E}\right] = \mathrm{tr}(H\rho)$, regardless of the quantum state ρ in question. Also, estimating individual terms in this sum is both cheap and easy. Convergence of the sum, however, does depend on the underlying quantum state and the correlations within. To illustrate this, we choose $\lambda \in (0, 1)$ and set

$$\rho(\lambda) = (1 - \lambda)|\psi\rangle\langle\psi| + \lambda|\phi\rangle\langle\phi|,$$

where $|\psi\rangle = |00 \cdots 00\rangle$ is the Ising ground state and $|\phi\rangle = |01 \cdots 01\rangle$ is a Néel state. These states obey $\langle\psi|H|\psi\rangle = -J(N-1)$ (ground state) and $\langle\phi|H|\phi\rangle = +J(N-1)$ (highest excited state), so

$$\mathrm{tr}(H\rho(\lambda)) = -J(n-1)(1-2\lambda).$$

The task is to approximate this expectation value based on computational basis measurements. For each measurement, we either obtain outcome $0 \cdots 0$ (with probability $1 - p$) or outcome $01 \cdots 01$ (with probability p). This dichotomy extends to our estimator

$$\hat{E} = \begin{cases} \langle\psi|H|\psi\rangle = -J(N-1) & \text{with prob. } 1-\lambda, \\ \langle\phi|H|\phi\rangle = +J(N-1) & \text{with prob. } \lambda. \end{cases}$$

and we are effectively faced with estimating the (re-scaled) expectation value of a biased coin. The associated variance of such a coin toss can be easily computed and amounts to

$$\mathrm{Var}\left[\hat{E}\right] = \mathbb{E}\left[\hat{E}^2\right] - \left(\mathbb{E}\left[\hat{E}\right]\right)^2 = 4J^2(N-1)^2\lambda(1-\lambda).$$

Unless $\lambda \neq 0, 1$ (where the variance vanishes), this variance it is proportional to $L^2 = (N-1)^2$ and controls the rate of convergence. Asymptotically, a total number of

$$T \geq \mathrm{Var}\left[\hat{E}\right]/\epsilon^2 = 4J^2L^2\lambda(1-\lambda)/\epsilon^2 = \Omega(L^2/\epsilon^2)$$

independent coin tosses are necessary (and sufficient) to ϵ-approximate the true expectation value $\mathbb{E}\left[\hat{E}\right] = \mathrm{tr}(\rho(\lambda)H)$. This is a consequence of the central limit theorem and showcases that a measurement budget scaling with the number L of Hamiltonian terms is unavoidable in general.

We emphasize that this is a contrived worst-case argument that showcases how correlated measurements can affect the approximation quality of a sum of many simple terms, while each term individually is cheap and easy to evaluate. A generalization to the Heisenberg Hamiltonian consideblue in the main text, see Eq. (3.2), is straightforward.

B.1.5 Proof of Theorem 4

Theorem 4 is a consequence of the following concentration inequality. Let $\|O\|_\infty$ denote the operator/spectral norm of an observable. We will also use $\|\cdot\|_1$ to denote the trace norm.

Theorem 5. *Fix a collection of L range-k observables O_l with $\|O_l\|_\infty \leq 1$, a quantum state ρ and let $\hat{\rho} = \frac{1}{T}\sum_{t=1}^{T}\hat{\rho}^{(t)}$ be a classical shadow estimate thereof. Then, for $\epsilon \in (0,1)$,*

$$\Pr\left[\max_{1\leq l\leq L}\left|\mathrm{tr}\left(O_l\hat{\rho}\right)-\mathrm{tr}\left(O_l\rho\right)\right|\geq\epsilon\right]\leq 2L\exp\left(-\frac{\epsilon^2 T}{4^{k+1}}\right).$$

This large deviation bound is a consequence of another well-known tail bound, see e.g. Ref. [FR13, Theorem 7.30].

Theorem 6 (Bernstein inequality)**.** *Let $X^{(1)},\ldots,X^{(T)}$ be independent, centeblue (i.e. $\mathbb{E}\left[X_t\right]=0$) random variables that obey $|X^{(t)}|\leq R$ almost surely. Then, for $\epsilon>0$*

$$\Pr\left[\left|\frac{1}{T}\sum_{t=1}^{T}X^{(t)}\right|\geq\epsilon\right]\leq 2\exp\left(-\frac{\epsilon^2T^2/2}{\sigma^2+RT\epsilon}\right),$$

where $\sigma^2=\sum_{t=1}^{T}\mathbb{E}\left[\left(X^{(t)}\right)^2\right]$.

Proof of Theorem 5. Fix an observable $O=O_l$ with $1\leq l\leq L$ and define $X^{(t)}=\mathrm{tr}\left(O\hat{\rho}^{(t)}\right)-\mathrm{tr}\left(O\rho\right)$. Then, by construction of classical shadows, each $X^{(t)}$ is an independent random variable that also obeys $\mathbb{E}\left[X^{(t)}\right]=0$, courtesy of Eq. (B.2). Next, let $A\subseteq\{1,\ldots,N\}$ with $|A|=k$ be the subsystem on which the range-k observable acts nontrivially, i.e. $O=O_A\otimes\mathbb{I}_{\neg A}$ and $\|O\|_\infty=\|O_A\|_\infty\leq 1$. Then, Hoelder's inequality ($|\mathrm{tr}\left(O_A\rho_A\right)|\leq\|O_A\|_\infty\|\rho_A\|_1$) asserts

$$\begin{aligned}\left|X^{(t)}\right|&=\left|\mathrm{tr}\left(O_A\hat{\rho}_A^{(t)}\right)-\mathrm{tr}\left(O_A\rho_A\right)\right|\\&\leq\|O_A\|_\infty\left(\|\rho_A\|_1+\left\|\hat{\rho}_A^{(t)}\right\|_1\right)\\&=\|O_A\|_\infty\left(1+\prod_{a\in A}\left\|3|s_a^{(t)}\rangle\langle s_a^{(t)}|-\mathbb{I}\right\|_1\right)\\&\leq\left(1+2^{|A|}\right)=1+2^k=R,\end{aligned}$$

where we have also used $\|\rho_A\|_1=\mathrm{tr}(\rho_A)=1$ and the specific form of subsystem classical shadows (B.3) that factorizes nicely into tensor products. Estimating the variance is more difficult by comparison. However, Ref. [**?** , Proposition S3] asserts

$$\mathbb{E}\left[\left(X^{(t)}\right)^2\right]\leq\|O\|_{\mathrm{shadow}}^2\leq 4^k\|O\|_\infty=4^k.$$

In turn, $\sigma^2\leq T4^k$ and we conclude

$$\begin{aligned}&\Pr\left[\left|\mathrm{tr}\left(O\hat{\rho}\right)-\mathrm{tr}\left(O\rho\right)\right|\geq\epsilon\right]\\=&\Pr\left[\left|\frac{1}{T}\sum_{t=1}^{T}X^{(t)}\right|\geq\epsilon\right]\\\leq&2\exp\left(-\frac{\epsilon^2T^2/2}{T4^k+(1+2^k)T\epsilon}\right)\\\leq&2\exp\left(-\frac{\epsilon^2T}{4^{k+1}}\right),\end{aligned}$$

where the last line is a rough simplification of the exponent. Such a tail bound is valid for any $O = O_l$ and the advertised statement follows from taking a union bound (also known as Boole's inequality) over all possible deviations:

$$\begin{aligned}&\Pr\left[\max_{1\leq l\leq L}\left|\mathrm{tr}\left(O_l\hat{\rho}\right)\right) - \mathrm{tr}\left(O_l\rho\right)\right| \geq \epsilon\right]\\ \leq&\sum_{l=1}^{L}\Pr\left[\left|\mathrm{tr}\left(O_l\hat{\rho}\right)\right) - \mathrm{tr}\left(O_l\rho\right)\right| \geq \epsilon\right]\\ \leq&2L\exp\left(-\frac{\epsilon^2 T}{4^{k+1}}\right).\end{aligned}$$

□

B.2 Unitary t-designs

Here, we briefly review the notion of unitary t-designs. The Haar measure is the unique left/right invariant measure on the unitary group $U(d)$, where d here stands for the dimension of the full Hilbert space, $d = 2^N$. Unitary t-designs are ensembles of unitaries that approximate moments of the Haar measure. More precisely, let $\mathcal{E}$ be an ensemble of unitaries, i.e. a subset of $U(d)$ equipped with a probability measure. For an operator O acting on the t-fold Hilbert space $\mathcal{H}^{\otimes t}$, the t-fold channel with respect to $\mathcal{E}$ is defined as

$$\Phi^t_{\mathcal{E}}(O) = \int_{\mathcal{E}} \mathrm{d}U U^{\otimes t}(O)U^{\dagger\otimes t}. \tag{B.10}$$

Essentially, we are asking when the average of an operator O over the ensemble $\mathcal{E}$ equals an average over the full unitary group. A unitary t-design [DCEL09a, GAE07] is an ensemble $\mathcal{E}$ for which the t-fold channels are equal for all operators O,

$$\Phi^t_{\mathcal{E}}(O) = \Phi^t_{\mathrm{Haar}}(O).$$

Being a t-design means we exactly capture the first t moments of the Haar measure with larger t better approximating the full unitary group. There are known constructions of t-designs for $t = 2$ and $t = 3$ [DCEL09b, CLLW16, KG15, Web15, Zhu17]. For $t = 1$, it is known that any basis for the algebra of operators of $\mathcal{H}$, including the Pauli group, is a 1-design. In practice, one is more interested in when the ensemble of unitaries is close to forming a t-design. With this, given a tolerance $\epsilon_t > 0$ one refers to the ensemble $\mathcal{E}$ as being an approximate t-design if

$$\left\|\Phi^t_{\mathcal{E}} - \Phi^t_{\mathrm{Haar}}\right\|_\diamond \leq \epsilon_t,$$

where $\|\cdot\|_\diamond$ is the diamond norm – a worst-case distance measure that is very popular in quantum information theory, see e.g. [Wat18]. In the quantum machine learning literature the distance between the two t-fold channels is known as the expressibility of the ensemble $\mathcal{E}$ [HSCC21], the smaller the distance the more expressive the ensemble is.

B.3 Entanglement and unitary 2-designs

Random unitary operators have often been used to approximate late-time quantum dynamics. In the crudest approximation, it is assumed that the unitary matrix is directly drawn from the

Haar measure. Although flawed – energy, for instance, is not conserved – this model has led to new insights into black hole physics [Pag93, HP07, SS08] and produced computable models of information spreading and entanglement dynamics [NRVH17, NVH18, HQRY16, vKRPS18].

In what follows, we consider a weaker situation where the random unitary operator is drawn from an ensemble $\mathcal{E}$ forming a 2-design, and focus on the entanglement properties of N-qubits random pure states

$$|\psi\rangle = U|\psi_0\rangle, \tag{B.11}$$

with $U \sim \mathcal{E}$. These results have been previously obtained, see for example [PSW06, ODP07, DOP07] and references therein.

Given a bipartition $(A, \neg A)$ of the system, we begin by studying the distance of the blueuced density matrix ρ_A to the maximally entangled state $\rho_A^\infty = \mathbb{I}_A/d_A$, where d_A is the dimension of the Hilbert space $\mathcal{H}_A$ associated with region A. The full Hilbert space dimension is denoted by $d = 2^N$.

B.3.1 Bounding the expected trace distance

Let us recall the following inequality relating the 1-norm (trace distance) $\|M\|_1 = \operatorname{tr}\sqrt{M^\dagger M}$, and the 2-norm (Frobenius norm) $\|M\|_2 = \sqrt{\operatorname{tr}(M^\dagger M)}$

$$\|M\|_2 \leq \|M\|_1 \leq \sqrt{d}\|M\|_2. \tag{B.12}$$

We are interested in bounding $\mathbb{E}_{\mathcal{E}}\big(\|\rho_A - \mathbb{I}_A/d_A\|_1\big)^2$. To do so we first use Jensen's inequality and afterwards employ the inequality (B.12),

$$\begin{aligned}\mathbb{E}_{\mathcal{E}}\big(\|\rho_A - \mathbb{I}_A/d_A\|_1\big)^2 &\leq \mathbb{E}_{\mathcal{E}}\big(\|\rho_A - \mathbb{I}_A/d_A\|_1^2\big)\\ &\leq d_A \mathbb{E}_{\mathcal{E}}(\|\rho_A - \mathbb{I}_A/d_A\|_2^2).\end{aligned} \tag{B.13}$$

The last term on the right hand side is related to the purity:

$$\mathbb{E}_{\mathcal{E}}(\|\rho_A - \mathbb{I}_A/d_A\|_2^2) = \mathbb{E}_{\mathcal{E}}(\operatorname{tr}\rho_A^2) - 1/d_A. \tag{B.14}$$

As we see, the only non-trivial dependence on U comes from the purity of the blueuced density matrix. Let $\{|I\rangle = |i_A, j_{\neg A}\rangle\}_{i,j}$ be the computational basis for the Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_{\neg A}$ (such that it respects the bipartition).

Let us now proceed with the calculation of the average purity. We first compute the blueuced density matrix ρ_A and write it as a sum over products of matrix elements of the unitary operator U:

$$\begin{aligned}\rho_A &= \sum_{j_{\neg A}}^{d_{\neg A}} \langle j_{\neg A}|\rho|j_{\neg A}\rangle = \sum_{j_{\neg A}}^{d_{\neg A}} \sum_{J,I}^{d} \rho_{I,K}\langle j_{\neg A}|I\rangle\langle K|j_{\neg A}\rangle,\\ &= \sum_{i_A,k_A} \sum_{j_{\neg A}} \rho_{(i_A,j_{\neg A}),(k_A,j_{\neg A})}|i_A\rangle\langle k_A|,\\ &= \sum_{i_A,k_A} \sum_{j_{\neg A}} U_{(i_A,j_{\neg A}),(0,0)} U^*_{(k_A,j_{\neg A}),(0,0)}|i_A\rangle\langle k_A|,\end{aligned}$$

where the last line follows from Eq. (B.11).

Afterwards, it can be easily verified that $\mathrm{tr}(\rho_A^2)$ reads

$$\mathrm{tr}\Big(\rho_A^2\Big) = \sum_{i_A,k_A} \sum_{j_{\neg A},p_{\neg A}} U_{(i_A,j_{\neg A}),(0,0)} U_{(k_A,p_{\neg A}),(0,0)} U^*_{(k_A,j_{\neg A}),(0,0)} U^*_{(i_A,p_{\neg A}),(0,0)}. \tag{B.15}$$

Using the following identities for the first and second moment of the unitary group endowed with the Haar measure

$$\begin{aligned}
&\int_{U(n)} dU_H U_{i,j} U^*_{i_1,j_1} = \delta_{i,i_1}\delta_{j,j_1}/d, \\
&\int_{U(n)} dU_H U_{i,j} U_{l,m} U^*_{i_1,j_1} U^*_{l_1,m_1} = \\
&\frac{1}{d^2-1}(\delta_{i,i_1}\delta_{l,l_1}\delta_{j,j_1}\delta_{m,m_1} + \delta_{i,l_1}\delta_{l,i_1}\delta_{j,j_1}\delta_{m,m_1}) - \\
&\frac{1}{d(d^2-1)}(\delta_{i,i_1}\delta_{l,l_1}\delta_{j,m_1}\delta_{m,j_1} + \delta_{i,l_1}\delta_{l,i_1}\delta_{j,j_1}\delta_{m,m_1}),
\end{aligned} \tag{B.16}$$

we get that the following simple expression for the expected purity

$$\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\,\rho_A^2) = \frac{d_A + d_{\neg A}}{1 + d_A d_{\neg A}}. \tag{B.17}$$

Finally, substituting Eq. (B.17) into Eq. (B.14) we obtain

$$\mathbb{E}_{\mathcal{E}}\Big(\|\rho_A - \mathbb{I}_A/d_A\|_1\Big) \leq \sqrt{\frac{d_A^2 - 1}{d_A d_{\neg A} + 1}} \sim \mathcal{O}(\sqrt{d_A/d_{\neg A}}) \tag{B.18}$$

Note that the above result implies that when the complementary subsystem $\neg A$ is (significantly) larger than A, the expected deviation of ρ_A from the maximally mixed state is exponentially small.

B.3.2 Bounding the expected second Rényi entropy

Let us now explore the average value of the second Rényi entropy which, as mentioned in the main text, can be easily estimated using the classical shadows protocol by **?** .

Computing the exact average value of the second Rényi is a complicated task. Hence, we instead provide a lower and an upper bound for it. On one hand, via Jensen's inequality, we have that

$$-\ln \mathbb{E}_{\mathcal{E}}(\mathrm{tr}\,\rho_A^2) \leq \mathbb{E}_{\mathcal{E}}(S_2(\rho_A)), \tag{B.19}$$

which changes the focus of our attention to the expectation value of the purity of the blueuced density matrix $\mathbb{E}_{\mathcal{E}}(\mathrm{tr}\,\rho_A^2)$. Using the result from the previous subsection Eq. (B.17) and taking the logarithm, we get the following lower bound

$$-\ln \mathbb{E}_{\mathcal{E}}(\mathrm{tr}\,\rho_A^2) = -\ln \frac{d_A + d_{\neg A}}{1 + d_A d_{\neg A}}. \tag{B.20}$$

Taking the large d limit and writing everything in terms of $d_A/d_{\neg A}$ we find

$$-\ln \mathbb{E}_{\mathcal{E}}(\mathrm{tr}\,\rho_A^2) \approx \ln d_A - \frac{d_A}{d_{\neg A}} + \mathcal{O}\left(\frac{d_A^2}{d_{\neg A}^2}\right). \tag{B.21}$$

On the other hand, we have that for any state ρ_A the following inequality holds

$$S_2(\rho_A) \leq S(\rho_A) = -\ln \rho_A \,\mathrm{tr}\,\rho_A,$$

where $S(\rho_A)$ is the von Neumann entropy of ρ_A. Taking averages doesn't change this relation and we conclude $\mathbb{E}_{\mathcal{E}}(S_2(\rho_A)) \leq \mathbb{E}_{\mathcal{E}}(S(\rho_A))$. The expectation value of the von Neumann entropy is upper bounded by the *Page entropy*:

$$S^{\text{Page}}(d_A, d) = \frac{1}{\ln 2}\Big(- \frac{d_A - 1}{2}\frac{d_A}{d} + \sum_{j=d/d_A+1}^{d} \frac{1}{j}\Big). \tag{B.22}$$

(author?) conjectublue that this analytical formula accurately captures the von Neumann entropy of a Haar random state. This conjecture was subsequently proven in Ref. [FK94]. Putting everything together, we obtain

$$-\ln \frac{d_A + d_{\neg A}}{1 + d_A d_{\neg A}} \leq \mathbb{E}_{\mathcal{E}}(S_2(\rho_A)) \leq S^{\text{Page}}(d_A, d). \tag{B.23}$$

Considering now that the number of qubits inside region A is equal to k and assuming that $d_A/d_{\neg A} = 1/2^{N-2k} \ll 1$ we arrive at the expression in Theorem 1, that is

$$k \ln 2 - \frac{1}{2^{N-2k}} \leq \mathbb{E}_{\mathcal{E}}(S_2) \leq k \ln 2 - \frac{1}{2}\frac{1}{2^{N-2k}}. \tag{B.24}$$

We see that whenever the unitary ensemble $\mathcal{E}$ forms a 2-design, the expected value of the second Rényi entropy is close to the Page entropy.

B.4 Entanglement growth and learning rate

Here we detail the derivation of Eq. (3.8). We first upper bound the trace distance via

$$T(\rho_A, \sigma_A) \leq T(|\psi\rangle, |\phi\rangle) = \sqrt{1 - f(|\psi\rangle, |\phi\rangle)}, \tag{B.25}$$

where f stands for the pure state fidelity $f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta} + \boldsymbol{\delta})\rangle) = |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta} + \boldsymbol{\delta})\rangle|^2$. Taylor expanding the pure state fidelity around $\boldsymbol{\theta}$ we get

$$f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta} + \boldsymbol{\delta})\rangle) = 1 - \frac{1}{4}\boldsymbol{\delta}^T \mathcal{F}(\boldsymbol{\theta})\boldsymbol{\delta} + \mathcal{O}(\boldsymbol{\delta}^4), \tag{B.26}$$

where $\mathcal{F}(\boldsymbol{\theta})$ is the quantum Fisher information matrix (QFIM) given by

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = 4\,\text{Re}\{\langle\partial_i\psi|\partial_j\psi\rangle - \langle\partial_i\psi|\psi\rangle\langle\psi|\partial_j\psi\rangle\}. \tag{B.27}$$

Assuming $\boldsymbol{\delta} \ll 1$ we can neglect higher order terms in $\boldsymbol{\delta}$ and so

$$T(\rho_A, \sigma_A) \lesssim \sqrt{\frac{1}{4}\boldsymbol{\delta}^T \mathcal{F}(\boldsymbol{\theta})\boldsymbol{\delta}} = \sqrt{\frac{\eta^2}{4}(\nabla_{\boldsymbol{\theta}} E)^T \mathcal{F}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} E}, \tag{B.28}$$

where in the last equality we plugged in the parameter change under GD (Eq. (3.3)), $\boldsymbol{\delta} = -\eta\nabla_{\boldsymbol{\theta}} E$.

B.5 Algorithm performance for SYK model

In this section we show the numerical results for the VQE applied to the ground state search of the Sachdev-Ye-Kitaev (SYK) model [Kit15]. The SYK model provides a canonical example for a volume-law model where the ground state is nearly maximally entangled [HG19].

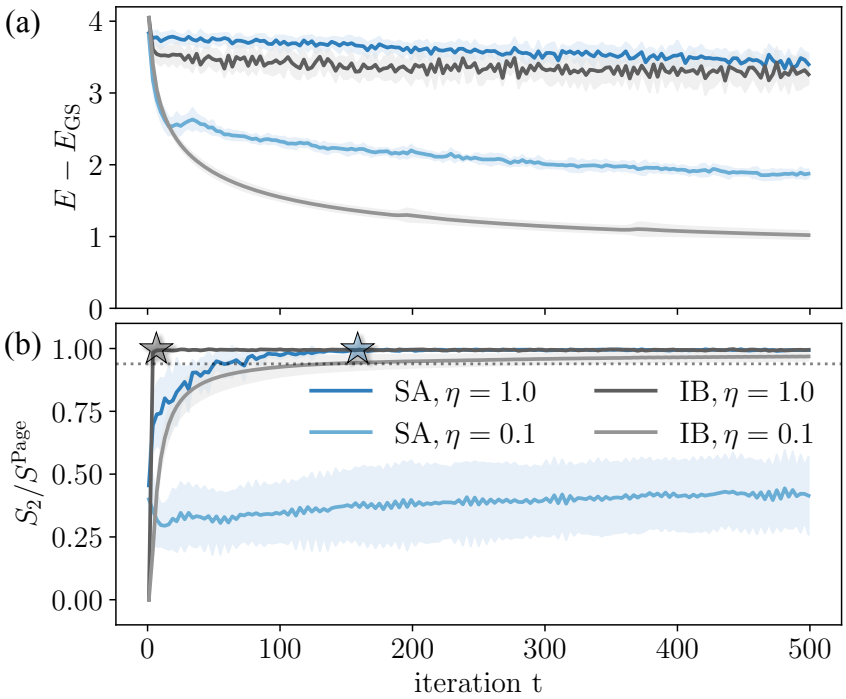

Figure B.1: (a-b) The application of our Algorithm to the problem of finding the ground state of the SYK model. For the initialization we consider the small-angle (SA) ($\epsilon_\theta = 0.1$) and identity block (IB) initialization [GWOB19] (using one block). We can see that only through the reset of the learning rate η, as suggested by Algorithm 1, WBPs are avoided during the optimization. The entanglement entropy of the target state is nearly maximal (indicated by the dotted line), we omit the WBP line for $\alpha = 1$ for improved visibility. We measure energy in units of J and use a system size of $N = 10$, subsystem size $k = 2$ and a random circuit from Eq. (3.1) with circuit depth $p = 100$. Data was averaged over 100 random instances.

The non-local nature of the Hamiltonian does not allow for an efficient estimation of the energy expectation value of this model using classical shadows. Thus, this model may be viewed as a theoretical example that shows that application of our algorithm is not limited to area-law entangled states. We use a small-angle initialization as well as the identity-block initialization [GWOB19] to illustrate our method.

The SYK model is a quantum mechanical model of $2N$ spinless Majorana fermions χ_i satisfying the following anti-commutation relations $\{\chi_i, \chi_j\} = \delta_{ij}$. The SYK model was introduced by Kitaev [Kit15] as a simplified variant of a model introduced by Sachdev and Ye [SY93]. The Hamiltonian of the model is

$$H_{\mathrm{SYK}} = \sum_{i,j,k,l}^{2N} J_{i,j,l}\chi_i\chi_j\chi_k\chi_l, \tag{B.29}$$

where the couplings $J_{i,j,k,l}$ are taken randomly from a Gaussian distribution with zero mean and variance

$$\mathrm{var}[J_{i,j,k,l}] = \frac{3!}{(N-3)(N-2)(N-1)}J^2.$$

We can study Majorana fermions using spin chain variables by a nonlocal change of basis known as the Jordan-Wigner transformation:

$$\chi_{2i} = \frac{1}{\sqrt{2}}\sigma_1^x \cdots \sigma_{i-1}^x \sigma_i^y, \quad \chi_{2i-1} = \frac{1}{\sqrt{2}}\sigma_1^x \cdots \sigma_{i-1}^x \sigma_i^z, \tag{B.30}$$

such that $\{\chi_i, \chi_j\} = \delta_{i,j}$. With this representation, encoding $2N$ Majorana fermions requires N qubits. For our studies, we set $J = 1$ and consider a system of $N = 10$ qubits.

We study performance of VQE for SYK model using two different initializations. Fig. B.1 (a)-(b) shows that the a WBP is avoided during optimization for if the learning rate is chosen appropriately. For a large learning rate ($\eta = 1$) both initializations encounter a WBP during the optimization (indicated by the gray and blue star). Once the learning rate is decreased ($\eta = 0.1$) the entanglement entropy slowly grows to the nearly maximal value associated with the ground state of the SYK model (dotted line) instead of uncontrollably reaching the Page value. For this model, it is important to use $\alpha = 1$ (the default value) such that the entanglement entropy can grow during the optimization. Only if there is some a priori knowledge of the properties of the ground state, α can be chose to be smaller.

The identity block initialization [GWOB19] here leads to the best optimization performance. We attribute this to the fact that the identity block initialization allows for a faster growth in entanglement since the parameter values are highly fine tuned. Our results suggest that sensitivity of the initialization ansatz to small perturbations may be beneficial for the cases when the ground state is nearly maximally entangled. These results highlight the advantage of using our Algorithm. The tracking of the second Rényi entanglement entropy during the optimization reveals that the larger learning rates encounter a WBP while the smaller learning rates successfully avoid it.

APPENDIX C

Appendices to Chapter 3

C.1 Restricting QAOA parameter space by symmetries

In this Appendix, we find the symmetry properties of the cost function

$$E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle \boldsymbol{\beta}, \boldsymbol{\gamma} | H_C | \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle$$

for the QAOA$_p$ (i.e. QAOA with circuit depth p) ansatz. Here we use bold notation for both β and γ parameters to denote a length-p vector of angles, i.e. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. The use of symmetries allows to restrict the manifold of variational parameters, leading to a more efficient exploration of the QAOA landscape. This section expands upon previous results by [ZWC+20].

We begin by rewriting the exponents of both classical and mixing Hamiltonian as:

$$e^{-\mathrm{i}\beta_l H_B} = \prod_{k=1}^{n} e^{-\mathrm{i}\beta_l \sigma_k^x} = (\cos\beta_l - \mathrm{i}\sin\beta_l\, \sigma^x)^{\otimes n}, \tag{C.1}$$

$$e^{-\mathrm{i}\gamma_l H_C} = \prod_{\langle j,k \rangle} e^{-\mathrm{i}\gamma_l \sigma_j^z \sigma_k^z} = \prod_{\langle j,k \rangle} (\cos\gamma_l - \mathrm{i}\sin\gamma_l \sigma_j^z \sigma_k^z). \tag{C.2}$$

From here it is apparent that adding π to any of the parameters, $\beta_l, \gamma_l \to \beta_l + \pi, \gamma_l + \pi$ for all $l \in [1, p]$ does not change the cost function value $E(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Indeed, this leads to an appearance of an overall negative sign that cancels within the expectation value of the classical Hamiltonian. Therefore we can easily restrict the search space to (i) $\beta_l, \gamma_l \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

For β parameters we can restrict the parameter space even further. In Ref. [ZWC+20] the authors restrict the parameters as $\beta_l \in [-\frac{\pi}{4}, \frac{\pi}{4}]$ due to the following considerations. Consider adding $\frac{\pi}{2}$ to β, the exponent $e^{-\mathrm{i}\left(\beta_l + \frac{\pi}{2}\right) H_B} = e^{-\mathrm{i}\beta_l H_B} e^{-\mathrm{i}\frac{\pi}{2} H_B}$ leads to an additional product of all σ^x operators,

$$e^{-\mathrm{i}\frac{\pi}{2} H_B} = (-\mathrm{i}\sigma^x)^{\otimes n}. \tag{C.3}$$

this operator flips all spins, effectively being a generator of the Z_2 symmetry of the classical Ising Hamiltonian, H_C. Therefore, such a shift of β_l will have no effect on the cost function and we restrict (ii) $\beta_l \in [-\frac{\pi}{4}, \frac{\pi}{4}]$.

Yet another symmetry is recovered by taking the complex conjugate of the energy. As both classical and mixing Hamiltonians are real-valued, one has

$$E^*(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle \boldsymbol{\beta}, \boldsymbol{\gamma} | H_C | \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle^* = E(-\boldsymbol{\beta}, -\boldsymbol{\gamma}). \tag{C.4}$$

And because the energy is also real-valued (H_C is Hermitian), we recover another symmetry of the cost function: (iii) $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \to (-\boldsymbol{\beta}, -\boldsymbol{\gamma})$.

The symmetries (i)-(iii) introduced above were discussed in Refs. [ZWC+20, SS21a]. But we can restrict the search space even further. In particular, we demonstrate that for the QAOA cost function for 3-regular random graphs (RRG3) the following *additional* symmetry holds:

(iv) Flipping sign of any of the $\beta_l \to -\beta_l$ for any $l \in [1, p]$ together with shifts of $\gamma_{l,l+1}$ angles, as $\gamma_{l,l+1} \to \gamma_{l,l+1} \pm \frac{\pi}{2}$. Note that for $l = p$ only the γ_p angle has to be shifted.

Let us prove this property for regular graphs with odd connectivity (i.e. 3-regular, 5-regular, ...). In order to demonstrate the property (iv) for $j < p$, it is enough to show that:

$$e^{-\mathrm{i}\frac{\pi}{2}H_C} e^{\mathrm{i}\beta H_B} e^{-\mathrm{i}\frac{\pi}{2}H_C} \sim e^{-\mathrm{i}\beta H_B}, \tag{C.5}$$

where $\sim$ stands for equivalence up to a global phase. In other words, we use the property that $e^{-\mathrm{i}\frac{\pi}{2}H_C} \sim \prod_i \sigma_i^z$ acts as a product of σ^z operators over all spins, that relies on the fact that each vertex is connected to an odd number of edges (interaction terms). This leads to the relation

$$e^{-\mathrm{i}\frac{\pi}{2}H_C} e^{\mathrm{i}\beta H_B} e^{-\mathrm{i}\frac{\pi}{2}H_C} \sim e^{-\mathrm{i}\beta H_B}. \tag{C.6}$$

Thus, the change of sign of β_k can be compensated by the shifts of "adjacent" angles $\gamma_{k,k+1}$ by $\pi/2$, leading to the property (iv) when $j < p$. In the particular case of $j = p$, the property (iv) for $j = p$ is obtained using the following relation

$$e^{\mathrm{i}\frac{\pi}{2}H_C} e^{-\mathrm{i}\beta H_B} H_C e^{\mathrm{i}\beta H_B} e^{-\mathrm{i}\frac{\pi}{2}H_C} \tag{C.7}$$

$$\sim e^{\mathrm{i}\frac{\pi}{2}H_C} e^{-\mathrm{i}\beta H_B} e^{\mathrm{i}\frac{\pi}{2}H_C} H_C e^{-\mathrm{i}\frac{\pi}{2}H_C} e^{\mathrm{i}\beta H_B} e^{-\mathrm{i}\frac{\pi}{2}H_C} \tag{C.8}$$

$$= e^{\mathrm{i}\beta H_B} H_C e^{-\mathrm{i}\beta H_B}. \tag{C.9}$$

Finally, let us rewrite the property (iv) by sequentially applying this symmetry for all indices j starting from k and ending at p. Then we obtain the following property equivalent to (iv) and dubbed (iv'):

(iv') $\forall j = [k, p] : \ \beta_j \to -\beta_j, \ \gamma_j \to \gamma_j \pm \frac{\pi}{2}$.

This allows us to restrict all γ angles to the region $[-\frac{\pi}{4}, \frac{\pi}{4}]$. Moreover, the sign-flip symmetry (iii) allows us to make one of the γ angles, for instance, γ_1, positive, cutting the search space in half.

In addition, let us apply property (iv') for $k = 1$ (i.e. including all layers of the unitary circuit) and supplement it with a global sign flip, operation (iii). As a result, we obtain the following symmetry:

$$\gamma_1 \to \pm\frac{\pi}{2} - \gamma_1, \ \forall j = [2, p] : \ \gamma_j \to -\gamma_j \tag{C.10}$$

This indicates that there is a p-dimensional plane in the landscape with coordinates $\boldsymbol{\gamma} = (\pm\frac{\pi}{4}, \mathbf{0}_{p-1})$ which acts as a mirror. This plane is characterized by a vanishing gradient of the cost function and the Hessian having p vanishing eigenvalues. However, it is located on the edge of our search space and it has a vanishing expectation value of the cost function, corresponding to the approximation ratio $r = 0$, which is very far from the good-quality local minima.

In summary, collecting all symmetries discussed above, we restrict the fundamental search region to

$$\beta_l \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right], \ \forall l \in [1,p], \tag{C.11}$$

$$0 < \gamma_1 < \frac{\pi}{4}, \tag{C.12}$$

$$\gamma_j \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right], \ \forall j \in [2,p]. \tag{C.13}$$

C.2 Construction of transition states

In this section, we show how to use a local minimum of the QAOA$_p$ to construct a set of $2p+1$ transition states (TS) at circuit depth $p+1$. These are stationary points with all but one Hessian eigenvalue being positive. More precisely, we show the following statement:

Theorem 7 (TS construction, full version)**.** *Let $\mathbf{\Gamma}^p_{\min} = (\boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star) = (\beta^\star_1, \ldots, \beta^\star_p, \gamma^\star_1, \ldots, \gamma^\star_p)$ be a local minimum of QAOA$_p$. Define the following $2p+1$ points by padding this vector with zeroes at distinguished positions:*

$$\begin{aligned}\mathbf{\Gamma}^{p+1}_{TS}(i,j) = (&\beta^\star_1, ..., \beta^\star_{j-1}, 0, \beta^\star_j, ..., \beta^\star_p, \\ &\gamma^\star_1, ..., \gamma^\star_{i-1}, 0, \gamma^\star_i, ..., \gamma^\star_p)\end{aligned} \tag{C.14}$$

with $i \in [1, p+1]$ and $j = i$ or $j = i+1$. Then each of these points is either (i) a TS for QAOA$_{p+1}$ or (ii) has a non-regular Hessian.

Theorem 3 in the main text is a streamlined version of this statement that does not mention the possibility of degenerate Hessians. We expect that the Hessian matrix of a local minimum of QAOA$_p$ is non-degenerate in the absence of symmetries and provided the circuit is not overparametrized [LJGM$^+$21] (if there exists some combination of variational angles, such that its changes do not influence the quantum state, it leads to vanishing eigenvalue of Hessian). Analogously, in the case of the Hessian at the TS of QAOA$_{p+1}$, we numerically find that option (ii) never happens. Below, we relate the two new additional eigenvalues of the Hessian at the TS to the expectation value of a physical operator over the variational state. This expectation value is non-zero in the absence of special symmetries or fine-tuning, providing a physical justification for why we do not observe zero eigenvalues in the Hessian spectra of our TS.

C.2.1 Cost function gradient

Let us start by computing the energy gradient $\nabla E(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Derivatives of the quantum state with respect to parameters β_l, γ_l are given by the following expressions:

$$\begin{aligned}\partial_{\beta_l}|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle &= -\mathrm{i} U_{>l} H_B U_{\leq l}|+\rangle, \\ \partial_{\gamma_l}|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle &= -\mathrm{i} U_{\geq l} H_C U_{<l}|+\rangle,\end{aligned} \tag{C.15}$$

where $U_{\geq l} = U_B(\beta_p)U_C(\gamma_p)\cdots U_B(\beta_l)U_C(\gamma_l)$, $U_{\leq l} = U_B(\beta_l)U_C(\gamma_l)\cdots U_B(\beta_1)U_C(\gamma_1)$ and analogously for $U_{<l}$, and $U_{>l}$. For simplified notation we use write $|+\rangle$ instead of $|+\rangle^{\otimes n}$. We can now deduce the components of the energy gradient $\nabla E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ from Eq. (C.15). They read

$$\begin{aligned}\partial_{\beta_l} E(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \mathrm{i}\langle+|U^\dagger_{\leq l}[H_B, U^\dagger_{>l} H_C U_{>l}]U_{\leq l}|+\rangle, \\ \partial_{\gamma_l} E(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \mathrm{i}\langle+|U^\dagger_{<l}[H_C, U^\dagger_{\geq l} H_C U_{\geq l}]U_{<l}|+\rangle.\end{aligned} \tag{C.16}$$

Our goal is to prove that given a local minimum $\mathbf{\Gamma}^p_{\text{min}} = (\beta^\star_1, \ldots, \beta^\star_p, \gamma^\star_1, \ldots, \gamma^\star_p)$ for a QAOA$_p$ the set of $2p+1$ points

$$\begin{aligned}\mathbf{\Gamma}^{p+1}_{\text{TS}}(l,k) = (&\beta^\star_1, ..., \beta^\star_{l-1}, 0, \beta^\star_l, ..., \beta^\star_p,\\ &\gamma^\star_1, ..., \gamma^\star_{k-1}, 0, \gamma^\star_k, ..., \gamma^\star_p),\end{aligned} \tag{C.17}$$

with l ranging from 1 to $p+1$ and either $k=l$ or $k=l+1$ are all TSs. The first step is to prove that they are all stationary points. That is, each such point leads to a vanishing gradient. From the above expression, it follows that we only have to consider gradient components where the zero insertion is made since the others are zero due to the point $\mathbf{\Gamma}^p_{\text{min}}$ being a local minimum (i.e. derivatives are vanishing). For the derivatives over newly introduced angles using Eq. (C.15), we see that

$$\begin{aligned}\partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(l,l)} &= \partial_{\beta_{l-1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}},\\ \partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(l,l+1)} &= \partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}},\\ \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(l,l)} &= \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}},\\ \partial_{\gamma_{l+1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(l,l+1)} &= \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}},\end{aligned} \tag{C.18}$$

where the index l ranges from 1 to $p+1$ for the (l,l) case and from 1 to p in the $(l,l+1)$ case.

These observations reduce the derivatives over the new angles to derivatives over angles from local minima of QAOA$_p$. And these vanish by definition because we started in a local minimum which is itself a stationary point, that is

$$\nabla E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} = 0. \tag{C.19}$$

We emphasize that these arguments do not apply to two special cases that should be treated separately.

In particular, Eq. (C.15) does not provide any information for: (i) the gradient component $\partial_{\beta_1}[\cdot]$ when considering TS $\mathbf{\Gamma}^{p+1}_{\text{TS}}(1,1)$ and $\mathbf{\Gamma}^{p+1}_{\text{TS}}(1,2)$, and (ii) the gradient component $\partial_{\gamma_{p+1}}[\cdot]$ when considering points $\mathbf{\Gamma}^{p+1}_{\text{TS}}(p+1,p+1)$. For case (i), we use that $H_B|+\rangle = n|+\rangle$ with n being the number of qubits, to show that

$$\partial_{\beta_1}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(1,k)} = -\,\mathrm{i}n|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} \tag{C.20}$$

for $k=1,2$. This in turn implies

$$\partial_{\beta_1}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{1}_{\text{TS}}(1,k)} = (\mathrm{i}n - \mathrm{i}n)\langle\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} = 0, \tag{C.21}$$

as desired. For case (ii) we have that

$$\partial_{\gamma_{p+1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(p+1,p+1)} = -\mathrm{i}H_C|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\text{min}}}, \tag{C.22}$$

which handles the second special case:

$$\partial_{\gamma_{p+1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}(p+1,p+1)} = (\mathrm{i}-\mathrm{i})E(\mathbf{\Gamma}^{\boldsymbol{p}}_{\text{min}}) = 0. \tag{C.23}$$

Putting everything together implies that all energy partial derivatives vanish for every $\mathbf{\Gamma}_{\text{TS}}^{p+1}$ introduced in Theorem 3:

$$\nabla E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,l)} = \nabla E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,l+1)} = 0 \tag{C.24}$$

for all $l \in [1, p+1]$ except the pair $(p+1, p+2)$ which exceeds the index range. In other words: these $2(p+1)-1 = 2p+1$ points must all be stationary points.

C.2.2 Cost function Hessian

We now proceed with the study of the Hessian for each of the stationary states in the set $\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,k)$ with l ranging from 1 to $p+1$ and k being l or $l+1$. Using basic row and column operations we decompose the Hessian as follows:

$$H[\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,k)] = \begin{pmatrix} H(\mathbf{\Gamma}_{\text{min}}^{p}) & v(l,k) \\ v^T(l,k) & h(l,k) \end{pmatrix}, \tag{C.25}$$

where $H(\mathbf{\Gamma}_{\text{min}}^{p}) \in \mathbb{R}^{2p\times 2p}$, $v(l,k) \in \mathbb{R}^{2p\times 2}$ and, $h(l,k) \in \mathbb{R}^{2\times 2}$. It is important to note that the determinant of the Hessian at the point $\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,k)$ remains unchanged by such reordering of rows and columns. To see this, recall that switching two rows or columns causes the determinant to switch signs. Since we switch x rows and x columns, we realize that the overall sign does not change after all. In terms of matrix elements, $v(l,k) \in \mathbb{R}^{2p\times 2}$ reads

$$v(l,k) = \begin{pmatrix} \partial_{\beta_1}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\beta_1}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\beta_{l-1}}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\beta_{l-1}}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \partial_{\beta_{l+1}}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\beta_{l+1}}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\beta_{p+1}}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\beta_{p+1}}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \partial_{\gamma_1}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\gamma_1}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\gamma_{k-1}}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\gamma_{k-1}}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \partial_{\gamma_{k+1}}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\gamma_{k+1}}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\gamma_{p+1}}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\gamma_{p+1}}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \end{pmatrix},$$

while $h(l,k) \in \mathbb{R}^{2\times 2}$ becomes

$$h(l,k) = \begin{pmatrix} \partial_{\beta_l}\partial_{\beta_l} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\beta_l}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \\ \partial_{\beta_l}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} & \partial_{\gamma_k}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} \end{pmatrix}.$$

Our goal is to restrict the properties of the Hessian (C.25) using the fact that the Hessian at circuit depth p is a positive-definite matrix, a consequence of the fact that we start at a local minimum $\mathbf{\Gamma}_{\text{min}}^{p}$. To this end, we use a powerful theorem from matrix analysis.

Theorem 8 (Eigenvalue interlacing theorem [Bel97] (Theorem 4 on page 117)). *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and $B \in \mathbb{R}^{m\times m}$ with $m < n$ be a principal submatrix (obtained by removing both the i-th column and i-th row for some values of i). Suppose A has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and B has eigenvalues $\kappa_1 \leq \cdots \leq \kappa_m$. Then*

$$\lambda_k \leq \kappa_k \leq \lambda_{k+n-m}, \tag{C.26}$$

for $k = 1, m$.

The eigenvalue interlacing theorem relates the ordered set of Hessian eigenvalues $\{\lambda_i^{p+1}\}$ for QAOA$_{p+1}$ to the Hessian eigenvalues $\{\lambda_i^p\}$ of QAOA$_p$ in the following way:

$$\lambda_k^{p+1} \leq \lambda_k^p \leq \lambda_{k+2}^{p+1}. \tag{C.27}$$

Using the fact that $H_p(\mathbf{\Gamma}_{\text{min}}^p)$ has $\lambda_k^p > 0$ for all k, we see that the Hessian of QAOA$_{p+1}$ at point $\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,k)$ has at most two negative eigenvalues, $\lambda_1^{p+1}, \lambda_2^{p+1} < \lambda_1^p$, whereas $0 < \lambda_1^p < \lambda_j^{p+1}$ for $j \geq 3$. In what follows we establish that among these two eigenvalues, exactly one is negative and the other one is positive. This is achieved by demonstrating that the full Hessian matrix has a negative determinant,

$$det\, H\Big[\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,k)\Big] < 0, \tag{C.28}$$

which rules out the possibility that the remaining eigenvalues $\lambda_{1,2}^{p+1}$ have the same sign (which would cancel in the determinant).

Below we first prove Relation (C.28) for the cases where the insertion of the zeros is made at the first (i) or at the last (ii) layer of the unitary circuit. We then conclude by considering the general case (iii), where zeros are inserted in the "bulk" of the unitary circuit. Moreover, whenever is clear from context, we will drop the indices (l,k) for better readability. Furthermore, for all the cases considered below, we introduce a specific short-hand notation for the following second-order derivative

$$b = \partial_{\beta_l}\partial_{\gamma_k} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}}. \tag{C.29}$$

This matrix element will play a special role in the calculation of det$H(\mathbf{\Gamma}_{\text{TS}}^{p+1}(l,k))$. It is important to note, that while the specific expression of b differs for all the stationary points in the set given by Eq. (C.17), it has a non-zero value, $b \neq 0$. Indeed, below we express b as an expectation value of a non-vanishing operator over the QAOA variational state, that is non-zero in the absence of special symmetries.

Case (i): $l = k = p+1$

The first step is to compute the matrix elements of $v(p+1,p+1)$. From now on we drop the quantifying index and simply write v and h to reduce notational overhead. The first column of v corresponds to $v_{\beta_j,\beta_{p+1}} = \partial_{\beta_j}\partial_{\beta_{p+1}} E(\boldsymbol{\beta},\boldsymbol{\gamma})$ evaluated at the TS $\mathbf{\Gamma}_{\text{TS}}^{p+1}$:

$$\begin{aligned} &\partial_{\beta_j}\partial_{\beta_{p+1}} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} = \\ &\langle +|U_{\leq j}^\dagger [U_{>j}^\dagger [H_B, H_C] U_{>j}, H_B] U_{\leq j}|+\rangle = a_j, \end{aligned} \tag{C.30}$$

where we introduced the short-hand notation a_j for better readability. Analogously, considering matrix elements of the form $v_{\gamma_j,\beta_{p+1}} = \partial_{\gamma_j}\partial_{\beta_{p+1}} E(\boldsymbol{\beta},\boldsymbol{\gamma})$, we obtain

$$\partial_{\gamma_j}\partial_{\beta_{p+1}} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}_{\text{TS}}^{p+1}} = \langle +|U_{<j}^\dagger [U_{\geq j}^\dagger [H_B, H_C] U_{\geq j}, H_C] U_{<j}|+\rangle = a_{p+1+j}. \tag{C.31}$$

Evaluating the second derivatives on Eq. (C.30) and Eq. (C.31) at $j = p+1$ corresponds to the first column of the 2×2 matrix h. In particular, evaluating Eq. (C.30) at $j = p+1$ leads to $U_{>j} = \mathbb{I}$ and $U_{\leq j} = U$ which in turn implies that

$$\partial^2_{\beta_{p+1}} E(\boldsymbol{\beta}, \boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}} = \langle \mathbf{\Gamma}^{p}_{\mathsf{min}} | [[H_B, H_C], H_B] | \mathbf{\Gamma}^{p}_{\mathsf{min}} \rangle = a_{p+1}. \tag{C.32}$$

Note that above we used $U_{>p+1} = \mathbb{I}$. This is because when the derivative is taken concerning the last layer $(p+1)$ of the unitary circuit, there is no unitary to the left of it which, in the notation introduced on Eq.(C.15) is equivalent to $U_{>p+1} = \mathbb{I}$. Doing the same on Eq. (C.31) gives

$$\partial_{\gamma_{p+1}} \partial_{\beta_{p+1}} E(\boldsymbol{\beta}, \boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}} = \langle \mathbf{\Gamma}^{p}_{\mathsf{min}} | [[H_B, H_C], H_C] | \mathbf{\Gamma}^{p}_{\mathsf{min}} \rangle = b. \tag{C.33}$$

Finally, let us look at the matrix elements of the form $v_{\beta_j, \gamma_{p+1}} = \partial_{\beta_j} \partial_{\gamma_{p+1}} E(\vec{\beta}, \vec{\gamma})$ and analogously $v_{\gamma_j, \gamma_{p+1}}$, corresponding to the second column of v. Let us first inspect $\partial_{\gamma_{p+1}} E(\vec{\beta}, \vec{\gamma})$:

$$\partial_{\gamma_{p+1}} E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathrm{i} \langle + | U^\dagger_{<p+1} [H_C, U^\dagger_{p+1} H_C U_{p+1}] U_{<p+1} | + \rangle. \tag{C.34}$$

When evaluated at point $\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}$, we obtain that $[H_C, U^\dagger_{p+1} H_C U_{p+1}] = 0$ since $U_{p+1} = \mathbb{I}$ and H_C commutes with itself. Hence, we see that as long as the second derivative is taken with respect to an element (β or γ) at index $j < p+1$ the final result will be zero. As we already saw in Eq. (C.33), $\partial_{\gamma_{p+1}} \partial_{\beta_{p+1}} E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is equal to b. Using similar arguments, we show that $\partial_{\gamma_{p+1}} \partial_{\gamma_{p+1}} E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ which corresponds to the $h_{\gamma_{p+1}, \gamma_{p+1}}$ matrix element of h. We are then ready to construct the Hessian at the TS under consideration:

$$H(\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}) = \begin{pmatrix} H(\mathbf{\Gamma}^{p}_{\mathsf{min}}) & v \\ v^T & h \end{pmatrix}, \tag{C.35}$$

with

$$v^T = \begin{pmatrix} a_1 & \cdots & a_{2p+1} \\ 0 & \cdots & 0 \end{pmatrix} \quad \text{and} \quad h = \begin{pmatrix} a_{p+1} & b \\ b & 0 \end{pmatrix}. \tag{C.36}$$

Using the expression for the determinant of a block matrix [Bel97]

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A) \det(D - CA^{-1}B), \tag{C.37}$$

we rewrite the determinant of the full Hessian as follows

$$\det\Big[H(\Gamma^{p+1}_{TS})\Big] = \det \begin{pmatrix} a_{p+1} & b \\ b & 0 \end{pmatrix} \det\Big[H(\Gamma^{p}_{min}) - v h^{-1} v^T\Big] = -b^2 \det\Big[H(\Gamma^{p}_{min})\Big]. \tag{C.38}$$

We used that $v h^{-1} v^T = 0$ in the last line. We then see that as long as $b \neq 0$ the determinant of the Hessian at the TS is negative, $\det[H(\Gamma^{p+1}_{TS})] < 0$. The explicit expression (C.33) for b relates it to the expectation value of the commutator $[[H_B, H_C], H_C]$ over the variational wave function. Since this commutator is a non-vanishing operator, its expectation value is generically non-zero, $b \neq 0$. This concludes the proof of Theorem 3 for the case when zeros are inserted at the last layer of the unitary circuit.

Case (ii): $l = k = 1$

As before, we focus on computing the matrix elements of $v = v(1,1)$ and $h = h(1,1)$. Starting from the first column of v, with matrix elements v_{β_j,β_1} and v_{γ_j,β_1} for $j \in [2, p+1]$ we find

$$\begin{aligned}\partial_{\beta_j}\partial_{\beta_1} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} &= \langle +|[H_B, U_{\leq j}^\dagger [U_{>j}^\dagger H_C U_{>j}, H_B] U_{\leq j}]|+\rangle = 0,\\ \partial_{\gamma_j}\partial_{\beta_1} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} &= \langle +|[H_B, U_{<j}^\dagger [U_{\geq j}^\dagger H_C U_{\geq j}, H_C] U_{<j}]|+\rangle = 0.\end{aligned} \tag{C.39}$$

Moving onto the second column of v, with matrix elements v_{β_j,γ_1} and v_{γ_j,γ_1} for $j \in [2, p+1]$ we obtain

$$\begin{aligned}\partial_{\gamma_1}\partial_{\beta_j} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} &= \langle +|[H_C, U_{\leq j}^\dagger [U_{>j}^\dagger H_C U_{>j}, H_B] U_{\leq j}]|+\rangle = c_j,\\ \partial_{\gamma_j}\partial_{\gamma_1} E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} &= \langle +|[H_c, U_{<j}^\dagger [U_{\geq j}^\dagger H_C U_{\geq j}, H_C] U_{<j}]|+\rangle = c_{p+1+j}\end{aligned} \tag{C.40}$$

where for better readability we introduced the short-hand notation c_j with $j \in [2, p]$. Finally, evaluating the above expressions Eq. (C.39) and Eq. (C.40) at $j = 1$ leads to the matrix elements of the 2×2 matrix h. Altogether, we find

$$v^T(1,1) = \begin{pmatrix} 0 & \cdots & 0 \\ c_1 & \cdots & c_{2p+2} \end{pmatrix}, \quad h(1,1) = \begin{pmatrix} 0 & b \\ b & c_{p+2} \end{pmatrix},$$

where

$$b = \langle +|[H_C, [U^\dagger H_C U, H_B]]|+\rangle \tag{C.41}$$

and the value of c_{p+2} follows from evaluating Eq. (C.40) at $j = 1$.

Invoking once again the expression for the determinant of a block matrix Eq. (C.37) we get

$$\begin{aligned}\mathsf{det}\Big[H(\Gamma_{TS}^{p+1})\Big] &= \mathsf{det}\Big[H(\Gamma_{\min}^p))\mathsf{det}(h + v^T H(\Gamma_{\min}^p) v\Big)\\ &= \mathsf{det}\left[\begin{pmatrix} 0 & b \\ b & c_{p+2} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \mathsf{const} \end{pmatrix}\right] \mathsf{det}\Big[H(\Gamma_{\min}^p)\Big] = -b^2 \mathsf{det}\Big[H(\Gamma_{\min}^p)\Big].\end{aligned} \tag{C.42}$$

Using that the point $\boldsymbol{\Gamma}_{min}^p$ is a local minimum (with the Hessian being non-singular), we see that as long as $b \neq 0$ the determinant of the Hessian at the TS is negative. The fact that the parameter b in Eq. (C.41) is non-vanishing can be inferred from the similar argument to the one used at the end of Appendix C.2.2

Case (iii): $l, k \in 2, p$

So far we have proven that when the zeros insertion is made at the initial (I) or last (II) layer of the unitary circuit the corresponding points $\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}$ of QAOA$_{p+1}$ are TS. In both cases, we proved that the determinant of the Hessian of QAOA$_{p+1}$ at the given points is negative. In order to do this, we used that one of the columns of the $2p \times 2$ matrix v was zero which greatly simplified the computation of the determinant. In what follows, we show that these simplifications, unfortunately, do not occur when the zeros insertion is made in the bulk of the unitary circuits. However, we instead observe that the matrix $v(l,k)$ is constructed by taking the l-th (β_l) and $p+1+k$-th (γ_k) columns of the Hessian of QAOA$_p$ at the local minimum

$\mathbf{\Gamma}^p_{\text{min}}$. This fact, together with the invariance of the determinant under linear operations performed on rows or columns leads to the desired result.

We begin by explicitly computing the matrix elements of $h(l,k)$ and $v(l,k)$ and then relating them to matrix elements of the Hessian $H(\mathbf{\Gamma}^p_{\text{min}})$. For the sake of concreteness, we focus on the particular case of symmetric TS, i.e. $k=l$. The other case, i.e. $k=l+1$ can be covered by an analogous chain of arguments. As before, in what follows we drop the quantifying indices for better readability. Starting from h, we obtain

$$\begin{aligned} h &= \begin{pmatrix} \partial_{\beta_l}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}} & \partial_{\beta_l}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}} \\ \partial_{\beta_l}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}} & \partial_{\gamma_l}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}} \end{pmatrix} = \begin{pmatrix} \partial^2_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} & b \\ b & \partial^2_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} \end{pmatrix} \\ &= \begin{pmatrix} H(\mathbf{\Gamma}^p_{\text{min}})_{\beta_{l-1},\beta_{l-1}} & b \\ b & H(\mathbf{\Gamma}^p_{\text{min}})_{\gamma_l,\gamma_l} \end{pmatrix}, \end{aligned} \tag{C.43}$$

where

$$b = \langle +|U^\dagger_{\leq l-1}[H_C,[H_B,U^\dagger_{>l-1}H_C U_{>l-1}]]U_{\leq l-1}|+\rangle. \tag{C.44}$$

One might be tempted by looking at the properties listed in Eq. (C.18) to relate $\partial_{\beta_l}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p+1}_{\text{TS}}}$ to $\partial_{\beta_{l-1}}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\text{min}}}$. However, upon closer inspection, we can see that these are not the same. More specifically, we get

$$\partial_{\beta_{l-1}}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} = \langle +|U^\dagger_{\leq l-1}[H_B,[H_C,U^\dagger_{>l-1}H_C U_{>l-1}]]U_{\leq l-1}|+\rangle. \tag{C.45}$$

Comparing the above expression with Eq. (C.44) we realize that although not equal, they are related via the Jacobi identity

$$[A,[B,C]]+[B,[C,A]]+[C,[A,B]]=0, \tag{C.46}$$

for operators A,B and C. More specifically, we obtain

$$b-\partial_{\beta_{l-1}}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\text{min}}} = \langle +|U^\dagger_{\leq l-1}[U^\dagger_{>l-1}H_C U_{>l-1},[H_B,H_C]]U_{\leq l-1}|+\rangle = \bar{b}. \tag{C.47}$$

Considering now the matrix elements of v we get

$$v = \begin{pmatrix} \partial_{\beta_1}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\beta_1}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \vdots & \vdots \\ \partial_{\beta_{l-1}}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\beta_{l-1}}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \partial_{\beta_l}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\beta_l}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \vdots & \vdots \\ \partial_{\beta_p}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\beta_p}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \partial_{\gamma_1}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\gamma_1}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \vdots & \vdots \\ \partial_{\gamma_{l-1}}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\gamma_{l-1}}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \partial_{\gamma_l}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\gamma_l}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \\ \vdots & \vdots \\ \partial_{\gamma_p}\partial_{\beta_{l-1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} & \partial_{\gamma_p}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^p_{\mathsf{min}}} \end{pmatrix} = \begin{pmatrix} H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\beta_1,\beta_{l-1}} & H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\beta_1,\gamma_l} \\ \vdots & \vdots \\ H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\beta_p,\beta_{l-1}} & H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\beta_p,\gamma_l} \\ H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\gamma_1,\beta_{l-1}} & H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\gamma_1,\gamma_l} \\ \vdots & \vdots \\ H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\gamma_p,\beta_{l-1}} & H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{\gamma_p,\gamma_l} \end{pmatrix}. \tag{C.48}$$

Hence, we find that the $2p \times 2$ rectangular matrix v corresponds to the matrix formed by taking columns $H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{m,\beta_{l-1}}$ and $H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})_{m,\gamma_l}$ with $m = 1,\ldots,2p$ of $H(\boldsymbol{\Gamma}^{\boldsymbol{p}}_{\mathsf{min}})$. Using this result and the fact that the determinant is invariant under linear operations performed on rows or columns, we get that

$$\mathsf{det}(H(\boldsymbol{\Gamma}^{\boldsymbol{p+1}}_{\mathsf{TS}})) = \mathsf{det}\begin{pmatrix} H(\boldsymbol{\Gamma}^{p}_{\mathsf{min}}) & v(l,k) \\ 0 & \overline{h}(l,l) \end{pmatrix}, \tag{C.49}$$

where we subtracted rows $H(\boldsymbol{\Gamma}^{p}_{\mathsf{min}})_{\beta_{l-1},m}$ and $H(\boldsymbol{\Gamma}^{p}_{\mathsf{min}})_{\gamma_l,m}$ with $m = 1,\ldots,2p$ from v^T, and introduced

$$\overline{h} = \begin{pmatrix} 0 & \bar{b} \\ \bar{b} & 0 \end{pmatrix}, \tag{C.50}$$

Using once again the expression for the determinant of a block matrix Eq. (C.37), and the fact that $\mathsf{det}(\overline{h}(l,l)) = -\bar{b}^2$ is negative ($\bar{b} \neq 0$ due to similar argument as in Appendix C.2.2) we obtain

$$\mathsf{det}\Big[H(\boldsymbol{\Gamma}^{p+1}_{\mathsf{TS}})\Big] = -\bar{b}^2\mathsf{det}\Big[H(\boldsymbol{\Gamma}^{p}_{\mathsf{min}})\Big] < 0, \tag{C.51}$$

concluding our proof for the general TS.

C.3 Counting of unique minima

The number of minima found in the initialization graph construction presented in the main text, naïvely scales as $N_{\mathsf{min}}(p) = 2^{p-1}p!$. This follows from our recursive construction. Each local minimum of QAOA$_p$ is used to construct $p+1$ symmetric TS and for each TS we then find two new minima of QAOA$_{p+1}$, see Figs. 4.1 and 4.2. This factorial growth is, however, only sustained if every TS produces two new minima that are all distinct from each other. Numerically, we find that this is not the case and that the number of unique minima is

significantly smaller. The increase in the number of unique minima is consistent with an exponential dependence proportional to $e^{\kappa p}$ [we find that $N_{\text{min}}(p)$ can be approximated as $N_{\text{min}}(p) \approx 0.19e^{0.98p}$]. However, the limited range of p does not allow us to completely rule out factorial growth, see Fig. C.1. The much smaller number of unique minima, compared to the naïve counting demonstrates that different TS often lead to similar minima, as illustrated in Fig. 4.4.

C.4 Properties of the index-1 direction

The index-1 direction is the direction of negative curvature at a TS in a QAOA$_{p+1}$ which we use to find two new minima in QAOA$_{p+1}$, as illustrated in Fig. 4.2(a). The index-1 direction is obtained by finding the eigenvector corresponding to the unique negative eigenvalue of the Hessian, $H(\mathbf{\Gamma}_{TS}^{p+1})$. Numerically we showed in Fig. 4.2(b) that optimization initialized along the $\pm$ index-1 direction either heals or enhances the perturbation introduced by a creation of the TS from the local minima of QAOA$_p$.

Interestingly, we find that the index-1 vector has dominant components at positions where zero angles were inserted as well as the positions of adjacent angles. In contrast, all other components of the index-1 vector have nearly zero weight, as illustrated in Fig. C.2. The large contribution along the component corresponding to the zero insertion can be physically motivated by the fact that the gate with the zero parameter does initially not have any effect for driving the initial state $|+\rangle^{\otimes n}$ towards the ground state of H_C. Hence, the energy can be lowered by 'switching on' the action of this gate by moving the value of the corresponding variational angle away from zero. Interestingly, we see that the neighboring gates with non-zero parameters are also changed along the index-1 direction. The next nearest neighboring gates appear to be not involved in this process. We note that this numerical observation allows to *a priori* guess the index-1 direction without having to diagonalize the Hessian $H(\mathbf{\Gamma}_{TS}^{p+1})$. This may be useful for the practical implementation of our initialization on available quantum computers.

C.5 Description of GREEDY algorithm

In the following, we provide a detailed description for the GREEDY QAOA initialization, as well as the `subroutines` required to implement the algorithm. To this end, we first provide

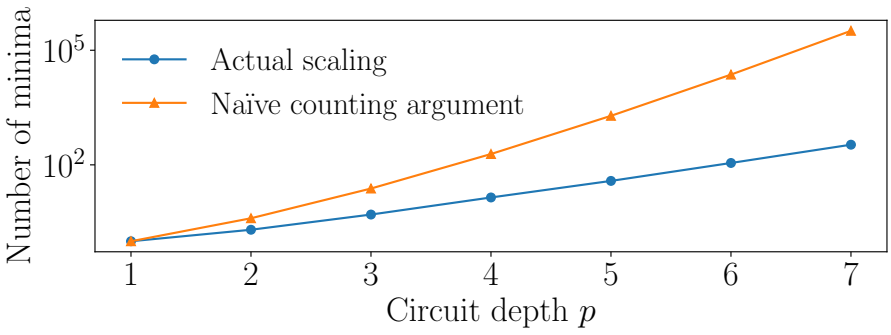

Figure C.1: Number of minima found in the initialization graph in Fig. 4.2 with system size $n = 10$. The orange line describes a naïve counting argument ($2^{p-1}p!$) while the blue line lists the actual number of distinct minima that can be approximated as $0.19\, e^{0.98p}$.

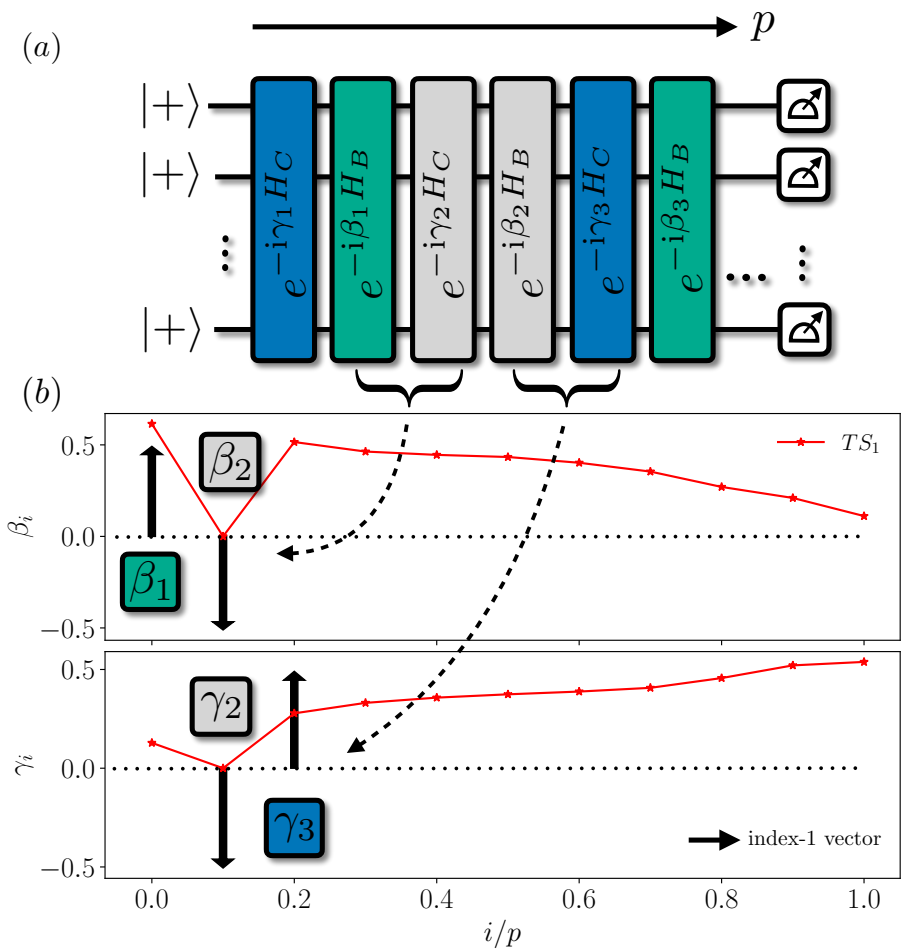

Figure C.2: (a) Illustration of the circuit implementing the QAOA at a TS. Gray gates correspond to the zero insertion. The index-1 direction has mainly weight at the position of the zeros as well as the two adjacent gates. (b) Numerical example of the index-1 vector and the QAOA parameter pattern at the TS. Arrows correspond to the magnitude and sign of the entries in the index-1 direction. Only entries at $\beta_1, \beta_2, \gamma_2$ and γ_3 have a large magnitude, all other entries are nearly zero.

a pseudo-code for a gradient-based QAOA parameter optimization routine. The algorithm is a so-called variational hybrid algorithm, which implies that the quantum computer is used in a closed feedback loop with a classical computer. There the quantum computer is used to implement a variational state and measure observables while the classical computer is used to keep track of the variational parameters and update them in order to minimize the energy expectation value.

Algorithm 2 `QAOA subroutine`

1: Given the circuit depth p, choose initial parameters $\mathbf{\Gamma}^p_{\text{init.}} = (\boldsymbol{\beta}_{\text{init.}}, \boldsymbol{\gamma}_{\text{init.}})$
2: **repeat**
3: Implement $|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle$ on a quantum device
4: Estimate $E(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle \boldsymbol{\beta}, \boldsymbol{\gamma} | H_C | \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle$
5: Estimate gradient $\nabla E(\boldsymbol{\beta}, \boldsymbol{\gamma})$
6: Update $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ using gradient information
7: **until** $E(\boldsymbol{\beta}, \boldsymbol{\gamma})$ has converged
8: Return minimum $\mathbf{\Gamma}^p_{\text{min}}$

For very shallow circuit depths, such as $p = 1$, the optimization landscape is sufficiently low dimensional and simple such that global optimization routines can be used to find the optimal parameters. One of the most straightforward global optimization routines is the so-called grid search. There, the parameters are initialized on a dense grid and a parameter optimization routine, such as the `QAOA sub-routine` is carried out for each point in the grid. Then, only the lowest energy local minimum is kept.

Algorithm 3 `Grid search subroutine`

1: Given a circuit depth p, construct an evenly spaced grid on the fundamental region:

$$\beta_i \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]; \;\; \gamma_1 \in \left(0, \frac{\pi}{4}\right), \;\; \gamma_j \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right], \tag{C.52}$$

with $i \in [1, p]$ and $j \in [2, p]$

2: `QAOA subroutine` initialized from each point in grid

3: Return local minimum with the lowest energy $\mathbf{\Gamma}^p_{\min}$

Using the two `subroutines` presented above we can provide a detailed pseudo-code for the GREEDY QAOA algorithm, see Fig. C.3 for a visualization.

Algorithm 4 `Greedy QAOA`

1: Choose maximum circuit depth $p_{\max}$
2: Choose small offset $\epsilon \ll 1$
3: Grid search for $p = 1$ to find $\mathbf{\Gamma}^{p=1}_{\min}$ ▷ See `Grid search subroutine`
4: **repeat**
5: Construct $p + 1$ symmetric TS $\mathbf{\Gamma}^{i,p+1}_{TS}$ from $\mathbf{\Gamma}^p_{\min}$
6: Compute or approximate the index-1 unit vector $\hat{\boldsymbol{v}}$ for each TS
7: Construct points $\mathbf{\Gamma}^{i,p+1}_{\pm} = \mathbf{\Gamma}^{i,p+1}_{TS} \pm \epsilon \hat{\boldsymbol{v}}_i$ for each TS
8: Run QAOA init. from $\mathbf{\Gamma}^{i,p+1}_{\pm}$ ▷ See `QAOA subroutine`
9: Keep local minimum with the lowest energy $\mathbf{\Gamma}^{p+1}_{\min}$
10: $p \longleftarrow p + 1$
11: **until** $p = p_{\max}$
12: Return minimum $\mathbf{\Gamma}^{p=p_{\max}}_{\min}$

The index-1 direction $\hat{\boldsymbol{v}}_i$ can either be found explicitly by diagonalizing the Hessian matrix or using the heuristic approximation outlined in the previous section. While explicit diagonalization incurs classical computation costs that scale polynomially with p, and thus can be done efficiently, an approximation to index-1 direction is expected to give a similar performance of QAOA subroutine at a lower classical computational cost.

C.6 Additional graph ensembles and system size scaling

In the main text, we numerically investigated the performance of our method on random 3-regular graphs (RRG3) with system size $n = 10$. In the following, we present results for larger system sizes as well as two more graph types. Namely, weighted 3-random regular graphs (RWRG3) where the Hamiltonian is given by $H_C = \sum_{\langle i,j \rangle \in E} w_{ij} \sigma^z_i \sigma^z_j$ and w_{ij} are random

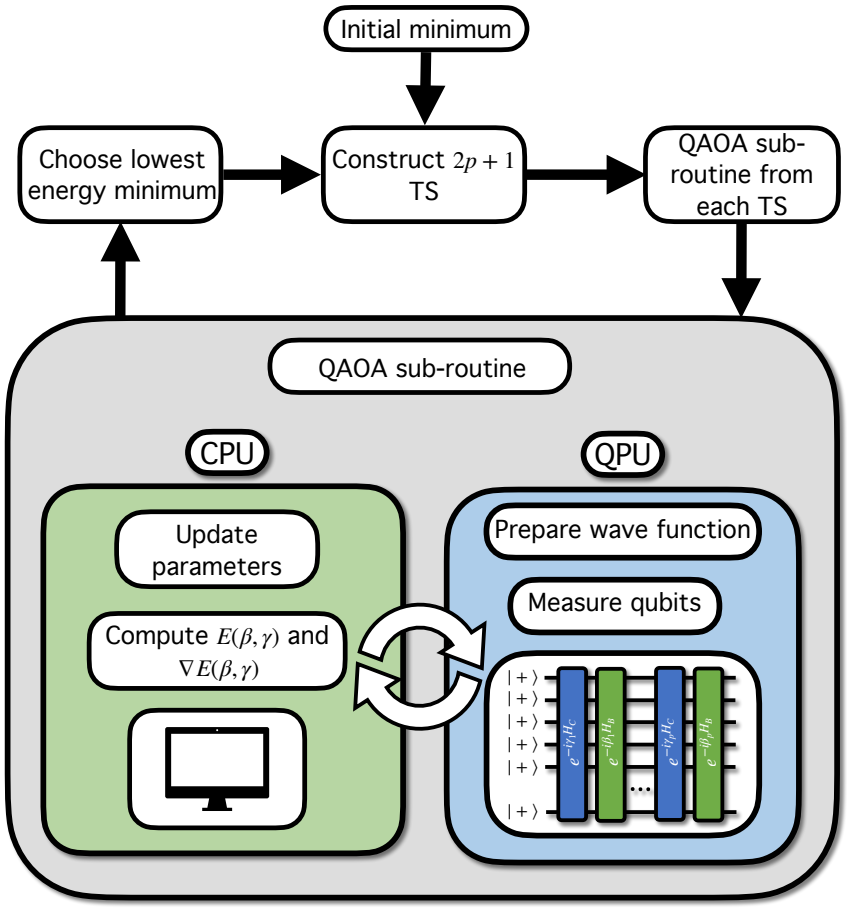

Figure C.3: Flow diagram to visualize the Greedy QAOA initialization algorithm presented in Algorithm 4.

weights $w_{ij} \in [0, 1)$, as well as random ErdAos-Rényi graphs (RERG) with edge probability $p_E = 0.5$.

Fig. C.4 shows the performance comparison between Greedy, TQA, and Interp on RWRG3 and RERG. We can see that for RWRG3 the performance of the three methods is comparable, while for RERG the TQA performs worse than the other two methods. Greedy and Interp yield (nearly) the same performance for both graph ensembles on the system size that we considered ($n = 10$).

Fig. C.5 compares the performance for RRG3 with different system sizes. Interp and Greedy yield very similar performance for smaller system sizes ($n = 8$ indicated by light color) while they yield the same performance for larger system sizes ($n = 16$ indicated by dark color). TQA performs slightly worse than Greedy and Interp for all system sizes considered. We can furthermore see that the gain in performance from every additional layer is becoming less for bigger system sizes. This is because for the QAOA to "see" the whole graph, a circuit depth p scaling as $p \sim \log n$ is required [FGG20a].

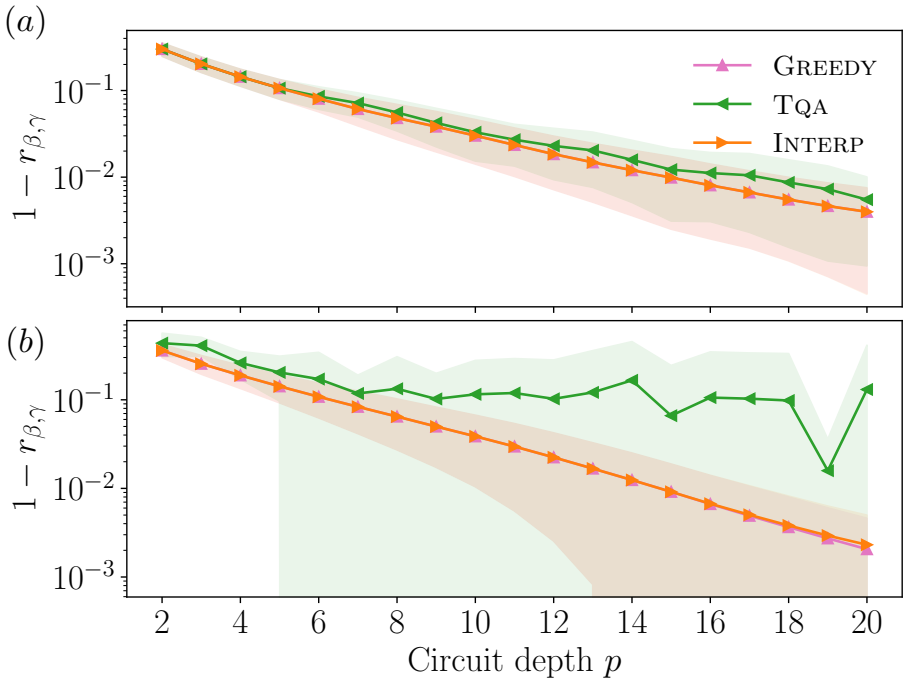

Figure C.4: Performance comparison on (a) RWRG3 and (b) RERG with system size $n = 10$. Data is averaged over 19 non-isomorphic graphs.

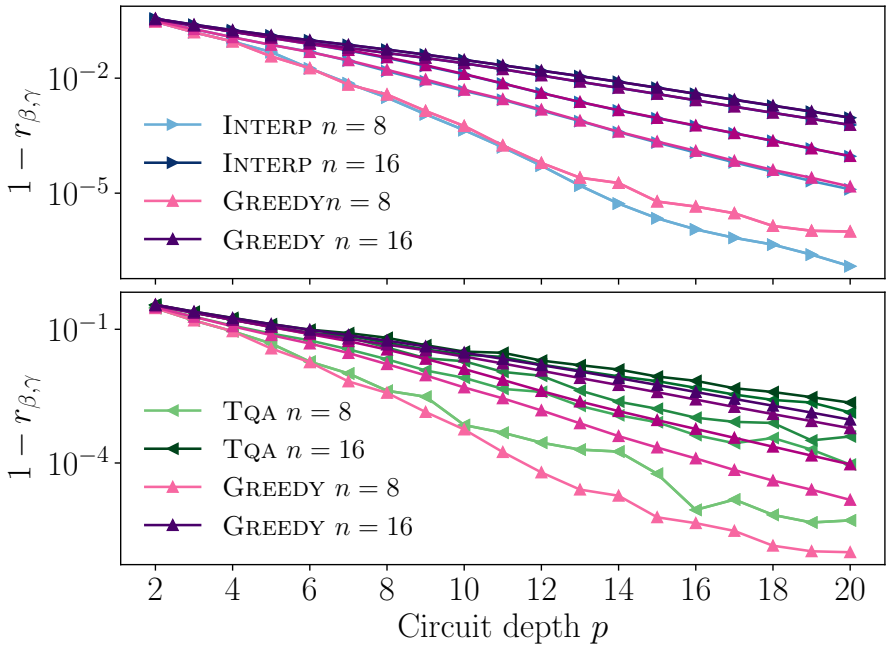

Figure C.5: System size scaling for performance comparison on RRG3. Color shade indicates system size, light color is $n = 8$ and dark color is $n = 16$. System size changes in steps of two between those values. Data is averaged over 19 non-isomorphic RRG3 graphs.

APPENDIX D

Appendices to Chapter 4

D.1 Numerical simulations

All the simulations performed in this work were conducted using the Julia programming language [BEKS17] and the package `QAOALandscapes.jl`[Med24], which was developed by one of the authors. This package is designed to apply the QAOA to solve general combinatorial optimization problems by encoding the classical problem into a k-spin classical Hamiltonian. It relies on matrix-free operations to enhance speed and reduce memory usage. Currently, it supports only the Pauli-X mixer operator (see Eq.(5.3)); however, additional mixers can be incorporated as is described in the documentation. The package includes support for both CPU and GPU backends, with GPU capabilities for CUDA and Metal-based devices through `CUDA.jl` [**?** **?**] and `Metal.jl` [**?**] packages respectively.

Numerical optimization was performed using the `Optim.jl` [MR18, MWR+24] package, and in particular using the BFGS [Bro70, Fle70, Gol70, Sha70] algorithm. For this, the package supports fast and exact gradient calculations through the use of automatic differentiation using the method introduced in [LLZW20, JG20]. Further details on how to use `QAOALandscapes.jl` can be found on the Readme and documentation in [Med24].

D.1.1 Quality of optimization using only $\Gamma^{p+1}_{\mathrm{TS}}(1,1)$

The GREEDY approach introduced in [SMKS23], requires one to launch optimization twice from each of the $2p+1$ TS constructed from a local minimum $\Gamma^p_{\min}$ of QAOA$_p$. The need to launch optimization from $2p+1$ distinct transition states and moving in two potential directions away from the saddle accounts for a $2(2p+1)$ overhead on top of heuristic initializations like INTERP and FOURIER [ZWC+20] with similar performance. Even though GREEDY comes with a guarantee of improvement at each circuit depth p, it is desirable to further reduce the optimization cost that it incurs.

We thus inspect given a local minimum $\Gamma^p_{\min}$ what fraction of the $2p+1$ TS constructed from it leads to the GREEDY solution after optimization. We numerically observe in Fig. D.1 that the number of TS that connect through optimization to the best local solution decreases with p but remains finite at approximately 0.7 at circuit depth $p=20$. In all cases, we observed that the transition state constructed by padding with zeros the first layers, i.e. $\Gamma^p_{\mathrm{TS}}(1,1)$ in the notation of [SMKS23], leads to the GREEDY solution as also shown in the figure, where

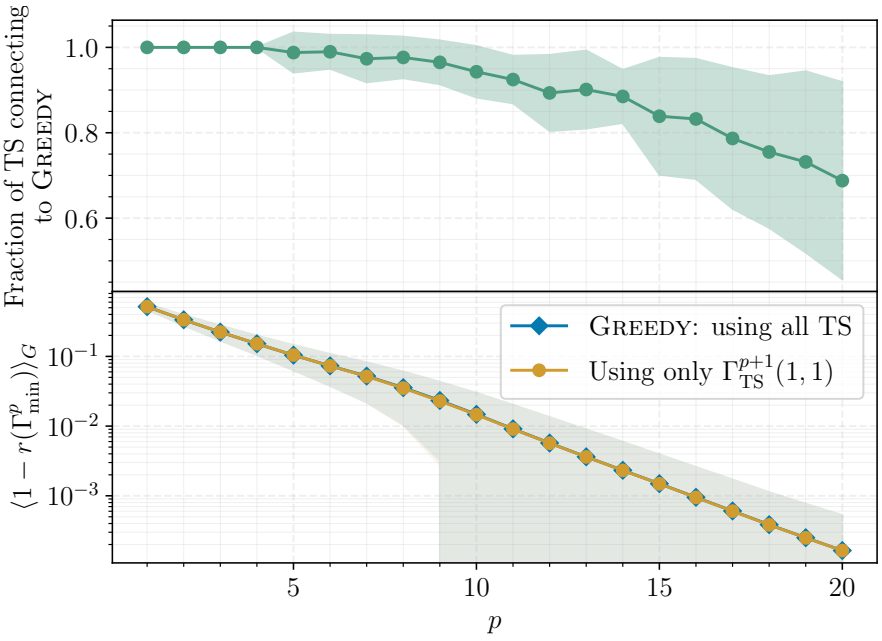

Figure D.1: (*Top*) Fraction of the $2p+1$ TS constructed from a local minima $\Gamma^p_{\min}$ that connect to the Greedy solution. The data corresponds to instances of random 3-regular unweighted graphs with $N = 12$ vertices. (*Bottom*) Performance of numerical optimization using only the transition state with zeros padded at indices $(\beta, \gamma) = (1, 1)$. The average performance, over instances of 3-regular unweighted graphs with $N = 12$ vertices seems effectively identical to that of the Greedy strategy [SMKS23] that uses the set of all $2p+1$ TS constructed from a local minima of QAOA$_p$.

we plot the average approximation ratio obtained from using the Greedy strategy, and the result coming from only using $\Gamma^p_{\mathrm{TS}}(1,1)$.

This observation motivates us to focus on the transition state $\Gamma^p_{\mathrm{TS}}(1,1)$ as it provides a reliable choice of an initial transition state. Fixing the transition state in such a way reduces the cost of optimization to a simple factor of two, while still keeping the guarantee of improvement. Furthermore, as we will show below, using $\Gamma^p_{\mathrm{TS}}(1,1)$ enables us to obtain a lower bound on the energy improvement of the QAOA between circuit consecutive circuit depths.

D.1.2 Converging to an excited state

Here we provide a specific example of QAOA converging to an excited state. To this end, we use a MaxCut instance studied in Ref. [ZWC+20]. This instance was used to highlight the performance of the Fourier and Interp strategies [ZWC+20] on "hard" instances of MaxCut. The authors used the minimum spectral gap $\Delta_{\min}$ of the annealing Hamiltonian

$$H_{\mathrm{QA}}(s) = sH_C + (1-s)H_B;\ s \in [0,1]$$

to distinguish between "hard" and "easy" instances for optimization. In particular, for the instance shown in Fig. D.2 one can show that $\Delta_{\min} < 10^{-3}$, which translates into prohibitively long annealing time [ZWC+20].

Both Fourier and Interp initialization strategies reuse a local minimum of the QAOA at circuit depth p to construct an initialization at circuit depth $p+1$. In the case of the Fourier initialization — which will also use here — the idea is to use a different parametrization

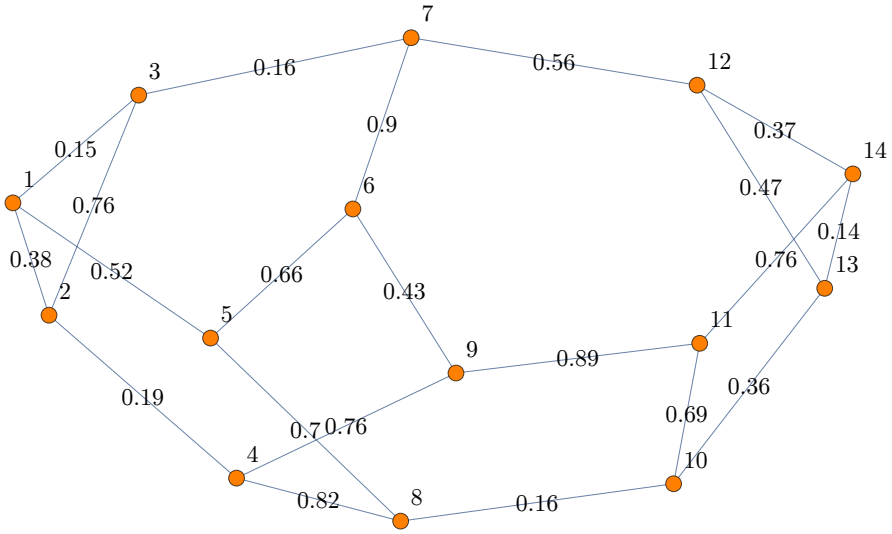

Figure D.2: Instance of MaxCut with $N = 14$ vertices where the QAOA algorithm gets trapped in local optima, and mostly converges to the first excited state of the cost Hamiltonian H_C.

of QAOA. Instead of using the $2p$-parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, Ref. [ZWC+20] considers the discrete cosine and sine transform of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively

$$\beta_i = \sum_{k=1}^{q} v_k \cos\Big[(k-\frac{1}{2})(i-\frac{1}{2})\frac{\pi}{p}\Big], \tag{D.1}$$

$$\gamma_i = \sum_{k=1}^{q} v_k \sin\Big[(k-\frac{1}{2})(i-\frac{1}{2})\frac{\pi}{p}\Big]. \tag{D.2}$$

Through such coordinate transformation, the new parameters become the amplitudes $(\boldsymbol{v}, \boldsymbol{u})$ of the frequency components for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. The basic $\text{Fourier}[\infty, 0]$ variant of the strategy, generates a good initial point for QAOA_{p+1} by adding a higher frequency component, initialized at zero amplitude, to the optimum at level p. Last, in the improved variant $\text{Fourier}[\infty, R]$ in addition to optimizing according to the basic strategy, we optimize QAOA_{p+1} from $R+1$ extra initial points, R of which are generated by adding random perturbations to the best of all local optima $(\boldsymbol{v}, \boldsymbol{u})$ found at level p (see the Appendix B.2 in Ref. [ZWC+20] for more details). It is crucial to note that the number of random perturbations in Fourier space, denoted by R, serves as a hyperparameter; optimizing its value is essential for enhanced convergence, requiring multiple runs of the optimization process with varied R settings to determine the most effective configuration.

In Fig. D.3 we compare the performance of the QAOA under the $\text{Fourier}[\infty, 0]$, $\text{Fourier}[\infty, 10]$, Greedy, and Γ_{TS}^{p+1} strategies. From the approximation ratio, we note that all strategies yield the same performance for circuit depths $p \in [1, 24]$, yet the improvement of the approximation ratio is stalled for $p \geq 10$. We attribute this behavior to the fact that QAOA prepares an excited state of the system with progressively increased fidelity, see the middle panel of Fig. D.3. Eventually, however, the QAOA is able to escape the local minimum that prepares an excited state and starts converging to the ground state.

When the QAOA escapes the local minimum that prepares an excited state of the classical Hamiltonian depends on the initialization scheme. The initialization $\text{Fourier}[\infty, 10]$ is the first to escape local optima at $p \sim 24$. Greedy follows next and only manages to escape

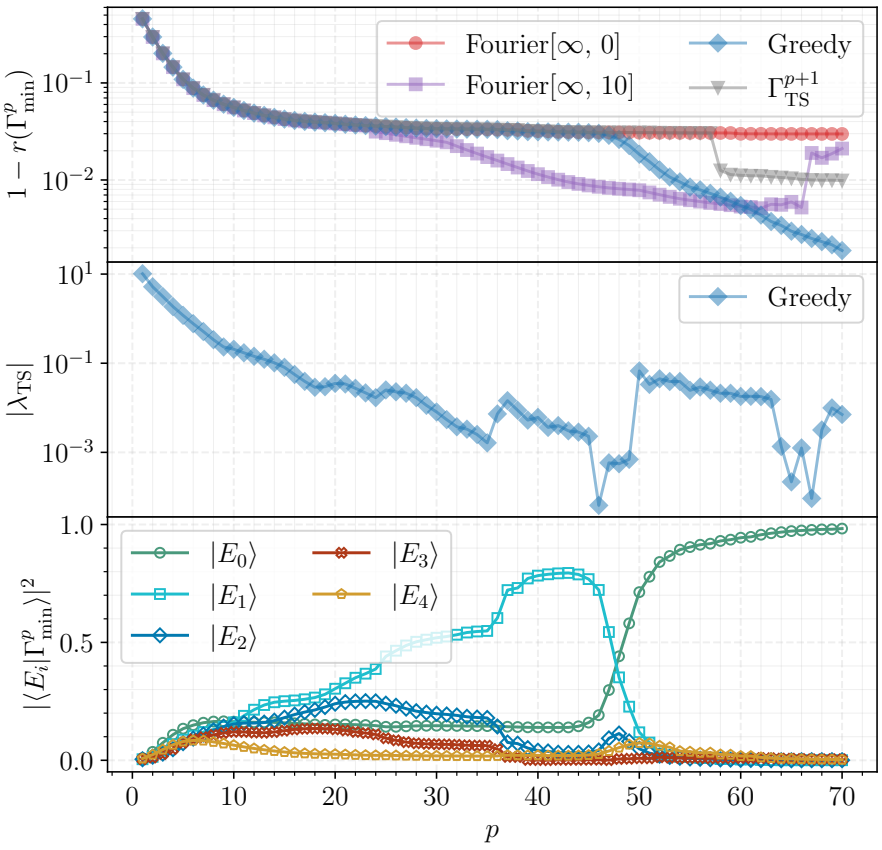

Figure D.3: (*Top*) Behavior of the approximation ratio as a function of the circuit depth, for different optimization strategies. (*Middle)* Probability of measuring the fifth lowest energy eigenstates as a function of the circuit depth for the Greedy strategy. The ground state population remains unchanged for a wide range of circuit depths, followed by a sudden increase which correlates with the QAOA overcoming local minima. (*Bottom*) Circuit depth dependence of the landscape curvature at the transition state defined in Eq. (5.6) following the Greedy strategy. The curvature displays a gradual decrease, followed by a significant increase when the QAOA overcomes local minima.

around $p \sim 46$. Despite this, we observe that the final approximation is consistently better in Greedy than in all other strategies. Interestingly, we see that using only the first transition state as an initialization yields worse performance than Greedy and does escape the local minima that prepares the excited state but at circuit depths $p \sim 58$. Moreover, the initial variant of the Fourier strategy fails to escape from local optima and gets stalled for all circuit depths explored.

Following our comparative analysis, we articulate two critical observations. First, we conclude that QAOA is capable of preparing low-lying excited states of the classical Hamiltonian. How soon QAOA escapes from such a trap depends on the initialization scheme used, but this phenomenon is present for all initialization routines considered here. Second, we conclude that the landscape curvature at the initial transition state $\Gamma^{p+1}_{\mathrm{TS}}$, quantified in Eq. (5.10), is a good indicator of such QAOA behavior. Indeed, the bottom panel in Fig. D.3 demonstrates that although the approximation ratio is stalling, the curvature keeps decreasing when QAOA prepares the excited state with progressively higher fidelity. As soon as QAOA starts converging to the true ground state, the curvature shows an increase and then continues to reduce.

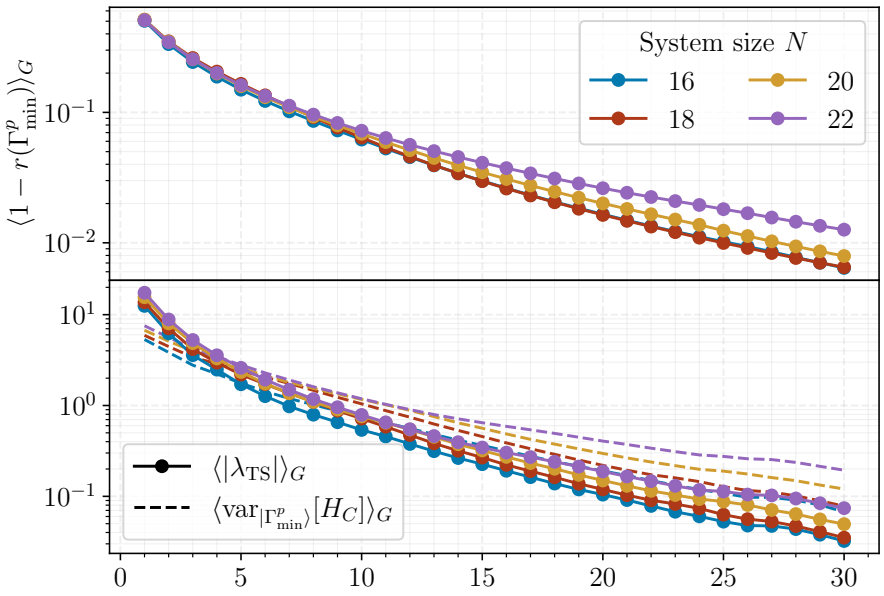

Figure D.4: (*Top*) Circuit depth dependence of the approximation ratio $r(\Gamma^p_{\min})$. The scaling of the approximation ratio with the circuit depth p matches the numerical results from [ZWC+20]. (*Bottom*) Relationship between the magnitude of the negative curvature around the transition state $\Gamma^{p+1}\mathrm{TS}$ and the energy variance $\mathrm{var}_{\Gamma^p_{\min}}[H_C]$ as functions of circuit depth p. Although there appears to be qualitative agreement between the curvature and the energy variance across varying system sizes N, it is not as close as for the unweighted instances.

All in all, our results suggest that there may be scenarios, particularly at large system sizes N, where the QAOA experiences stagnation, with negligible performance gains across a wide range of circuit depths p. This stagnation primarily arises because the QAOA effectively converges to a low-energy manifold of H_C, characterized by a small landscape curvature. In the context evaluated in this study Fig. D.2, this manifold principally consists of the ground state and the first excited state Fig. D.3. Our results suggest that the curvature provides useful insights into the QAOA behavior, complementary to the behavior of the approximation ratio.

D.1.3 Numerical results for weighted 3-regular graphs

In this section we present numerical results analogous to those shown in the main text, but for weighted instances of MaxCut on 3-regular graphs.

We start by examining the curvature of the QAOA energy landscape at the transition state from Eq.(5.6), alongside the variance of the cost Hamiltonian in the QAOA state. In Fig. D.4 we notice behavior similar to that for unweighted instances described in the main text, albeit with notable differences. First, the approximation ratio decays slower with the system size, aligning with findings from previous studies on similar MaxCut instances [ZWC+20]. Second, while there is a close qualitative relationship between the energy variance of the QAOA state and the landscape curvature at the transition state in Eq. (5.6), this relationship is not as accurate as for unweighted instances, and energy variance decays slower compared to the curvature with QAOA depth p.

Finally, we check the accuracy of the lower bound on energy improvement from Eq. (5.16). Similar to observations with unweighted MaxCut instances, the lower bound significantly underestimates the energy improvement achieved by QAOA between consecutive circuit depths.

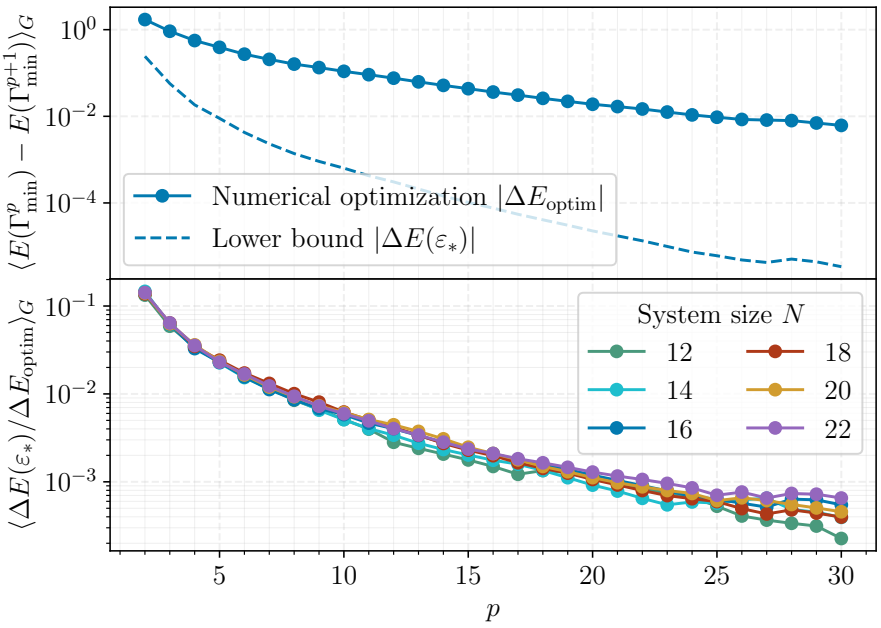

Figure D.5: (*Top*) Average energy improvement between local minima of QAOA$_p$ and QAOA$_{p+1}$ as a function of the circuit depth p. The lower bound Eq. (5.16), which relies on local information about the cost function landscape around index-1 saddle points overestimates the results obtained by numerically optimizing using the Greedy strategy of [SMKS23]. (*Bottom*) Averaged quality of the lower bound on the energy improvement, as given by $\Delta E(\varepsilon_*)/\Delta E_{\text{optim}}$, for systems sizes ranging from 12 to 22 vertices.

At the same time, the ratio between true energy improvement and our lower bound seems to be a universal function of p across a broad range of system sizes considered here.

D.2 Bounds on the Hessian eigenvalue and eigenvector.

In this Appendix, we first construct upper and lower bounds for the minimum Hessian eigenvalue at the TS. In the second part of the Appendix, we introduce an approximation for the eigenvector associated with the minimum Hessian eigenvalue and show that its expectation value provides a tighter upper bound for the minimum Hessian eigenvalue. For clarity, throughout this Appendix we focus on symmetric TS. That is, TS where the zero insertion is made at the same layer l for β and γ components. We fully describe our construction for the case $l \in [2, p+1]$ and only provide the final expression for the remaining eigenvectors.

D.2.1 Bound on the minimum Hessian eigenvalue

Given $\Gamma^p_{\min}$, a local minima of QAOA$_p$, let $\Gamma^{p+1}_{\text{TS}}(l,l)$ be the TS constructed by padding the l-th layer of the QAOA$_p$ circuit with zeros. For clarity, we will omit the layer index whenever possible.

The starting point is to apply a change of basis P that takes the Hessian at Γ^{p+1}_{TS} to the following generic form:

$$H(\Gamma^{p+1}_{\text{TS}}) \mapsto H_P(\Gamma^{p+1}_{\text{TS}}) = P^T H(\Gamma^{p+1}_{\text{TS}}) P = \begin{pmatrix} H(\Gamma^p_{\min}) & v(l,l) \\ v^T(l,l) & h(l,l) \end{pmatrix}. \tag{D.3}$$

Ref. [SMKS23] demonstrated that the $2p \times 2$ rectangular matrix $v(l,l)$ is constructed by taking the $l-1$-th and the $p+l$-th columns of $H(\Gamma^p_{\text{min}})$. Using this knowledge, we can apply a composition of two elementary transformations on the rows and columns of $H(\Gamma^p_{\text{min}})$. More specifically, let us define the following $D \times D$ matrices:

1. $R^D_{i,j}(m)$ is the identity matrix that has an additional non-zero entry m in the (i,j) position. Note that when applied on the left to a matrix A, the resulting matrix will have $[A]_{i,x} \mapsto [A]_{i,x} + m[A]_{j,x}$. The inverse of this matrix is simply $(R^D_{i,j}(m))^{-1} = R^D_{i,j}(-m)$
2. $C^D_{i,j}(m)$ is the identity matrix with an additional non-zero entry m in the (j,i) position. Note that $C^D_{i,j}(m)$=$(R^D_{i,j}(m))^T$.

Using the above definitions, we bring the Hessian at point Γ^{p+1}_{TS} to a block diagonal form:

$$H_{\text{block}} = \mathcal{R} H(\Gamma^{p+1}_{\text{TS}}) \mathcal{R}^T = \begin{pmatrix} H(\Gamma^p_{\text{min}}) & 0 \\ 0 & \bar{h} \end{pmatrix}, \tag{D.4}$$

$$\mathcal{R} = R^{2(p+1)}_{2(p+1),p+l}(-1) R^{2(p+1)}_{2p+1,l-1}(-1), \tag{D.5}$$

$$\bar{h} = \begin{pmatrix} 0 & \bar{b} \\ \bar{b} & 0 \end{pmatrix}, \tag{D.6}$$

where

$$\bar{b} = \partial_{\beta_l} \partial_{\gamma_l} E(\Gamma^{p+1}_{l,l}) - \partial_{\beta_{l-1}} \partial_{\gamma_l} E(\Gamma^p_{\text{min}}) = \langle + | U^\dagger [H_C, U_{\geq l} [H_C, H_B] U^\dagger_{\geq l}] U | + \rangle. \tag{D.7}$$

The transformation defined above subtracts rows $l-1$ and $p+l$ of $H(\Gamma^p_{\text{min}})$ to the first and second row of v^T respectively. Then, we apply the same operation but on the columns. It is important to note that the eigenvalues of $H_P(\Gamma^{p+1}_{\text{TS}})$ do change under $\mathcal{R}$. This is because the transformation we applied is not a similarity transformation. To see this, note that by definition $\mathcal{R}^T \neq \mathcal{R}^{-1}$. However, we can use this block diagonal form to get some useful information on the minimum eigenvalue of $H_P(\Gamma^{p+1}_{\text{TS}})$. Recall the definition of the minimum eigenvalue of a squared matrix M is

$$\lambda_{\text{min}}(M) = \inf_{\|\psi\|=1} \langle \psi | M | \psi \rangle.$$

Using this definition on $\lambda_{\text{min}}(H_{\text{block}})$ we get

$$\lambda_{\text{min}}(H_{\text{block}}) = \inf_{\|\psi\|=1} \langle \psi | H_{\text{block}} | \psi \rangle = \inf_{\|\psi\|=1} \langle \psi | \mathcal{R} H_P(\Gamma^{p+1}_{\text{TS}}) \mathcal{R}^T | \psi \rangle.$$

Multiplying by 1 in the form of $\|\mathcal{R}^T|\psi\rangle\|_2^2 / \|\mathcal{R}^T|\psi\rangle\|_2^2$ and further using that

$$\|\mathcal{R}^T|\psi\rangle\|_2 \leq \|\mathcal{R}^T\|_2 \||\psi\rangle\|_2 = \|\mathcal{R}^T\|_2,$$

we obtain

$$\|\mathcal{R}^T\|_2^{-2} \lambda_{\text{min}}(H_{\text{block}}) \geq \lambda_{\text{min}}(H_P(\Gamma^{p+1}_{\text{TS}})). \tag{D.8}$$

Doing the same on $\lambda_{\text{min}}(H_P(\Gamma^{p+1}_{\text{TS}}))$ we get

$$\lambda_{\text{min}}(H_P(\Gamma^{p+1}_{\text{TS}})) \geq \|(\mathcal{R}^{-1})^T\|_2^2 \lambda_{\text{min}}(H_{\text{block}}). \tag{D.9}$$

In conclusion, by utilizing the fact that $H(\Gamma^{p+1}_{\text{TS}})$ and $H_P(\Gamma^{p+1}_{\text{TS}})$ possess identical spectra, we can establish the following bounds for the minimum eigenvalue of the Hessian at the TS

$$-\|(\mathcal{R}^{-1})^T\|_2^2 |\bar{b}| \leq \lambda_{\text{min}}(H(\Gamma^{p+1}_{\text{TS}})) \leq -\|\mathcal{R}^T\|_2^{-2} |\bar{b}|, \tag{D.10}$$

where we used that $\lambda_{\text{min}}(H_{\text{block}}) = -|\bar{b}|$. With the above inequalities, the bound on the minimum eigenvalue reduces to obtaining the operator norm of $\mathcal{R}$. This is, in general, a tough task but due to the simple form of the matrix $\mathcal{R}$ defined in Eq. (D.5) in this particular case, we can compute it. The operator norm is the maximum singular value of $\mathcal{R}$, which equivalently corresponds to the maximum eigenvalue of $\mathcal{R}\mathcal{R}^T$, which we can easily compute. In particular, we calculate the spectrum of the matrix $\mathcal{R}\mathcal{R}^T$ to consist of three different eigenvalues, $\sqrt{\frac{1}{2}(3+\sqrt{5})}, \sqrt{\frac{1}{2}(3-\sqrt{5})}$ and 1, with multiplicities $2, 2$ and $2p-2$ respectively. Thus, we obtain that

$$|\mathcal{R}\|_2 = \sqrt{\frac{1}{2}(3+\sqrt{5})}. \tag{D.11}$$

The same is true for the inverse of $\mathcal{R}$. With this, we arrive at the following bound

$$-\frac{3+\sqrt{5}}{2}|\bar{b}| \leq \lambda_{\text{min}}(H(\Gamma_{\text{TS}}^{p+1})) \leq -\frac{2}{3+\sqrt{5}}|\bar{b}|. \tag{D.12}$$

This provides upper and lower bounds on the magnitude of the minimum Hessian eigenvalue at the TS. Below, we introduce an approximation for the eigenvector associated with the minimum Hessian eigenvalue. Moreover, we show that the corresponding Rayleigh coefficient improves the upper bound provided in Eq. (D.12).

D.2.2 Eigenvector approximation

We define an analogous matrix to $\mathcal{R}$ in Eq. (D.5) that acts on columns:

$$\mathcal{C} = C^{2(p+1)}_{2(p+1),p+l}(-1/2)C^{2(p+1)}_{2p+1,l-1}(-1/2). \tag{D.13}$$

We now apply the transformation $\mathcal{R}^{-1}\mathcal{C}$ to $H_P(\Gamma_{\text{TS}}^{p+1})$, which gives the following expression:

$$\tilde{H}_P(\Gamma_{\text{TS}}^{p+1}) = \mathcal{C}^{-1}\mathcal{R}H_P(\Gamma_{\text{TS}}^{p+1})\mathcal{R}^{-1}\mathcal{C} = \begin{pmatrix} H(\Gamma_{\text{min}}^{p}) + M & 0 \\ 0 & \bar{h}/2 \end{pmatrix} + \begin{pmatrix} 0 & u^T/4 \\ u & 0 \end{pmatrix}. \tag{D.14}$$

The matrix M features two non-zero columns at indices $j = l-1$ and $j = p+l$, mirroring the structure of the Hessian matrix at the local minimum, $H(\Gamma_{\text{min}}^{p})$. It includes an additive correction of $\bar{b}/2$ at specific elements $M_{l-1,p+l}$ and $M_{p+l,l-1}$. More specifically,

$$M_{i,j} = H(\Gamma_{\text{min}}^{p})_{i,j}\Big(\delta_{j,l-1}\Big(1+\frac{\bar{b}}{2}\delta_{i,p+l}\Big) + \delta_{j,p+l}\Big(1+\frac{\bar{b}}{2}\delta_{i,l-1}\Big)\Big). \tag{D.15}$$

Finally, we have the $2 \times 2p$ matrix u with non zero entries equal to $\bar{b}$ at positions $\{1, p+l\}$ and $\{2, l-1\}$. It is important to note that the applied transformation here, in contrast to the one constructed in Eq. (D.5), is a *similarity transformation*. Thus, it preserves the spectrum of the Hessian while the eigenvectors are transformed as $v \mapsto \tilde{v} := \mathcal{C}^{-1}\mathcal{R}v$.

We then define the approximate Hessian eigenstate to be

$$\tilde{v}_{\text{bound}} = \begin{pmatrix} \mathbf{0}_{2p} \\ 1/\sqrt{2} \\ -\text{sign}(\bar{b})/\sqrt{2} \end{pmatrix}. \tag{D.16}$$

Computing the expectation value of the Hessian on $\tilde{v}_{\text{bound}}$ we get:

$$\lambda_{\text{TS}} = -\Big|\bar{b}\Big|/2, \tag{D.17}$$

which, taking into account that $\frac{1}{2} > \frac{2}{3+\sqrt{5}}$, improves the previously obtained upper bound on the minimum Hessian eigenvalue at the TS in Eq. (D.12)

$$-\frac{3+\sqrt{5}}{2}|\bar{b}| \leq \lambda_{\min}(H(\Gamma_{\mathrm{TS}}^{p+1})) \leq \lambda_{\mathrm{TS}} = -\frac{|\bar{b}|}{2}. \tag{D.18}$$

The last step is then to find the expression of $\tilde{v}_{\mathrm{bound}}$ in the original basis $\delta_{\mathrm{TS}} = P^T\mathcal{R}^{-1}\mathcal{C}\tilde{v}_{\mathrm{bound}}$, that results in the following form

$$[\delta_{\mathrm{TS}}]_j = \begin{cases} \frac{1}{2} & \text{if } j = l-1, l \\ -\frac{\mathrm{sign}(\bar{b})}{2} & \text{if } j = p+l+1, \\ \frac{\mathrm{sign}(\bar{b})}{2} & \text{if } j = p+l+2, \\ 0 & \text{otherwise.} \end{cases} \tag{D.19}$$

Thus, we see that the approximate Hessian eigenvector has weights at layer l as well as the two adjacent gates. The sparse structure of $\boldsymbol{\delta}_{\mathrm{TS}}$ comes from the fact that $\mathcal{R}^{-1}\mathcal{C}$ has only four non-zero offdiagonal matrix elements in positions $(2(p+1), p+l), (2p+1, l-1), (p+l, 2(p+1))$, and $(l-1, 2p+1)$. Thus, we see that when acting $\tilde{v}_{\mathrm{bound}}$, we will get a vector with also four non-zero elements. Since $\mathcal{R}^{-1}\mathcal{C}$ is not an orthogonal matrix, then it is needed to normalize the resulting vector $\mathcal{R}^{-1}\mathcal{C}\tilde{v}_{\mathrm{bound}}$. Finally, the action of the permutation will just reorder the elements in the vector without changing its content.

Finally, we list without derivation eigenvector and eigenvalue approximations for the remaining TS:

1. For $l = 1$ we have that:

$$[\delta_{\mathrm{TS}}]_j = \begin{cases} -\frac{\mathrm{sign}(b)}{\sqrt{2}} & \text{if } j = l, \\ \frac{1}{2} & \text{if } j = p+1+l, \\ -\frac{1}{2} & \text{if } j = p+1+l+1, \\ 0 & \text{otherwise.} \end{cases} \tag{D.20}$$

   with $b = \langle +|[H_C, [H_B, U(\Gamma_{\min}^p)^\dagger H_C U(\Gamma_{\min}^p)]]|+\rangle$. The approximate eigenvalue in this case equals $-|b|/\sqrt{2}$.

2. For $l = p+1$ we have that:

$$[\delta_{\mathrm{TS}}]_j = \begin{cases} \frac{\mathrm{sign}(b)}{2} & \text{if } j = l-1, \\ -\frac{\mathrm{sign}(b)}{2} & \text{if } j = l, \\ \frac{1}{\sqrt{2}} & \text{if } j = p+1+l, \\ 0 & \text{otherwise.} \end{cases} \tag{D.21}$$

   with $b = \langle +|U(\Gamma_{\min}^p)^\dagger|[H_C, [H_B, H_C]]U(\Gamma_{\min}^p)|+\rangle$. The approximate eigenvalue in this case equals $-|b|/\sqrt{2}$.

We emphasize that the TS with $l = 1$ and $l = p+1$ where the identity gates are inserted at the edges of the optimized QAOA circuit are special since the bound has prefactor $1/\sqrt{2}$ in contrast to the remaining TS where the bound features a prefactor $1/2$. We also emphasize this by using a constant b rather than $\bar{b}$ used for the "bulk" transition states.

In the next section, we will use the eigenvector approximation Eq. (D.20) to derive a lower bound on the energy gain after each iteration of the QAOA algorithm. However, it is essential that before we obtain a simplified expression for $b = \langle +|[H_C, [H_B, U(\Gamma^p_{\min})^\dagger H_C U(\Gamma^p_{\min})]]|+\rangle$. This simplification leverages the specific forms of the mixing Hamiltonian H_B and the cost Hamiltonian H_C. We use the fact that both states $|+\rangle$ and $H_C|+\rangle$ are eigenvectors of the mixing Hamiltonian H_B with eigenvalues $-N$ and $-N+4$ respectively. This allows us to simplify the expression for $\bar{b}$ and arrive at the equation:

$$b = 8\langle +|\tilde{H}_C H_C|+\rangle, \tag{D.22}$$

where $\tilde{H}_C = U(\Gamma^p_{\min})^\dagger H_C U(\Gamma^p_{\min})$ can be thought as the cost Hamiltonian in the Heisenberg picture. Furthermore, the condition $\partial_{\gamma_1} E(\Gamma^{p+1}_{\text{TS}}) = 2\,\text{Im}\Big\{\langle +|\tilde{H}_C H_C|+\rangle\Big\} = 0$, which arises from the transition state being at a stationary point, is applied. It is important to note that the curvature in this scenario is described by $\lambda_{\text{TS}} = -|b|/\sqrt{2}$, equivalently expressed as:

$$\lambda_{\text{TS}} = -\text{sign}(b) b/\sqrt{2} = -4\sqrt{2}|\langle +|\tilde{H}_C H_C|+\rangle|. \tag{D.23}$$

D.3 Expansion of energy alongside the index-1 direction

Given a local minimum of QAOA$_p$, the transition states construction introduced in [SMKS23] guarantees that the energy has to decrease alongside the index-1 direction. Thus, in this section, we will use the approximate Hessian eigenvector introduced in Eq. (D.20) to provide a lower bound on the energy decrease after optimization of QAOA$_{p+1}$. Out of all possible $2p+1$ transition states available for a given local minima, we focus on the transition state constructed by inserting the zeros in the first layer of the QAOA circuit.

D.3.1 Energy

We begin by simplifying the expression for the QAOA$_{p+1}$ wave function obtained once we deviate from the transition state along the descent direction:

$$|\Gamma^{p+1}_{\text{TS}} + \varepsilon \boldsymbol{\delta}_{\text{TS}}\rangle = U(\Gamma^{p+1}_{\text{TS}} + \varepsilon \boldsymbol{\delta}_{\text{TS}})|+\rangle = \Big(\prod_{l=1}^{p+1} U_B(\beta_l) U_C(\gamma_l)\Big)|+\rangle = U(\Gamma^p_{\min}) U_\varepsilon |+\rangle, \tag{D.24}$$

where

$$U_\varepsilon = e^{-i\varepsilon/2 H_C} e^{i s_b \sqrt{2}\varepsilon/2 H_B} e^{i\varepsilon/2 H_C}, \tag{D.25}$$

and we introduce a short-hand notation:

$$s_b = \text{sign}(b). \tag{D.26}$$

The next step is to Taylor expand Eq. (D.25) around $\varepsilon = 0$. For this, we will make use of the following identity:

$$e^A B e^{-A} = B + [A,B] + \frac{1}{2}[A,[A,B]] + \cdots + \frac{1}{n!}[A,[A,\cdots[A,B]\cdots]]. \tag{D.27}$$

Truncating the above expansion up to the 2nd order, and setting $B = e^{i s_b \varepsilon \sqrt{2} H_B/2}$ and $A = -i\varepsilon H_C/2$ leads to

$$U_\varepsilon|+\rangle = e^{-i\varepsilon\phi}|+\rangle - i\frac{\varepsilon}{2}[H_C, e^{i s_b \varepsilon\sqrt{2} H_B/2}]|+\rangle - \frac{\varepsilon^2}{2^3}[H_C,[H_C, e^{i s_b \varepsilon\sqrt{2} H_B/2}]]|+\rangle, \tag{D.28}$$

where $\phi = s_b\sqrt{2}N/2$ and we used that $H_B|+\rangle = -N|+\rangle$.

To simplify the above expression we need to understand the action of the operator $e^{i\varepsilon \mathrm{sign}(b)\sqrt{2}H_B/2}$ on H_C and H_C^2 operators. For this, it is important to note that $H_C|+\rangle$ is an eigenvector of H_B with eigenvalue $(-N+4)$. With this at hand, we obtain that:

$$e^{is_b\varepsilon\sqrt{2}H_B/2}H_C|+\rangle = e^{is_b\varepsilon(-N+4)\sqrt{2}/2}H_C|+\rangle = e^{-i\varepsilon\phi}e^{i\varepsilon s_b 2\sqrt{2}}H_C|+\rangle. \tag{D.29}$$

The procedure is slightly more involved in the case of H_C^2. This is because H_C^2 is a sum of 4-local, 2-local, and 0-local (constant) Hamiltonian densities. More specifically, the squared cost function Hamiltonian is written as

$$H_C^2 = n_C^2\mathbb{I} + T_2 + T_4, \tag{D.30}$$

where T_k with $k = 2, 4$ is a sum involving $4n_{\mathcal{E}}(\mathcal{G})$ and $n_{\mathcal{E}}(\mathcal{G})(n_{\mathcal{E}}(\mathcal{G}) - 5)$ k-local terms respectively. This, together with the $n_{\mathcal{E}}(\mathcal{G})$ terms contributing to $n_C^2\mathbb{I}$ makes for a total of $n_{\mathcal{E}}(\mathcal{G})^2$ terms. Thus, we obtain

$$\begin{aligned} e^{is_b\varepsilon\sqrt{2}H_B/2}H_C^2|+\rangle &= e^{-i\varepsilon\phi}(T_0\mathbb{I} + e^{i\varepsilon s_b 2\sqrt{2}}T_2 + e^{i\varepsilon s_b 4\sqrt{2}}T_4)|+\rangle \\ &= e^{-i\varepsilon\phi}(H_C^2 + (e^{i\varepsilon s_b 2\sqrt{2}} - 1)T_2 + (e^{i\varepsilon s_b 4\sqrt{2}} - 1)T_4)|+\rangle \\ &= e^{-i\varepsilon\phi}H_C^2|+\rangle + e^{-i\varepsilon\phi}O_\varepsilon|+\rangle, \end{aligned} \tag{D.31}$$

where we defined

$$O_\varepsilon = (e^{i\varepsilon s_b 2\sqrt{2}} - 1)T_2 + (e^{i\varepsilon s_b 4\sqrt{2}} - 1)T_4).$$

Analogously, we obtain:

$$[H_C, [H_C, e^{-i\varepsilon s_b\sqrt{2}H_B/2}]]|+\rangle = 2e^{-i\varepsilon\phi}(1 - e^{i\varepsilon s_b 2\sqrt{2}})H_C^2|+\rangle + e^{-i\varepsilon\phi}O_\varepsilon|+\rangle. \tag{D.32}$$

Putting together Eq. (D.29), Eq. (D.31), and Eq. (D.32) we have a final expression for $U_\varepsilon|+\rangle$ that (up to a global phase) reads

$$U_\varepsilon|+\rangle = |+\rangle - i\frac{\varepsilon}{2}(1 - e^{i\varepsilon s_b 2\sqrt{2}})H_C|+\rangle - \frac{\varepsilon^2}{2^2}(1 - e^{i\varepsilon s_b 2\sqrt{2}})H_C^2|+\rangle - \frac{\varepsilon^2}{2^3}O_\varepsilon|+\rangle. \tag{D.33}$$

We can then use Eq. (D.33) to obtain the expression for the energy when moving along the index-1 direction as a function of ε:

$$\begin{aligned} E(\Gamma_{\mathrm{TS}}^{p+1} + \varepsilon\delta_{\mathrm{TS}}) = \langle+|U_\varepsilon^\dagger U^\dagger(\Gamma_{\min}^p)H_C U(\Gamma_{\min}^p)U_\varepsilon|+\rangle &= E(\Gamma_{\min}^p) - \varepsilon\sin\Big(2\sqrt{2}s_b\varepsilon\Big)\langle+|\tilde{H}_C H_C|+\rangle \\ &+ \frac{\varepsilon^2}{2}\sin\Big(\varepsilon s_b\sqrt{2}\Big)^2\partial_{\gamma_1}^2 E(\Gamma_{\mathrm{TS}}^{p+1}) - \frac{\varepsilon^2}{2}\sin\Big(\varepsilon 2\sqrt{2}s_b\Big)\,\mathrm{Im}\Big\{\langle+|\tilde{H}_C H_C^2|+\rangle\Big\} \\ &- \frac{\varepsilon^2}{4}\,\mathrm{Re}\Big\{\langle+|\tilde{H}_C O_\varepsilon|+\rangle\Big\}, \end{aligned} \tag{D.34}$$

where for ease of notation we introduced $\tilde{O} = U^\dagger(\Gamma)OU(\Gamma)$ for a generic Hermitian operator O. The next step in the calculation is to isolate terms depending on the magnitude of their contribution with circuit depth p (independently of the value of ε). For this, we need to further simplify the expectation value $\langle+|\tilde{H}_C O_\varepsilon|+\rangle$. After performing careful algebraic manipulations, we obtain a convoluted expression for the energy along the index-1 direction, represented as the fourth-order polynomial in the parameter ε. For enhanced readability, we present the terms

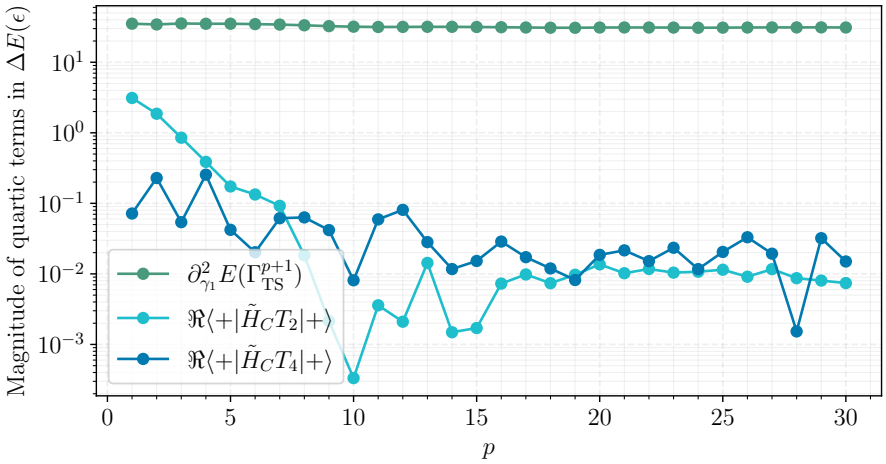

Figure D.6: Magnitude of the prefactors of three different quartic terms $\sim \varepsilon^4$ in the energy expansion along the index-1 direction as a function of the circuit depth p. The first term in the expansion is dominant.

of the expansion of $\Delta E(\varepsilon) = E(\Gamma^{p+1}_{\mathrm{TS}} + \varepsilon\delta_{\mathrm{TS}}) - E(\Gamma^{p}_{\mathrm{min}})$ separately, organized according to the power of ε with which they are associated:

$$\varepsilon^2 \to -\varepsilon \sin\Big(2\sqrt{2}s_b\varepsilon\Big)\langle+|\tilde{H}_C H_C|+\rangle \tag{D.35}$$

$$\varepsilon^3 \to -\frac{\varepsilon^2}{4}\sin\Big(2\sqrt{2}s_b\varepsilon\Big)\,\mathrm{Im}\Big\{\langle+|\tilde{H}_C T_2|+\rangle\Big\} - \frac{\varepsilon^2}{2}\sin\Big(2\sqrt{2}s_b\varepsilon\Big)\Big(1-\frac{\sin\Big(4\sqrt{2}s_b\varepsilon\Big)}{2\sin\Big(2\sqrt{2}s_b\varepsilon\Big)}\Big)\,\mathrm{Im}\Big\{\langle+|\tilde{H}_C T_4|+\rangle\Big\}. \tag{D.36}$$

$$\varepsilon^4 \to \frac{\varepsilon^2}{2}\sin^2(\varepsilon s_b\sqrt{2})\partial^2_{\gamma_1}E(\Gamma^{p+1}_{\mathrm{TS}}) + \frac{\varepsilon^2}{4}2\sin^2(\varepsilon s_b\sqrt{2})\,\mathrm{Re}\Big\{\langle+|\tilde{H}_C T_2|+\rangle\Big\} + \frac{\varepsilon^2}{2}\sin^2(2\varepsilon s_b\sqrt{2})\,\mathrm{Re}\Big\{\langle+|\tilde{H}_C T_4|+\rangle\Big\}. \tag{D.37}$$

From the above equations, we see that at the first non-trivial order in ε the only cubical $\sim \varepsilon^3$ contribution that remains is $\mathrm{Im}\Big\{\langle+|\tilde{H}_C T_2|+\rangle\Big\}$. The quartic $\sim \varepsilon^4$ term is however more involved, and we resort instead to numerically verifying the order of magnitude of each term involved as a function of circuit depth p.

In Figure D.6 we show the circuit depth dependence of three different terms, (1) $\partial^2_{\gamma_1}E(\Gamma^{p+1}_{\mathrm{TS}})$, (2) $\mathrm{Re}\Big\{\langle+|\tilde{H}_C T_2|+\rangle\Big\}$ and (3) $\mathrm{Re}\Big\{\langle+|\tilde{H}_C T_4|+\rangle\Big\}$ for a single MaxCut instance of a 3-regular weighted graph with $N = 14$ vertices (see Fig. D.2 for the details of the instance). The numerical data reveals that the term (1) $\partial^2_{\gamma_1}E(\Gamma^{p+1}_{\mathrm{TS}})$ dominates over terms (2)-(3) at all circuit depths.

Finally, we obtain a close and concise expression for the energy change along the index-1 direction:

$$\Delta E(\varepsilon) \approx -\varepsilon\sin\Big(2\sqrt{2}s_b\varepsilon\Big)\langle+|\tilde{H}_C H_C|+\rangle - \frac{\varepsilon^2}{4}\sin\Big(2\sqrt{2}s_b\varepsilon\Big)\,\mathrm{Im}\Big\{\langle+|\tilde{H}_C T_2|+\rangle\Big\} + \frac{\varepsilon^2}{2}\sin^2(\varepsilon s_b\sqrt{2})\partial^2_{\gamma_1}E(\Gamma^{p+1}_{\mathrm{TS}}). \tag{D.38}$$

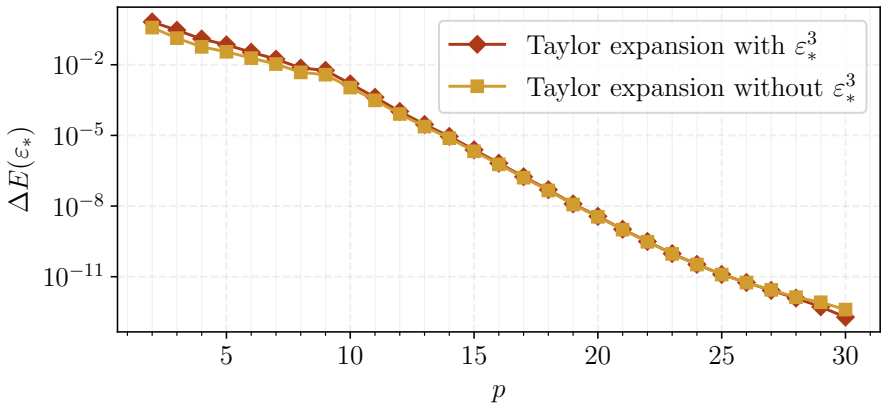

Figure D.7: The energy difference between the transition state and the local minima obtained along the descent direction shows little sensitivity to the presence of the cubic term in the expansion.

It is worth noting that in the above equation, Eq. (D.38) the quadratic term at small ε is negative and proportional to the (approximate) minimum curvature, i.e,

$$\varepsilon^2 \to -\varepsilon^2 2\sqrt{2}\text{sign}(b)b = \varepsilon^2 \frac{\lambda_{\text{TS}}}{2}. \tag{D.39}$$

The negative value of the second order term provided that the fourth order term is positive (see discussion in Sec. 5.4.2 after Eq. (5.17)), leads to an existence of non-trivial minimum in the expansion of the energy along the index-1 direction. Our argument as to why $\partial^2_{\gamma_1} E(\Gamma_{\text{TS}}^{p+1}) > 0$ lies on the observation that it can be approximated as the positive energy difference (see Eq. (5.17)) of the states $\frac{1}{n_C} U H_C |+\rangle$ and $U|+\rangle$.

As discussed in the main text, we discard the cubic $\sim \varepsilon^3$ term for simplicity. This approximation is justified by the negligible effect of the cubic term in the minima of the energy along the index-1 direction, as shown in Fig. D.7. We are then left with an expression for the energy $\Delta E(\varepsilon)_{\text{sym}}$ which has two degenerate global minima

$$\Delta E(\varepsilon^*)_{\text{sym}} = -\frac{\lambda_{\text{TS}}^2}{16\partial^2_{\gamma_1} E(\Gamma_{\text{TS}}^{p+1})}, \tag{D.40}$$

where $(\varepsilon^*)^2 = -\lambda_{\text{TS}}/4\partial^2_{\gamma_1}(\Gamma_{\text{TS}}^{p+1})$.

Bibliography

[Aar17] Scott Aaronson. Shadow Tomography of Quantum States. *arXiv e-prints*, page arXiv:1711.01053, November 2017.

[ACC+21] Andrew Arrasmith, M. Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J. Coles. Effect of barren plateaus on gradient-free optimization. *Quantum*, 5:558, October 2021.

[ACdF89] B. Apolloni, C. Carvalho, and D. de Falco. Quantum stochastic optimization. *Stochastic Processes and their Applications*, 33(2):233–244, 1989.

[AL18] Tameem Albash and Daniel A. Lidar. Adiabatic quantum computation. *Rev. Mod. Phys.*, 90:015002, Jan 2018.

[AR19] Scott Aaronson and Guy N. Rothblum. Gentle Measurement of Quantum States and Differential Privacy. *arXiv e-prints*, page arXiv:1904.08747, April 2019.

[ARCB21] V. Akshay, D. Rabinovich, E. Campos, and J. Biamonte. Parameter concentrations in quantum approximate optimization. , 104(1):L010401, July 2021.

[Aru20] Frank Arute *et al.* Hartree-Fock on a superconducting qubit quantum computer. *Science*, 369(6507):1084–1089, August 2020.

[ASS21] Atithi Acharya, Siddhartha Saha, and Anirvan M. Sengupta. Informationally complete POVM-based shadow tomography. *arXiv e-prints*, page arXiv:2105.05992, May 2021.

[AvDK+07] Dorit Aharonov, Wim van Dam, Julia Kempe, Zeph Landau, Seth Lloyd, and Oded Regev. Adiabatic quantum computation is equivalent to standard quantum computation. *SIAM Journal on Computing*, 37(1):166–194, 2007.

[BBB+21] Lucas T. Brady, Christopher L. Baldwin, Aniruddha Bapat, Yaroslav Kharkov, and Alexey V. Gorshkov. Optimal Protocols in Quantum Annealing and Quantum Approximate Optimization Algorithm Problems. , 126(7):070505, February 2021.

[BBF+18] Fernando G. S. L. Brandão, Michael Broughton, Edward Farhi, Sam Gutmann, and Hartmut Neven. For Fixed Control Parameters the Quantum Approximate Optimization Algorithm's Objective Function Value Concentrates for Typical Instances. *arXiv e-prints*, page arXiv:1812.04170, December 2018.

[BBRA99] J. Brooke, D. Bitko, F. T. Rosenbaum, and G. Aeppli. Quantum annealing of a disordered magnet. *Science*, 284(5415):779–781, 1999.

[BCK+21] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. Noisy intermediate-scale quantum (NISQ) algorithms. *arXiv e-prints*, page arXiv:2101.08448, January 2021.

[BCLK+22] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. Noisy intermediate-scale quantum algorithms. *Rev. Mod. Phys.*, 94:015004, Feb 2022.

[BEKS17] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[Bel97] Richard Bellman. *Introduction to Matrix Analysis (2nd Ed.)*. Society for Industrial and Applied Mathematics, USA, 1997.

[Ben80] Paul Benioff. The computer as a physical system: A microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. *Journal of Statistical Physics*, 22(5):563–591, 1980.

[Ben82] Paul Benioff. Quantum mechanical models of turing machines that dissipate no energy. *Phys. Rev. Lett.*, 48:1581–1585, Jun 1982.

[BFK+13] V. Bapst, L. Foini, F. Krzakala, G. Semerjian, and F. Zamponi. The quantum adiabatic algorithm applied to random optimization problems: The quantum spin glass perspective. *Physics Reports*, 523(3):127–205, 2013. The Quantum Adiabatic Algorithm Applied to Random Optimization Problems: The Quantum Spin Glass Perspective.

[BFM+22] Joao Basso, Edward Farhi, Kunal Marwaha, Benjamin Villalonga, and Leo Zhou. The quantum approximate optimization algorithm at high depth for maxcut on large-girth regular graphs and the sherrington-kirkpatrick model. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

[BGMZ22] Joao Basso, David Gamarnik, Song Mei, and Leo Zhou. Performance and limitations of the qaoa at constant levels on large sparse hypergraphs and spin glass models. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 335–343, 2022.

[BH13] Fernando G. S. L. Brandão and Michał Horodecki. An area law for entanglement from exponential decay of correlations. *Nature Physics*, 9(11):721–726, 2013.

[BK99] Piotr Berman and Marek Karpinski. On some tighter inapproximability results (extended abstract). In Jirí Wiedermann, Peter van Emde Boas, and Mogens Nielsen, editors, *Automata, Languages and Programming*, pages 200–209, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[BK21] Lennart Bittel and Martin Kliesch. Training Variational Quantum Algorithms Is NP-Hard. , 127(12):120502, September 2021.

[BLPS17] C. L. Baldwin, C. R. Laumann, A. Pal, and A. Scardicchio. Clustering of nonergodic eigenstates in quantum spin glasses. *Phys. Rev. Lett.*, 118:127201, Mar 2017.

[BLSF19a] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, November 2019.

[BLSF19b] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, November 2019.

[BM21] Sami Boulebnane and Ashley Montanaro. Predicting parameters for the quantum approximate optimization algorithm for max-cut from the infinite-size limit, 2021.

[BM22] Sami Boulebnane and Ashley Montanaro. Solving boolean satisfiability problems with the quantum approximate optimization algorithm, 2022.

[Bro70] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 03 1970.

[BV93] Ethan Bernstein and Umesh Vazirani. Quantum complexity theory. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '93, page 11–20, New York, NY, USA, 1993. Association for Computing Machinery.

[BVC21] Stefano Barison, Filippo Vicentini, and Giuseppe Carleo. An efficient quantum algorithm for the time evolution of parameterized circuits. *Quantum*, 5:512, July 2021.

[CAB+21] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, August 2021.

[CCHL21] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. *arXiv e-prints*, page arXiv:2111.05881, November 2021.

[CDD01] Nadia Creignou, Hervé Daudé, and Olivier Dubois. Approximating the satisfiability threshold for random k-xor-formulas. *CoRR*, cs.DM/0106001, 2001.

[CLLW16] Richard Cleve, Debbie Leung, Li Liu, and Chunhao Wang. Near-linear constructions of exact unitary 2-designs, 2016.

[CLSS21] Chi-Ning Chou, Peter J. Love, Juspreet Singh Sandhu, and Jonathan Shi. Limitations of Local Quantum Algorithms on Random Max-k-XOR and Beyond. *arXiv e-prints*, page arXiv:2108.06049, August 2021.

[CMNF16] ZhiHua Chen, ZhiHao Ma, Ismail Nikoufar, and Shao-Ming Fei. Sharp continuity bounds for entropy and conditional entropy. *Science China Physics, Mechanics & Astronomy*, 60(2):020321, 2016.

[CPSV20] Ignacio Cirac, David Perez-Garcia, Norbert Schuch, and Frank Verstraete. Matrix Product States and Projected Entangled Pair States: Concepts, Symmetries, and Theorems. *arXiv e-prints*, page arXiv:2011.12127, November 2020.

[Cro18] Gavin E. Crooks. Performance of the Quantum Approximate Optimization Algorithm on the Maximum Cut Problem. *arXiv e-prints*, page arXiv:1811.08419, November 2018.

[CRO+19] Yudong Cao, Jonathan Romero, Jonathan P. Olson, Matthias Degroote, Peter D. Johnson, Mária Kieferová, Ian D. Kivlichan, Tim Menke, Borja Peropadre, Nicolas P. D. Sawaya, Sukin Sim, Libor Veis, and Alán Aspuru-Guzik. Quantum chemistry in the age of quantum computing. *Chemical Reviews*, 119(19):10856–10915, 2019. PMID: 31469277.

[CSV+21] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12:1791, January 2021.

[CYZF21] Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia. Robust shadow estimation. *PRX Quantum*, 2:030348, Sep 2021.

[DBW+19] Alexandre G. R. Day, Marin Bukov, Phillip Weinberg, Pankaj Mehta, and Dries Sels. Glassy phase of optimal quantum control. *Phys. Rev. Lett.*, 122:020601, Jan 2019.

[DBW+21] James Dborin, Fergus Barratt, Vinul Wimalaweera, Lewis Wright, and Andrew G. Green. Matrix Product State Pre-Training for Quantum Machine Learning. *arXiv e-prints*, page arXiv:2106.05742, June 2021.

[DCEL09a] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Phys. Rev. A*, 80:012304, Jul 2009.

[DCEL09b] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Phys. Rev. A*, 80:012304, Jul 2009.

[DMB+23] Alexander M. Dalzell, Sam McArdle, Mario Berta, Przemyslaw Bienias, Chi-Fang Chen, András Gilyén, Connor T. Hann, Michael J. Kastoryano, Emil T. Khabiboulline, Aleksander Kubica, Grant Salton, Samson Wang, and Fernando G. S. L. Brandão. Quantum algorithms: A survey of applications and end-to-end complexities, 2023.

[DOP07] O C O Dahlsten, R Oliveira, and M B Plenio. The emergence of typical entanglement in two-party random processes. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8081–8108, Jun 2007.

[DP85] David Deutsch and Roger Penrose. Quantum theory, the church–turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 400(1818):97–117, 1985.

[DP89] David Elieser Deutsch and Roger Penrose. Quantum computational networks. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 425(1868):73–90, 1989.

[E+22] Sepehr Ebadi et al. Quantum Optimization of Maximum Independent Set using Rydberg Atom Arrays. *Science*, 376:1209, 2022.

[ECP10] J. Eisert, M. Cramer, and M. B. Plenio. Colloquium: Area laws for the entanglement entropy. *Reviews of Modern Physics*, 82(1):277–306, January 2010.

[EG20] Dax Enshan Koh and Sabee Grewal. Classical Shadows with Noise. *arXiv e-prints*, page arXiv:2011.11580, November 2020.

[EH12] Alexander Elgart and George A. Hagedorn. A note on the switching adiabatic theorem. *Journal of Mathematical Physics*, 53(10):102202, 09 2012.

[EHF19] Tim J. Evans, Robin Harper, and Steven T. Flammia. Scalable Bayesian Hamiltonian learning. *arXiv e-prints*, page arXiv:1912.07636, December 2019.

[EKH+20a] Andreas Elben, Richard Kueng, Hsin-Yuan Robert Huang, Rick van Bijnen, Christian Kokail, Marcello Dalmonte, Pasquale Calabrese, Barbara Kraus, John Preskill, Peter Zoller, and Benoît Vermersch. Mixed-State Entanglement from Local Randomized Measurements. , 125(20):200501, November 2020.

[EKH+20b] Andreas Elben, Richard Kueng, Hsin-Yuan (Robert) Huang, Rick van Bijnen, Christian Kokail, Marcello Dalmonte, Pasquale Calabrese, Barbara Kraus, John Preskill, Peter Zoller, and Benoît Vermersch. Mixed-state entanglement from local randomized measurements. *Phys. Rev. Lett.*, 125:200501, Nov 2020.

[EMW21] Daniel J. Egger, Jakub Mareček, and Stefan Woerner. Warm-starting quantum optimization. *Quantum*, 5:479, June 2021.

[Fey82] Richard P. Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6):467–488, 1982.

[FGG14] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A Quantum Approximate Optimization Algorithm. *arXiv e-prints*, page arXiv:1411.4028, November 2014.

[FGG15] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm applied to a bounded occurrence constraint problem, 2015.

[FGG20a] Edward Farhi, David Gamarnik, and Sam Gutmann. The Quantum Approximate Optimization Algorithm Needs to See the Whole Graph: A Typical Case. *arXiv e-prints*, page arXiv:2004.09002, April 2020.

[FGG20b] Edward Farhi, David Gamarnik, and Sam Gutmann. The Quantum Approximate Optimization Algorithm Needs to See the Whole Graph: A Typical Case. *arXiv e-prints*, page arXiv:2004.09002, April 2020.

[FGGN05] E. Farhi, J. Goldstone, S. Gutmann, and Daniel Nagaj. How to make the quantum adiabatic algorithm fail. *International Journal of Quantum Information*, 06:503–516, 2005.

[FGGS00] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, and Michael Sipser. Quantum Computation by Adiabatic Evolution. *arXiv e-prints*, pages quant–ph/0001106, January 2000.

[FGH+12] Edward Farhi, David Gosset, Itay Hen, A. W. Sandvik, Peter Shor, A. P. Young, and Francesco Zamponi. Performance of the quantum adiabatic algorithm on random instances of two optimization problems on regular hypergraphs. *Phys. Rev. A*, 86:052334, Nov 2012.

[FGS+94] A.B. Finnila, M.A. Gomez, C. Sebenik, C. Stenson, and J.D. Doll. Quantum annealing: A new method for minimizing multidimensional functions. *Chemical Physics Letters*, 219(5-6):343–348, Mar 1994.

[FK94] S. K. Foong and S. Kanno. Proof of page's conjecture on the average entropy of a subsystem. *Phys. Rev. Lett.*, 72:1148–1151, Feb 1994.

[FL11] Steven T. Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. *Phys. Rev. Lett.*, 106:230501, Jun 2011.

[Fle70] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 01 1970.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[Fra13] Eduardo Fradkin. *Field Theories of Condensed Matter Physics*. Cambridge University Press, 2 edition, 2013.

[FWS20] Matthew Fishman, Steven R. White, and E. Miles Stoudenmire. The ITensor software library for tensor network calculations, 2020.

[GAE07] D. Gross, K. Audenaert, and J. Eisert. Evenly distributed unitaries: on the structure of unitary designs. *J. Math. Phys.*, 48(5):052104, 22, 2007.

[Gol70] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[Got09] Daniel Gottesman. An introduction to quantum error correction and fault-tolerant quantum computation, 2009.

[Gro96] Lov K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, page 212–219, New York, NY, USA, 1996. Association for Computing Machinery.

[GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, nov 1995.

[GWOB19] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, December 2019.

[GZCW21] Julien Gacon, Christa Zoufal, Giuseppe Carleo, and Stefan Woerner. Simultaneous Perturbation Stochastic Approximation of the Quantum Fisher Information. *Quantum*, 5:567, October 2021.

[Har21a] Matthew P. Harrigan *et al.* Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, 17(3):332–336, January 2021.

[Har21b] Matthew P. Harrigan *et al.* Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, 17(3):332–336, February 2021.

[HBC+21] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean. Quantum advantage in learning from experiments. *arXiv e-prints*, page arXiv:2112.00778, December 2021.

[HBK21] Tobias Haug, Kishor Bharti, and M. S. Kim. Capacity and Quantum Geometry of Parametrized Quantum Circuits. *PRX Quantum*, 2(4):040309, October 2021.

[HCT+19] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. , 567(7747):209–212, March 2019.

[HG19] Yichen Huang and Yingfei Gu. Eigenstate entanglement in the sachdev-ye-kitaev model. *Phys. Rev. D*, 100:041901, Aug 2019.

[HKL+20] Philipp Hauke, Helmut G Katzgraber, Wolfgang Lechner, Hidetoshi Nishimori, and William D Oliver. Perspectives of quantum annealing: methods and implementations. *Reports on Progress in Physics*, 83(5):054401, May 2020.

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, June 2020.

[HKP21] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Efficient estimation of pauli observables by derandomization. *Phys. Rev. Lett.*, 127:030503, Jul 2021.

[HKT+21] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. Provably efficient machine learning for quantum many-body problems. *arXiv e-prints*, page arXiv:2106.12627, June 2021.

[HLZ04] Eran Halperin, Dror Livnat, and Uri Zwick. Max cut in cubic graphs. *Journal of Algorithms*, 53(2):169–185, 2004.

[HP07] Patrick Hayden and John Preskill. Black holes as mirrors: quantum information in random subsystems. *Journal of High Energy Physics*, 2007(09):120–120, Sep 2007.

[HQRY16] Pavan Hosur, Xiao-Liang Qi, Daniel A. Roberts, and Beni Yoshida. Chaos in quantum channels. *Journal of High Energy Physics*, 2016(2), Feb 2016.

[HÅ01] Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, jul 2001.

[HSCC21] Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J. Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *arXiv e-prints*, page arXiv:2101.02138, January 2021.

[JCKK21] Nishant Jain, Brian Coyle, Elham Kashefi, and Niraj Kumar. Graph neural network initialisation of quantum approximate optimisation. *arXiv e-prints*, page arXiv:2111.03016, November 2021.

[JDM+21] Sonika Johri, Shantanu Debnath, Avinash Mocherla, Alexandros Singk, Anupam Prakash, Jungsang Kim, and Iordanis Kerenidis. Nearest centroid classification on a trapped ion quantum computer. *npj Quantum Information*, 7:122, January 2021.

[JG20] Tyson Jones and Julien Gacon. Efficient calculation of gradients in classical simulations of variational quantum algorithms, 2020.

[JKSZ10] Thomas Jörg, Florent Krzakala, Guilhem Semerjian, and Francesco Zamponi. First-order transitions and the performance of quantum algorithms in random optimization problems. *Phys. Rev. Lett.*, 104:207206, May 2010.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980, December 2014.

[KG15] Richard Kueng and David Gross. Qubit stabilizer states are complex projective 3-designs. *arXiv e-prints*, page arXiv:1510.02767, October 2015.

[Kit15] Alexei Kitaev. A simple model of quantum holography. Talks at KITP, April 7, 2015 and May 27, 2015., 2015.

[KMT+17] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. , 549(7671):242–246, September 2017.

[KMvB+19] C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos, and P. Zoller. Self-verifying variational quantum simulation of lattice models. *Nature*, 569(7756):355–360, 2019.

[KN98] Tadashi Kadowaki and Hidetoshi Nishimori. Quantum annealing in the transverse Ising model. *Phys. Rev. E*, 58:5355–5363, Nov 1998.

[KO21a] Joonho Kim and Yaron Oz. Entanglement Diagnostics for Efficient Quantum Computation. *arXiv e-prints*, page arXiv:2102.12534, February 2021.

[KO21b] Joonho Kim and Yaron Oz. Quantum Energy Landscape and VQA Optimization. *arXiv e-prints*, page arXiv:2107.10166, July 2021.

[LCS+21] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo. Diagnosing barren plateaus with tools from quantum optimal control, 2021.

[LDG+21] Sheng-Hsuan Lin, Rohit Dilip, Andrew G. Green, Adam Smith, and Frank Pollmann. Real- and imaginary-time evolution with compressed quantum circuits. *PRX Quantum*, 2:010342, Mar 2021.

[LHA+20] Nathan Lacroix, Christoph Hellings, Christian Kraglund Andersen, Agustin Di Paolo, Ants Remm, Stefania Lazar, Sebastian Krinner, Graham J. Norris, Mihai Gabureac, Johannes Heinsoo, Alexandre Blais, Christopher Eichler, and Andreas Wallraff. Improving the performance of deep quantum optimization algorithms with continuous gate sets. *PRX Quantum*, 1:110304, Oct 2020.

[LJGM+21] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. Theory of overparametrization in quantum neural networks, 2021.

[LLL20] Daniel Liang, Li Li, and Stefan Leichenauer. Investigating quantum approximate optimization algorithms under bang-bang protocols. *Physical Review Research*, 2(3):033402, September 2020.

[Llo96] Seth Lloyd. Universal quantum simulators. *Science*, 273(5278):1073–1078, 1996.

[LLZW20] Xiu-Zhe Luo, Jin-Guo Liu, Pan Zhang, and Lei Wang. Yao.jl: Extensible, Efficient Framework for Quantum Algorithm Design. *Quantum*, 4:341, October 2020.

[LMSS15] C. R. Laumann, R. Moessner, A. Scardicchio, and S. L. Sondhi. Quantum annealing: The fastest route to quantum computation? *The European Physical Journal Special Topics*, 224(1):75–88, 2015.

[LO04] José Ignacio Latorre and Román Orús. Adiabatic quantum computation and quantum phase transitions. *Phys. Rev. A*, 69:062302, Jun 2004.

[LSS08] C. Laumann, A. Scardicchio, and S. L. Sondhi. Cavity method for quantum spin glasses on the Bethe lattice. *Phys. Rev. B*, 78:134424, Oct 2008.

[LvDX12] Wei Li, Jan von Delft, and Tao Xiang. Efficient simulation of infinite tree tensor network states on the Bethe lattice. *Phys. Rev. B*, 86:195137, Nov 2012.

[MBB+18] Nikolaj Moll, Panagiotis Barkoutsos, Lev S Bishop, Jerry M Chow, Andrew Cross, Daniel J Egger, Stefan Filipp, Andreas Fuhrer, Jay M Gambetta, Marc Ganzhorn, Abhinav Kandala, Antonio Mezzacapo, Peter Müller, Walter Riess, Gian Salis, John Smolin, Ivano Tavernelli, and Kristan Temme. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology*, 3(3):030503, June 2018.

[MBS+18] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9:4812, November 2018.

[MBS+22] Antonio Anna Mele, Glen Bigan Mbeng, Giuseppe Ernesto Santoro, Mario Collura, and Pietro Torta. Avoiding barren plateaus via transferability of smooth solutions in Hamiltonian Variational Ansatz. *arXiv e-prints*, page arXiv:2206.01982, June 2022.

[MD19] Joshua Morris and Borivoje Dakić. Selective Quantum State Tomography. *arXiv e-prints*, page arXiv:1909.05880, September 2019.

[Med24] Raimel A. Medina. QAOALandscapes.jl. `https://github.com/RaimelMedina/QAOALandscapes/tree/main`, 2024.

[Mey21] Johannes Jakob Meyer. Fisher Information in Noisy Intermediate-Scale Quantum Applications. *Quantum*, 5:539, September 2021.

[MH21] Kunal Marwaha and Stuart Hadfield. Bounds on approximating Max kXOR with quantum and classical local algorithms. *arXiv e-prints*, page arXiv:2109.10833, September 2021.

[MHS+17] S. C. Morampudi, B. Hsu, S. L. Sondhi, R. Moessner, and C. R. Laumann. Clustering in Hilbert space of a quantum optimization problem. *Phys. Rev. A*, 96:042303, Oct 2017.

[Mi 21] Xiao Mi *et al.* Information Scrambling in Computationally Complex Quantum Circuits. *arXiv e-prints*, page arXiv:2101.08870, January 2021.

[MM09] Marc Mezard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., New York, NY, USA, 2009.

[MNKF18] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. , 98(3):032309, September 2018.

[Mon16] Ashley Montanaro. Quantum algorithms: an overview. *npj Quantum Information*, 2(1):15023, 2016.

[MR18] Patrick K. Mogensen and Asbjørn N. Riseth. Optim: A mathematical optimization package for julia. *Journal of Open Source Software*, 3(24):615, 2018.

[MRBAG16] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, February 2016.

[MRZ02] M. Mezard, F. Ricci-Tersenghi, and R. Zecchina. Alternative solutions to diluted p-spin models and XORSAT problems. *arXiv e-prints*, pages cond–mat/0207140, July 2002.

[MS21] Raimel Medina and Maksym Serbyn. Duality approach to quantum annealing of the 3-variable exclusive-or satisfiability problem (3-XORSAT). *Phys. Rev. A*, 104:062423, Dec 2021.

[MS24] Raimel A. Medina and Maksym Serbyn. A Recursive Lower Bound on the Energy Improvement of the Quantum Approximate Optimization Algorithm. *arXiv*, 2405.10125, May 2024.

[MVS21] Raimel Medina, Romain Vasseur, and Maksym Serbyn. Entanglement transitions from restricted boltzmann machines. *Phys. Rev. B*, 104:104205, Sep 2021.

[MWR+24] Patrick Kofod Mogensen, John Myles White, Asbjørn Nilsen Riseth, Tim Holy, Miles Lubin, Christof Stocker, Andreas Noack, Antoine Levitt, Christoph Ortner, Benoît Legat, Blake Johnson, Christopher Rackauckas, Yichao Yu, Kristoffer Carlsson, Dahua Lin, Arno Strouwen, Josua Grawitter, Takafumi Arakaki, Benoît Pasquier, Thomas R. Covert, Ron Rock, Michael Creel, cossio, Jeffrey Regier, Ben Kuhn, Alexey Stukalov, Alex Williams, and Kenta Sato. Julianlsolvers/optim.jl: v1.9.3+doc1, March 2024.

[NCV+21] Antoine Neven, Jose Carrasco, Vittorio Vitale, Christian Kokail, Andreas Elben, Marcello Dalmonte, Pasquale Calabrese, Peter Zoller, Benot Vermersch, Richard Kueng, and Barbara Kraus. Symmetry-resolved entanglement detection using partial transpose moments. *npj Quantum Information*, 7(1):152, 2021.

[NFG+08] Daniel Nagaj, Edward Farhi, Jeffrey Goldstone, Peter Shor, and Igor Sylvester. Quantum transverse-field Ising model on an infinite tree from matrix product states. *Phys. Rev. B*, 77:214431, Jun 2008.

[Not] In translational invariant systems all subsystems of the same size related by a translation are equivalent. In the present case, we expect the same behavior to be true on average.

[NRVH17] Adam Nahum, Jonathan Ruhman, Sagar Vijay, and Jeongwan Haah. Quantum entanglement growth under random unitary dynamics. *Phys. Rev. X*, 7:031016, Jul 2017.

[NVH18] Adam Nahum, Sagar Vijay, and Jeongwan Haah. Operator spreading in random unitary circuits. *Phys. Rev. X*, 8:021014, Apr 2018.

[ODP07] R. Oliveira, O. C. O. Dahlsten, and M. B. Plenio. Generic entanglement can be generated efficiently. *Physical Review Letters*, 98(13), Mar 2007.

[OGB21] Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Structure optimization for parameterized quantum circuits. *Quantum*, 5:391, January 2021.

[OKW20] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement Induced Barren Plateaus. *arXiv e-prints*, page arXiv:2010.15968, October 2020.

[Pag93] Don N. Page. Average entropy of a subsystem. *Phys. Rev. Lett.*, 71:1291–1294, Aug 1993.

[PK19] Marco Paini and Amir Kalev. An approximate description of quantum states. *arXiv e-prints*, page arXiv:1910.10543, October 2019.

[PMS+14] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1):4213, 2014.

[PNGY21] Taylor L. Patti, Khadijeh Najafi, Xun Gao, and Susanne F. Yelin. Entanglement devised barren plateau mitigation. *Physical Review Research*, 3(3):033090, July 2021.

[Pre18] John Preskill. Quantum Computing in the NISQ era and beyond. *arXiv e-prints*, page arXiv:1801.00862, January 2018.

[PSW96] Boris G. Pittel, Joel Spencer, and Nicholas C. Wormald. Sudden emergence of a giant-core in a random graph. *J. Comb. Theory, Ser. B*, 67(1):111–151, 1996.

[PSW06] Sandu Popescu, Anthony J. Short, and Andreas Winter. Entanglement and the foundations of statistical mechanics. *Nature Physics*, 2(11):754–758, November 2006.

[RBMV21] Aniket Rath, Cyril Branciard, Anna Minguzzi, and Benoît Vermersch. Quantum Fisher Information from Randomized Measurements. , 127(26):260501, December 2021.

[SBG+19] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. , 99(3):032331, March 2019.

[SBSW20] Maria Schuld, Alex Bocharov, Krysta M. Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. , 101(3):032308, March 2020.

[Sch78] Thomas J. Schaefer. The complexity of satisfiability problems. In *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing*, STOC 78, pages 216–226, New York, NY, USA, 1978. Association for Computing Machinery.

[Sch11] Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96 – 192, 2011. January 2011 Special Issue.

[Sha70] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[Sho94] P.W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 124–134, 1994.

[Sho95] Peter W. Shor. Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer. *arXiv e-prints*, pages quant–ph/9508027, August 1995.

[Sho97] Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509, October 1997.

[SHS+24] Shree Hari Sureshbabu, Dylan Herman, Ruslan Shaydulin, Joao Basso, Shouvanik Chakrabarti, Yue Sun, and Marco Pistoia. Parameter Setting in Quantum Approximate Optimization of Weighted Problems. *Quantum*, 8:1231, January 2024.

[SIKC20] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum Natural Gradient. *Quantum*, 4:269, May 2020.

[Sim94] D.R. Simon. On the power of quantum computation. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 116–123, 1994.

[SJ16] Yiğit Subaşıand Christopher Jarzynski. Nonperturbative embedding for highly nonlocal Hamiltonians. *Phys. Rev. A*, 94:012342, Jul 2016.

[SJAG19] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.

[SMKS23] Stefan H. Sack, Raimel A. Medina, Richard Kueng, and Maksym Serbyn. Recursive greedy initialization of the quantum approximate optimization algorithm with guaranteed improvement. *Phys. Rev. A*, 107:062404, Jun 2023.

[SMM+20] Andrea Skolik, Jarrod R. McClean, Masoud Mohseni, Patrick van der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *arXiv e-prints*, page arXiv:2006.14904, June 2020.

[SMM+22] Stefan H. Sack, Raimel A. Medina, Alexios A. Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3:020365, Jun 2022.

[SS08] Yasuhiro Sekino and L Susskind. Fast scramblers. *Journal of High Energy Physics*, 2008(10):065–065, Oct 2008.

[SS21a] Stefan H. Sack and Maksym Serbyn. Quantum annealing initialization of the quantum approximate optimization algorithm. *arXiv e-prints*, page arXiv:2101.05742, January 2021.

[SS21b] Stefan H. Sack and Maksym Serbyn. Quantum annealing initialization of the quantum approximate optimization algorithm. *Quantum*, 5:491, July 2021.

[ST06] Giuseppe E Santoro and Erio Tosatti. Optimization using quantum mechanics: quantum annealing through adiabatic evolution. 39(36):R393–R431, Sep 2006.

[SY93] Subir Sachdev and Jinwu Ye. Gapless spin-fluid ground state in a random quantum heisenberg magnet. *Phys. Rev. Lett.*, 70:3339–3342, May 1993.

[SYS+21] Michael Streif, Sheir Yarkoni, Andrea Skolik, Florian Neukart, and Martin Leib. Beating classical heuristics for the binary paint shop problem with the quantum approximate optimization algorithm. , 104(1):012403, July 2021.

[TB06] Andrew G. Taube and Rodney J. Bartlett. New perspectives on unitary coupled-cluster theory. *International Journal of Quantum Chemistry*, 106(15):3393–3401, 2006.

[UB20] Alexey Uvarov and Jacob Biamonte. On barren plateaus and cost function locality in variational quantum algorithms. *arXiv e-prints*, page arXiv:2011.10530, November 2020.

[VBM+19] Guillaume Verdon, Michael Broughton, Jarrod R. McClean, Kevin J. Sung, Ryan Babbush, Zhang Jiang, Hartmut Neven, and Masoud Mohseni. Learning to learn with quantum neural networks via classical neural networks. *arXiv e-prints*, page arXiv:1907.05415, July 2019.

[VC21] Tyler Volkoff and Patrick J Coles. Large gradients via correlation in random parameterized quantum circuits. *Quantum Science and Technology*, 6(2):025008, jan 2021.

[vDMV01] W. van Dam, M. Mosca, and U. Vazirani. How powerful is adiabatic quantum computation? *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, 2001.

[VdNDVB07] M. Van den Nest, W. Dür, G. Vidal, and H. J. Briegel. Classical simulation versus universality in measurement-based quantum computation. *Phys. Rev. A*, 75:012337, Jan 2007.

[Vid03] Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, 91:147902, Oct 2003.

[vKRPS18] C.W. von Keyserlingk, Tibor Rakovszky, Frank Pollmann, and S.L. Sondhi. Operator hydrodynamics, otocs, and entanglement growth in systems without conservation laws. *Physical Review X*, 8(2), Apr 2018.

[Wal04] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge Molecular Science. Cambridge University Press, 2004.

[Wat18] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

[Web15] Zak Webb. The Clifford group forms a unitary 3-design. *arXiv e-prints*, page arXiv:1510.02769, October 2015.

[WFC+20] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J. Coles. Noise-Induced Barren Plateaus in Variational Quantum Algorithms. *arXiv e-prints*, page arXiv:2007.14384, July 2020.

[WHJR18] Zhihui Wang, Stuart Hadfield, Zhang Jiang, and Eleanor G. Rieffel. Quantum approximate optimization algorithm for MaxCut: A fermionic view. , 97(2):022304, February 2018.

[WL21] Jonathan Wurtz and Peter Love. MaxCut quantum approximate optimization algorithm performance guarantees for $p > 1$. , 103(4):042612, April 2021.

[WL22] Jonathan Wurtz and Peter J. Love. Counterdiabaticity and the quantum approximate optimization algorithm. *Quantum*, 6:635, January 2022.

[WSW24] Jonathan Wurtz, Stefan Sack, and Sheng-Tao Wang. Solving non-native combinatorial optimization problems using hybrid quantum-classical algorithms, 2024.

[WVG+22] Johannes Weidenfeller, Lucia C. Valor, Julien Gacon, Caroline Tornow, Luciano Bello, Stefan Woerner, and Daniel J. Egger. Scaling of the quantum approximate optimization algorithm on superconducting qubit based hardware. *arXiv e-prints*, page arXiv:2202.03459, February 2022.

[WZCK21] Roeland Wiersema, Cunlu Zhou, Juan Felipe Carrasquilla, and Yong Baek Kim. Measurement-induced entanglement phase transitions in variational quantum circuits. *arXiv e-prints*, page arXiv:2111.08035, November 2021.

[YBL20] Jiahao Yao, Marin Bukov, and Lin Lin. Policy gradient based quantum approximate optimization algorithm, 2020.

[ZBM24] Leo Zhou, Joao Basso, and Song Mei. Statistical estimation in the spiked tensor model via the quantum approximate optimization algorithm, 2024.

[Zhu17] Huangjun Zhu. Multiqubit clifford groups are unitary 3-designs. *Phys. Rev. A*, 96:062336, Dec 2017.

[ZTB+22] Linghua Zhu, Ho Lun Tang, George S. Barron, F. A. Calderon-Vargas, Nicholas J. Mayhall, Edwin Barnes, and Sophia E. Economou. An adaptive quantum approximate optimization algorithm for solving combinatorial problems on a quantum computer, 2022.

[ZWC+20] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D. Lukin. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Phys. Rev. X*, 10:021067, Jun 2020.