Improving Variational Quantum Algorithms: Innovative Initialization Techniques and Extensions to Qudit Systems

by

Stefan H. Sack

October, 2023

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. Hryhoriy Polshyn, Chair
Prof. Maksym Serbyn
Prof. Johannes M. Fink
Prof. Richard Kueng

The thesis of Stefan H. Sack, titled *Improving Variational Quantum Algorithms: Innovative Initialization Techniques and Extensions to Qudit Systems*, is approved by:

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

Signature: _____

Committee Member: Prof. Johannes M. Fink, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. Richard Kueng, Johannes Kepler University, Linz, Austria

Signature: _____

Defense Chair: Prof. Hryhoriy Polshyn, 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: _____

Stefan H. Sack
October, 2023

Signed page is on file

Abstract

This Ph.D. thesis presents a detailed investigation into Variational Quantum Algorithms (VQAs), a promising class of quantum algorithms that are well suited for near-term quantum computation due to their moderate hardware requirements and resilience to noise. Our primary focus lies on two particular types of VQAs: the Quantum Approximate Optimization Algorithm (QAOA), used for solving binary optimization problems, and the Variational Quantum Eigensolver (VQE), utilized for finding ground states of quantum many-body systems.

In the first part of the thesis, we examine the issue of effective parameter initialization for the QAOA. The work demonstrates that random initialization of the QAOA often leads to convergence in local minima with sub-optimal performance. To mitigate this issue, we propose an initialization of QAOA parameters based on the Trotterized Quantum Annealing (TQA). We show that TQA initialization leads to the same performance as the best of an exponentially scaling number of random initializations.

The second study introduces Transition States (TS), 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 increasing number of circuit layers.

In the third section, we extend the QAOA to qudit systems, which are higher-dimensional generalizations of qubits. This chapter provides theoretical insights and practical strategies for leveraging the increased computational power of qudits in the context of quantum optimization algorithms and suggests a quantum circuit for implementing the algorithm on an ion trap quantum computer.

Finally, we propose an algorithm to avoid "barren plateaus", regions in parameter space with vanishing gradients that obstruct efficient parameter optimization. This novel approach relies on defining a notion of weak barren plateaus based on the entropies of local reduced density matrices and showcases how these can be efficiently quantified using shadow tomography. To illustrate the approach we employ the strategy in the VQE and show that it allows to successfully avoid barren plateaus in the initialization and throughout the optimization.

Taken together, this thesis greatly enhances our understanding of parameter initialization and optimization in VQAs, expands the scope of QAOA to higher-dimensional quantum systems, and presents a method to address the challenge of barren plateaus using the VQE. These insights are instrumental in advancing the field of near-term quantum computation.

Acknowledgements

I acknowledge support from IBM for the IBM Ph.D. Fellowship 2022 in Quantum Computing for the grant entitled "Quantum Quantum Circuits and Software Variational quantum algorithms on NISQ devices" as well as the European Research Council (ERC) for the grant entitled "Non-Ergodic Quantum Matter: Universality, Dynamics and Control (NEQuM)" grant number 850899. In addition, I thank my supervisor Prof. Serbyn for the freedom and support during the Ph.D. which allowed me to freely pursue my research interests. I thank Prof. Kueng for the fruitful collaborations and guidance. I thank R. A. Medina for being a great officemate, scientific collaborator, and friend. Finally, I would like to thank my parents and grandparents for their continuous support that has allowed me to fully pursue my interests. Without them, this would not have been possible.

About the Author

Stefan completed a B.Sc. in Physics from the University of Technology in Vienna. After that, he moved to Zurich and completed an M.Sc. in Physics at ETH Zurich. This was followed by a research internship at Cambridge Quantum Computing in London. After that Stefan moved back to Austria and joined ISTA in February 2020 as an intern at Prof. Serbyn's group where he became a Ph.D. student in September 2020. His main research interest is near-term quantum algorithms. During his Ph.D. studies, Stefan published three peer-reviewed papers in Quantum, PRXQ, and PRA and presented the work at numerous conferences and invited talks. In 2022 Stefan was awarded the IBM Ph.D. fellowship on quantum computing, this led to an independent collaboration with IBM Research Zurich which also resulted in publication. Lastly, in the summer of 2023, Stefan joined QuEra Computing in Boston for a research internship to work on neutral atom quantum computing.

List of Collaborators and Publications

Stefan H. Sack and Maksym Serbyn. Quantum annealing initialization of the quantum approximate optimization algorithm. *Quantum*, 5:491, July 2021

Stefan H. Sack, Raimel A. Medina, Alexios A. Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding Barren Plateaus Using Classical Shadows. *PRX Quantum*, 3(2):020365, June 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

Table of Contents

Abstract vii

Acknowledgements viii

About the Author ix

List of Collaborators and Publications x

Table of Contents xi

List of Figures xii

List of Algorithms xix

1 Introduction 1
- 1.1 Brief history of quantum computing . . . 1
- 1.2 General motivation for Variational Quantum Algorithms . . . 4
- 1.3 Quantum Approximate Optimization Algorithm . . . 5
- 1.4 Variational Quantum Eigensolver . . . 8
- 1.5 Ansatz Representability and Generalization to Higher Dimensional Systems 9
- 1.6 Parameter Optimization . . . 9
- 1.7 Obstacles for Variational Quantum Algorithms . . . 11

2 Trotterized quantum annealing initialization of the quantum approximate optimization algorithm 13
- 2.1 Introduction . . . 13
- 2.2 Optimization landscape of the QAOA . . . 14
- 2.3 Trotterized quantum annealing as initialization . . . 16
- 2.4 Summary and Discussion . . . 19

3 Recursive greedy initialization of the quantum approximate optimization algorithm with guaranteed improvement 23
- 3.1 Introduction . . . 23
- 3.2 QAOA optimization landscape . . . 24
- 3.3 From transition states to QAOA intialization . . . 26
- 3.4 Summary and Discussion . . . 30

4 Generalization of the quantum approximate optimization algorithm to qudits 31
- 4.1 Introduction . . . 31
- 4.2 Qudits - Beyond Two-Level Systems . . . 32

4.3 Generalization of the QAOA to qudits . . . 33
4.4 Performance of Qudit-QAOA for graph coloring . . . 38
4.5 Summary and Discussion . . . 41

5 Avoiding barren plateaus using classical shadows 43
5.1 Introduction . . . 43
5.2 Avoiding barren plateaus in variational quantum optimization . . . 44
5.3 Weak barren plateaus and initialization of VQE . . . 49
5.4 Entanglement control during optimization . . . 52
5.5 Summary and Discussion . . . 56

6 Summary and Outlook 59
6.1 Summary of thesis content and open research questions . . . 59
6.2 Outlook for the future of quantum computing . . . 60

A Further numerical results for different graph ensembles and discussion on optimal TQA time 63
A.1 Optimization landscape for different graph ensembles . . . 63
A.2 Optimal time for TQA . . . 64
A.3 Patterns in optimized parameters . . . 66
A.4 Random vs TQA initialization for other graph ensembles . . . 67

B Proof of transition state properties and details for greedy algorithm 69
B.1 Restricting QAOA parameter space by symmetries . . . 69
B.2 Construction of transition states . . . 71
B.3 Counting of unique minima . . . 79
B.4 Properties of the index-1 direction . . . 80
B.5 Description of the GREEDY algorithm . . . 81
B.6 Additional graph ensembles and system size scaling . . . 83

C Mathematical details for Qudit-QAOA ansatz and fast numerical simulation of qudit noise 85
C.1 Comparison with previous formulations of the Qudit-QAOA ansatz . . . 85
C.2 Details on representation of Qudit-QAOA in terms of ion trap native gates 86
C.3 Generalized Pauli representation of depolarizing channel . . . 87
C.4 Fast noisy simulation of Qudits-QAOA using the Fast Walsh Hadamard Transform . . . 88
C.5 Cost function expectation value for random sampling . . . 90

D Mathematical details for classical shadows and further numerical results 93
D.1 Classical shadows and implementation details . . . 93
D.2 Unitary t-designs . . . 100
D.3 Entanglement and unitary 2-designs . . . 100
D.4 Entanglement growth and learning rate . . . 103
D.5 Algorithm performance for SYK model . . . 103

Bibliography 107

List of Figures

1.1 High level illustration of a VQA. The Quantum Processing Unit (QPU) prepares the variational wave function and measures the qubits. Wile the Classical Processing Unit (CPU) computes the energy expectation value from the measurement data as well as the update step. Arrows indicate the iterative nature of this process. 5

1.2 Quantum circuit of the QAOA circuit. The parameter p controls the depth of the circuit. The variational parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are updated in an iterative loop with a classical computer to minimize the energy expectation value $\langle H_C \rangle$. Each qubit is associated with one vertex V in the graph G. Sampled bitstrings thus correspond to graph partitions, the MaxCut is the bitstring with lowest corresponding energy. 7

1.3 Illustration of a HEA circuit, R_x, R_y and R_z indicate randomly chosen single-qubit rotation gates, W_l is the entangling layer. 8

1.4 Illustration of an optimization landscape with multiple local minima. The goal of variational algorithms is to find parameter values corresponding to a point with the lowest possible cost function value. The shape of the landscape depends on the problem instance as well as the variational ansatz that is used. $\boldsymbol{\theta}_{\text{init.}}$ indicates initial parameter values, $\boldsymbol{\theta}^{\star}$ converges final parameters values. The two points are connected by dashed lines that indicate a path that an optimization algorithm might take. Figure generated using Midjourney's generative AI. 10

2.1 (a) The circuit that prepares a quantum state in the QAOA is parametrized by a set of $2p$ angles γ_i, β_i. (b) The optimization of $\langle H_C \rangle$ is launched from a certain guess of parameters and state preparation and measurements are iterated until the algorithm converges to a set of optimized angles γ_i^*, β_i^*. (c) The cartoon of the cost function $\langle H_C \rangle$ landscape as a function of variational parameters shows that random initializations are prone to converge to sub-optimal local minima. In contrast, the family of TQA initializations proposed in this work converges to the (nearly) optimal minimum. 15

2.2 Joint probability distribution of distance to the global minimum in parameter space $d_{\vec{\gamma},\vec{\beta}}$ and in terms of approximation ratio $\Delta r_{\vec{\gamma},\vec{\beta}}$ reveals that the most probable outcome of random initialization is a convergence to sub-optimal local minima (yellow region). The orange dot corresponds to average values of $d_{\vec{\gamma},\vec{\beta}}, \Delta r_{\vec{\gamma},\vec{\beta}}$ for random initialization. In contrast, TQA initialization leads to a local minima with a better approximation ratio that occasionally outperforms the best random initialization (red dot, shifted from slightly negative values to $\Delta r_{\vec{\gamma},\vec{\beta}} = 0$ for improved visibility). The data is averaged over 50 random unweighted 3-regular graphs with $N = 12$ vertices and QAOA at level $p = 5$. 16

2.3 Optimal time of TQA evolution T^* increases linearly with number of discretization steps p. Top inset illustrates that optimal performance of TQA at time T^* is followed by the rapid decrease in approximation ratio at longer times T^*. Data is shown for $N = 12$. Bottom inset shows finite size scaling of the time step δt, determined by the slope of the T^* vs p dependence, that assumes approximately constant value with the graph size. All averaging is performed over 50 random instances of unweighted 3-regular graphs. 17

2.4 (a) Approximation ratio of the $p = 5$ QAOA as a function of TQA initialization time T reveals that a range of initialization times $[T^*_{\mathsf{min}}, T^*_{\mathsf{max}}]$ (green triangle and star) yield the performance within 1% of the minimal $1 - r_{\vec{\gamma},\vec{\beta}}$. On the other hand, the study of the distance between TQA initialization and converged value of angles reveals the existence of a time T^*_d where the QAOA performs the smallest parameter updates. (b) All three times T^*_{min}, T^*_{max}, and T^*_d defined in panel (a) increase linearly with QAOA circuit depth p. Moreover, the T^*_d is very close to the time where TQA protocol itself achieves optimal performance, T^*_{TQA}, see Fig. 2.2. Data was obtained for $N = 12$ and averaged over 50 random graphs. 18

2.5 A single optimization run of the QAOA with TQA initialization with time $\delta t\, p$ yields equivalent performance to the best out of 2^p random initializations. System size is $N = 12$. Inset reveals that the comparable performance persists over the entire range of considered system sizes. Averaging was performed over 50 random graphs. 20

3.1 (a) Circuit diagram that implements the QAOA ansatz state with circuit depth p, see Eq. (1.8). Gray boxes indicate the identity gates that are inserted when constructing a TS, as indicated in Theorem 1. (b) Local minima $\mathbf{\Gamma}^p_{\mathrm{min}}$ of QAOA$_p$ generate a TS $\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}$ for QAOA$_{p+1}$ that connects to two *new local minima*, $\mathbf{\Gamma}^{p+1}_{\mathrm{min}_{1,2}}$ with lower energy. 24

3.2 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. 3.1(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, and 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. 27

3.3 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 [SS21b] 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 B.6. 28

3.4 (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. 3.2 for $p = 10$. 29

4.1 Illustration of a graph with six vertices that is colorable with three colors (yellow, red, and blue). In the Qudit-QAOA each qudit is assigned to one vertex in the graph and each qudit state represents one color. Here the initial state $|+\rangle^{\otimes n}$ is an equal superposition of all possible colorings. 32

4.2 Circuit diagram of a Qudit-QAOA with ion trap native gates. The initial state $|+\rangle^{\otimes n}$ is the equal superposition of all qudit levels. Vertical lines indicate the native qudit entangling gates $G(\gamma)$ as defined in Eq. (4.10), they are applied to pairs of qudits $i, j \in E$ to implement $e^{-i\gamma_t C}$ (light blue box). Generalized Hadamard gates H are used to implement a basis transformation into the X-basis (green boxes) where phase shift gates $Z_a(-\beta_a^t)$ are applied consecutively to implement $\prod_{a\neq 0} Z_a(-\beta_a^t)$ (dark red boxes). $H^\dagger$ gates are used to transform back into the computational basis , which completes the implementation of $U_B(\boldsymbol{\beta}_t)^{\otimes n}$ (light red box). This pattern (gray box) is repeated p-times to implement a QAOA of circuit depth p. We omit the angle in phase shift gates and qudit entangling gates in the cartoon for simplicity. 36

4.3 Decision diagram for approximating a single-qudit depolarizing channel as a random choice experiment. State vectors sampled according to the diagram approximate the single-qudit depolarizing channel provided a large enough number of samples are used. 38

4.4 Inset shows the triangular graph used for graph coloring. The graph is three colorable, we use red, blue, and yellow as an example. The coloring is sixfold degenerate since the coloring can be permuted $3!$ times. (a) The overlap of the optimized QAOA-Qudit state with the eigenstates of the Potts model, Eq. (4.4), for a noise-free simulation with circuit depths $p = 1$ (blue) and $p = 2$ (green). We can see that for $p = 1$ the ground state cannot be expressed exactly and higher energy excitations are present. For $p = 2$ an exact superposition of the degenerate ground states is prepared and no higher energy contributions can be observed. (b) Under the presence of noise, this is no longer the case and both $p = 1$ (orange) and $p = 2$ (red) have higher energy contributions that reduce the overlap with the ground state. For the noise, we consider a single qudit depolarization error with $p_{\text{err}} = 0.01$ acting on each qudit after the entangling layer. The result is obtained using the Monte Carlo technique described in Sec. 4.3.6. We use 100 random samples. 39

4.5 Inset shows a 4-regular graph with six vertices that is three colorable illustrated using yellow, blue, and red color. (a) Cost function value during the parameter optimization of the Qudit-QAOA for different circuit depths. (b) Result for a noisy simulation. We use a single qudit depolarizing error $p_{\text{err}} = 0.01$ applied to each qudit after the entangling layer and 100 random samples. 40

4.6 Probability of sampling a valid coloring for different circuit depths p, see Eq. (4.19). We can see that the noiseless simulation achieves a significantly higher value compared to the noisy simulation. The noiseless simulation converges to a probability above 0.9 beyond depth $p = 5$, while the noisy simulation reaches a maximum at $p = 5$ of around 0.5. We use the graph shown in Fig. 4.5 (a). For the noisy simulation, we use 100 random samples and an error probability of $p_{\text{err}} = 0.01$ for a single-qudit depolarizing noise. 41

5.1 (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 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 D.1. 45

5.2 (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 is 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. (5.1), looks similar. 51

5.3 (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. 5.2. Data is averaged over 100 random instances, variance is for the local term $\sigma_1^z\sigma_2^z$. 52

5.4 We numerically illustrate the continuity bound Eq. (5.6) 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 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 is averaged over 500 random instances. 53

5.5 (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 (b), indicated by colored stars, forcing the restart of the optimization with smaller value of η. For $\eta = 0.01$ the algorithm avoids large entanglement region 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. (1.15)) 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 gray dashed line, see the last paragraph of Sec. 5.3.1 for a discussion on the choice of α. Data is averaged over 100 random instances. 55

5.6 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 nonlocal 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 consider (green star). In contrast, SA avoids the WBP for both learning rates. Good performance and further convergence in the local minimum is only achieved through a smaller learning rate of $\eta = 0.01$. We use a system size of $N = 10$ and a random circuit from Eq. (1.15) with circuit depth $p = 100$. Data is averaged over 100 random instances. 57

A.1 Comparing the joint probability distribution of the distance to the global minimum in parameter space $d_{\vec{\gamma},\vec{\beta}}$ and in terms of approximation ratio $\Delta r_{\vec{\gamma},\vec{\beta}}$ for weighted 3-regular (top) and Erdős-Rényi graphs with edge probability 0.5 (bottom) reveals that the distribution is dependent on the initialization interval for weighted 3-regular graphs. We initialize the parameters for $k = 1$ (left) and $k = 2$ (right) and observe that for weighted 3-regular graphs the enlarged interval leads to an increased spread of the local optimas in $\Delta r_{\vec{\gamma},\vec{\beta}}$ (yellow region). The spread in $\Delta r_{\vec{\gamma},\vec{\beta}}$ for Erdős-Rényi graphs remains largely unaffected, as expected from the symmetry considerations. Similarly to Fig. 2.2, red squares correspond to the QAOA minimum achieved from TQA initialization (shifted from small negative values of $\Delta r_{\vec{\gamma},\vec{\beta}}$ to zero for improved visibility), orange dots correspond to the average performance of random initialization. Data is for 50 random graphs with $N = 10$ and $p = 5$. 64

A.2 (Top) Optimal time time step of TQA evolution δt is largely independent of system size and scales qualitatively similar to Eq. (A.1) shown in the bottom panel. . 65

A.3 Converged parameters $\vec{\gamma}^*$ (red) and $\vec{\beta}^*$ (orange) show only slight alterations from the TQA initialization indicated by the green and blue lines respectively. The QAOA optimization modifies parameters at small i, while they remain TQA-like in the rest of the protocol. The results were averaged over 50 random unweighted 3-regular graphs (a), weighted 3-regular graphs (b) and Erdős-Rényi graphs (c), all data is for $p = 10$ and $N = 10$. 66

A.4 TQA initialization leads to the same QAOA performance as the best of 2^p random initializations for both weighted 3-regular graphs (top) and Erdős-Rényi graphs (bottom). We average the results over 50 graph realizations, the main plot was obtained for system size $N = 10$, inset is for circuit depth $p = 10$. 67

B.1 Number of minima found in the initialization graph in Fig. 5.2 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 $0.19\, e^{0.98p}$. 79

B.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. . . . 80

B.3 Flow diagram to visualize the GREEDY QAOA initialization algorithm presented in Algorithm 5. 82

B.4 Performance comparison on (a) RWRG3 and (b) RERG with system size $n = 10$. Data is averaged over 19 non-isomorphic graphs. 83

B.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. 84

C.1 Circuit for $p = 1$ that is used to approximate the single qudit depolarizing channel in our simulations. We use a gFWHT to implement both the noise and unitary gates as vector-vector multiplication. This has both lower time and memory complexity than naive matrix-vector multiplication. 90

D.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. (1.15) with circuit depth $p = 100$. Data is averaged over 100 random instances. . . . 104

List of Algorithms

1 QAOA with TQA initialization 20
2 WBP-free optimization with classical shadows 48
3 QAOA sub-routine 81
4 Grid search sub-routine 81
5 GREEDY QAOA 82
6 Generalized Fast Walsh-Hadamard Transform (gFWHT) 89

CHAPTER 1

Introduction

1.1 Brief history of quantum computing

1.1.1 Schrödinger equation

Since the development of quantum mechanics over 100 years ago, it has greatly impacted the development of many modern-day technological advances. From the transistors in our computers to nuclear fission and LED light technology used to transfer data at light speed through optical fibers all over the world, quantum mechanics is the basic building block of many of our modern-day technologies. From a mathematical perspective, the central equation of quantum mechanics is the Schrödinger equation, named after its creator Erwin Schrödinger. It describes the time evolution of a quantum system and is a partial differential equation that governs the behavior of the wave function $|\psi\rangle$, a mathematical representation of the quantum state of a particle or a system of particles. The Schrödinger equation is formulated in two different forms: the time-dependent Schrödinger equation (TDSE) and the time-independent Schrödinger equation (TISE). The TDSE, given by

$$i\hbar\frac{\partial |\psi\rangle}{\partial t} = H |\psi\rangle , \tag{1.1}$$

and describes the dynamics of the system as it evolves over time, where i is the imaginary unit, $\hbar$ is the reduced Planck constant (which we set to 1 in the remainder of this work), and H is the Hamiltonian operator. The TISE, given by

$$H |\psi\rangle = E |\psi\rangle , \tag{1.2}$$

and is used for stationary states. Solving the Schrödinger equation allows physicists to predict how a quantum system will behave and the probabilities of various outcomes of quantum measurements. The two equations have a remarkably simple form from a mathematical perspective. We can readily write down the Hamiltonian of many interesting systems such as chemistry molecules or materials in condensed matter physics [Tay06, AM76].

Solving the Schrödinger equation analytically has, however, unfortunately only been achieved for a few simple systems, prominent examples are the hydrogen atom or the harmonic oscillator [GS18, CTDL97]. For more complex systems, analytical solutions are often not available, and one has to resort to numerical methods. However, these numerical techniques

face a significant challenge when applied to large systems, mainly due to the exponential scaling of the Hilbert space, which is the vector space of quantum states. The dimension of the Hilbert space grows exponentially with the number of particles in the system, leading to a rapid increase in the required computational resources. This problem, known as the "exponential wall" or "curse of dimensionality", severely limits the applicability of classical numerical methods for solving the Schrödinger equation in complex systems.

1.1.2 Feynman's idea for quantum computing

The challenge of simulating quantum systems efficiently prompted Richard Feynman to propose the concept of quantum computation. In his visionary 1982 paper [Fey82], Feynman argued that since conventional, classical computers struggle to simulate quantum systems efficiently due to the exponential growth of the Hilbert space, a more natural approach to tackle this challenge would be to use a quantum mechanical system, which inherently follows the same rules, to simulate another quantum system. This idea laid the foundation for the development of quantum computers and quantum algorithms.

Feynman's original idea was to build a quantum computer using quantum bits, or qubits, which can exist in a superposition of states, as opposed to classical bits that can only be in one of two states (0 or 1). The unique feature of qubits, along with quantum entanglement and the ability to perform quantum gate operations, would allow quantum computers to process and manipulate information in a fundamentally different way compared to classical computers. This would not only enable the simulation of arbitrary quantum systems beyond the reach of classical computation, but also introduce a new avenue for computation in general.

Feynman's original idea of building a quantum computer using quantum bits, or qubits, sparked the development of the field of quantum algorithms and quantum information science. Over the past few decades, numerous breakthroughs and advancements have taken place, shaping the landscape of quantum computing as we know it today.

1.1.3 Early works on quantum computing

One of the earliest quantum algorithms, the Deutsch-Josza algorithm, was developed by David Deutsch and Richard Jozsa in 1992 [DJ92]. This algorithm marked the beginning of quantum computing, as it demonstrated that quantum computers could solve certain problems with significantly fewer queries than their classical counterparts. The Deutsch-Josza algorithm efficiently determines if a function is constant or balanced, providing an exponential speed-up compared to classical algorithms. While this computational problem is not particularly relevant from a practical perspective, the Deutsch-Josza algorithm gave a first hint at the computational possibilties of quantum algorithms.

Another significant milestone in quantum computing was the discovery of Shor's algorithm by Peter Shor in 1994 [Sho94]. This quantum algorithm is capable of efficiently factoring large numbers, providing an exponential speed-up compared to the best-known classical algorithms. Prime factors are often the key components in many cryptographic schemes, particularly in public key cryptography. The classical hardness of prime factorization, or the difficulty of factoring large numbers into their prime components, plays a vital role in the security of these encryption methods. The most well-known example is the RSA cryptosystem, which relies on the product of two large prime numbers as its encryption key.

The security of RSA and many other encryption schemes hinges on the fact that it is computationally infeasible for an attacker to derive the prime factors of the public key, even when the key itself is known. Classically, the best algorithms for factoring large numbers grow exponentially with the number of digits, making it practically impossible to break encryption keys that are sufficiently large.

Following Shor's breakthrough, several other quantum algorithms were developed, such as Grover's algorithm [Gro96], which provides a quadratic speed-up for searching unsorted databases. Grover's algorithm further highlighted the potential of quantum computing by offering a faster solution to a problem with broad applicability.

Since those early days of quantum computing, great progress has been made in the development of quantum algorithms. A large number of quantum algorithms have been developed, addressing a diverse range of problems, from optimization and simulation to cryptography and machine learning. The abundance of quantum algorithms has even led to the creation of the "Quantum Algorithm Zoo", a comprehensive repository that catalogues these algorithms [Jor23].

These quantum algorithms, which are known to provide significant speed-ups over classical algorithms, typically require fault-tolerant quantum computation. Fault-tolerant quantum computing is a method of performing quantum computation that can effectively deal with noise and errors that inevitably arise in realistic quantum systems. The basic idea is to ensure that the computation remains accurate even in the presence of errors, thus enabling the implementation of reliable and large-scale quantum algorithms.

1.1.4 Fault tolerant quantum computation

A crucial component of fault-tolerant quantum computing is quantum error correction (QEC). QEC codes are designed to protect quantum information from decoherence and noise by encoding the quantum information in a larger Hilbert space, allowing errors to be detected and corrected without destroying the information [NC00, Got97, Sho95]. Many different error-correcting codes have been proposed, such as the surface code [BK98] and the toric code [Kit03], each with their own advantages and trade-offs.

Resource estimates for implementing popular fault-tolerant quantum algorithms have been studied to understand the requirements for building practical, large-scale quantum computers. For example, Shor's algorithm has been analyzed in terms of resource requirements, with estimates suggesting that millions of qubits and a large number of gates are needed [FMMC12, GLF19]. Similarly, resource estimates have been provided for other quantum algorithms, such as Grover's algorithm and quantum simulations, which have been investigated in the context of quantum chemistry and condensed matter physics [AGDLHG05, Kit02, BKWS20]. The resource requirements for these algorithms vary widely, depending on the specific problem being solved and the desired accuracy of the results.

The realization of a practical quantum computer capable of running these fault-tolerant quantum algorithms is still an ongoing challenge. Many experimental platforms for quantum computing have been developed, such as superconducting qubits [DS13], trapped ions [BR12], photonic qubits [KMN07], and neutral atoms [BSK$^+$17]. These diverse platforms offer unique advantages and challenges in the pursuit of building a scalable quantum computer. Each of these platforms has its own unique set of advantages and challenges in terms of scalability, coherence times, and gate fidelity. While significant progress has been made in recent years, building large-scale fault-tolerant quantum computers remains a major goal in the field.

1.1.5 Current state of quantum computing

In October 2019, Google's quantum computing team claimed to have achieved quantum supremacy using a 53-qubit superconducting processor [AAB+19]. This milestone demonstrated the potential of quantum computing, although it is important to note that quantum supremacy was achieved for a specific task with limited practical applications. In particular, they used the quantum computer to sample random bitstrings from a random quantum circuit that is believed to be hard to simulate classically due to its large non-trivial entanglement [BIS+18].

The current generation of quantum computers is often referred to as Noisy Intermediate-Scale Quantum (NISQ) devices, a term coined by John Preskill [Pre18]. NISQ devices are characterized by having a modest number of qubits, typically ranging from tens to a few hundred qubits, and relatively high error rates in comparison to the fault-tolerant quantum computers that researchers ultimately aim to build.

The limitation of NISQ devices has let to the development of so-called NISQ algorithms, which are quantum algorithms specifically developed to work around the limitation of the current generation of quantum computers.

1.2 General motivation for Variational Quantum Algorithms

On NISQ hardware, two qubit gates have an error rate of $\sim 1\%$ while the errors for single qubit gates are $\sim 0.1\%$ [LMR+17]. This severely limits the number of gates that we can coherently apply on current quantum hardware. One of the most popular schemes to work around this limitation is the framework of variational quantum algorithms (VQAs) [PMS+14a, MRBAG16]. They use the quantum computer to implement a variational wave function with only a limited number of gates

$$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\,|0\rangle^{\otimes n} = \prod_{j=1}^{p} U(\theta_j)\,|0\rangle^{\otimes n}\,, \tag{1.3}$$

which typically consists of a repeating pattern of shallow variational unitaries $U(\theta_j)$ applied to an initial state, typically $|0\rangle^{\otimes n}$. The term shallow refers to the fact that the unitary can be implemented with a small (polynomial scaling in system size n) number of gates on a quantum computer. The parameter p, called circuit depth, controls how many times the unitary is applied. The circuit depth p can thus be set such that noise does not corrupt the computation too much. The smaller the circuit depth, the smaller the number of errors that can occur. A more shallow circuit is however also typically associated with a less expressive ansatz, so that these two aspects thus have to be balanced [MEAG+20].

Let us consider a Hamiltonian operator H, which describes the energy of a quantum system. The ground-state energy E_{GS} is the lowest eigenvalue of H, and the corresponding eigenstate $|\psi_{\mathsf{GS}}\rangle$ is the ground-state wavefunction,

$$H\,|\psi_{\mathsf{GS}}\rangle = E_{\mathsf{GS}}\,|\psi_{\mathsf{GS}}\rangle\,, \tag{1.4}$$

using the variational principle from standard quantum mechanics [Dir58], we know that

$$\langle\psi_{\mathsf{trial}}|\,H\,|\psi_{\mathsf{trial}}\rangle \geq E_{\mathsf{GS}}, \tag{1.5}$$

meaning that for any trial wave function $|\psi_{\mathsf{trial}}\rangle$ the energy expectation value can only be greater than the ground state energy.

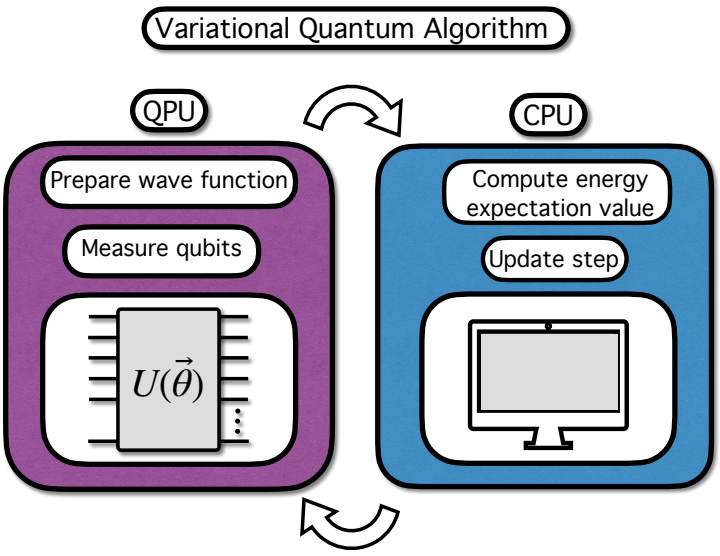


Figure 1.1: High level illustration of a VQA. The Quantum Processing Unit (QPU) prepares the variational wave function and measures the qubits. Wile the Classical Processing Unit (CPU) computes the energy expectation value from the measurement data as well as the update step. Arrows indicate the iterative nature of this process.

VQAs exploit this principle by using a hybrid quantum-classical approach to optimize a parametrized quantum circuit, called an ansatz, to approximate the ground-state energy and the corresponding quantum state of a given system. The ansatz is a trial wavefunction $|\psi_{\text{trial}}\rangle = |\psi(\boldsymbol{\theta})\rangle$ that depends on a set of classical parameters $\boldsymbol{\theta}$. By varying these parameters and minimizing the expectation value of the energy, the algorithm iteratively refines the ansatz to better approximate the true ground-state, see Fig. 1.1 for an illustration.

This idea has been applied in a variety of setting, ranging from using quantum computers for machine learning to variationally implementing a time evolution [HCT+19a, CLB21, GS19]. However the two most prominent settings, that we will focus on in this work, are solving classical optimization problems and finding ground states of chemistry Hamiltonians.

1.3 Quantum Approximate Optimization Algorithm

The Quantum Approximate Optimization Algorithm (QAOA) was first introduced by Farhi et al. [FGG14a] in 2014 as a near-term quantum algorithm for approximately solving classical combinatorial optimization problems. In particular, they introduced it for the graph partitioning problem, known as MaxCut. There, the task is to partition a graph $G = \{E, V\}$, consisting of edges E and vertices V, into two groups such that the partition cuts through a maximum number of edges. The problem is known to be NP-hard which implies that there is no known algorithm that can exactly solve the problem in polynomial time complexity (in system size). However, there is a number of approximate algorithms with polynomial time complexity. Mathematically we can frame the MaxCut problem as a minimization problem for the cost function

$$C = \sum_{i,j \in E} x_i x_j, \tag{1.6}$$

where x_i and x_j are binary variables. If two edges have the same value we get a contribution of 1 in the sum, otherwise it is 0. The cost function can be mapped to a diagonal cost

Hamiltonian by promoting the binary variable x_i to a quantum operator, $x_i \longrightarrow 2\sigma_i^z + \mathbb{I}$, this results in

$$H_C = \sum_{i,j \in E} \sigma_i^z \sigma_j^z, \tag{1.7}$$

where we dropped constant energy shifts and irrelevant factors, σ_i^z is the standard Pauli-Z matrix. Finding the MaxCut is thus equivalent to preparing the ground state of the diagonal Hamiltonian H_C. [FGG14a] propose to use the following variational ansatz state [see Fig. 1.2]

$$|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle = \prod_{j=1}^{p} e^{-i\beta_j H_B} e^{-i\gamma_j H_C} |+\rangle^{\otimes n}, \tag{1.8}$$

to variationally prepare the ground state. Here H_B is called the mixing Hamiltonian which is defined as

$$H_B = -\sum_i \sigma_i^x, \tag{1.9}$$

and the $|+\rangle^{\otimes n}$ state is its ground state, p is the circuit depth. Since H_C and H_B do not commute, the mixing term allows to change the energy, i.e. it allows for transitions between eigenstates of H_C. The idea of the QAOA is thus to first prepare an equal superposition of all possible graph partitions, i.e. $|+\rangle^{\otimes n}$, and evolve the state to the ground state of H_C using the unitaries $e^{-i\gamma_i H_C}$ and $e^{-i\beta_i H_B}$. The parameters γ_i and β_i are classical parameters that control the time for which the unitaries are applied. The goal is to find parameters that minimize the expectation value of the cost Hamiltonian with respect to the QAOA ansatz state

$$(\boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\gamma})} \langle \boldsymbol{\beta}, \boldsymbol{\gamma}| H_C |\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle . \tag{1.10}$$

Typically, the optimal parameters are found in a loop with a classical computer, where the quantum computer is used to implement the QAOA state and measure the expectation value, while the classical computer is used to store and update the variational parameters. Starting from some initial point, the parameters are iteratively updated in order to minimize the energy expectation value. This procedure is repeated until convergence, which implies that a local minima was found. Then, the QAOA state $|\boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star\rangle$ is measured in the computational basis, the obtained post-measurement states are bitstrings that encode a graph partition. The bitstring with the lowest associated cost value is then the approximate MaxCut solution. We illustrate this procedure in Fig. 1.2.

The performance of the QAOA is typically reported as a so-called approximation ratio

$$r_{\boldsymbol{\beta},\boldsymbol{\gamma}} = \frac{\langle \boldsymbol{\beta}, \boldsymbol{\gamma}| H_C |\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle}{C_{\min}}, \tag{1.11}$$

where $C_{\min}$ is the minimal cost function value for the MaxCut. For classical optimization algorithms the Goemans-Williamson (GW) algorithm provides a performance guarantee of 0.87856 for all graphs [GW95]. Unfortunately, due to the heuristic and variational nature of the algorithm, there are little known performance guarantees for the QAOA. In the original work [FGG14a] were able to prove a performance guarantee for $p = 1$ of 0.6924, subsequently [WL21] extended the result for up to $p = 3$ (under certain conditions on the graph). A general performance guarantee for all p has so far been out of reach. In numerical simulations, the QAOA has been reported to outperform the GW algorithm at circuit depths beyond around $p = 9$ [Cro18, ZWC+18]. This result is however purely heuristic and it remains to be shown if it holds beyond the system sizes that can be simulated classically. These promising numerical

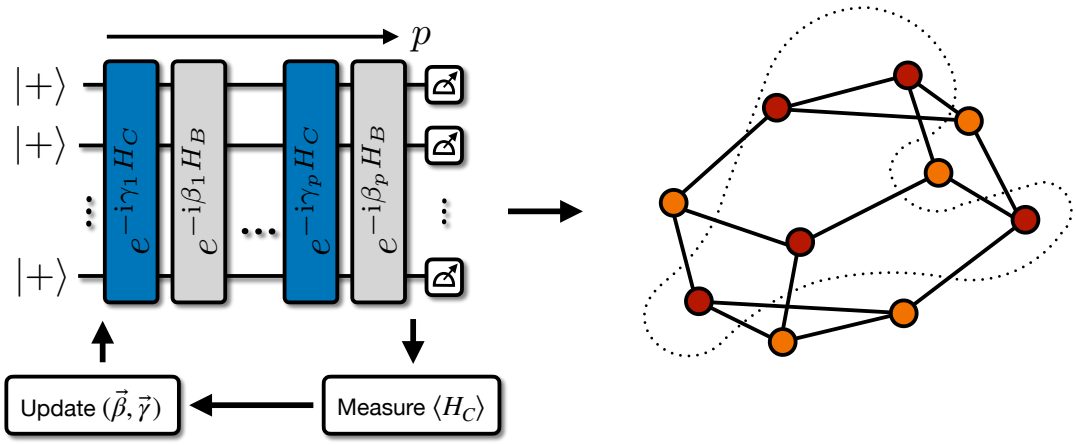


Figure 1.2: Quantum circuit of the QAOA circuit. The parameter p controls the depth of the circuit. The variational parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are updated in an iterative loop with a classical computer to minimize the energy expectation value $\langle H_C \rangle$. Each qubit is associated with one vertex V in the graph G. Sampled bitstrings thus correspond to graph partitions, the MaxCut is the bitstring with lowest corresponding energy.

results however give hope that the QAOA might be a promising algorithm for achieving a quantum advantage on real quantum hardware in the near term.

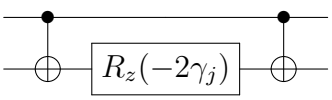
The QAOA is particularly well suited for near-term quantum computing since the depth of the quantum circuit can easily controlled by the circuit depth p such that the computation is not fully corrupted by noise. In addition the unitaries have simple representation in terms of gates that are easily implemented on both superconducting and ion trap hardware. In particular for the quantum term we have that

$$e^{-i\beta_j H_B} = e^{-i\beta_j \sum_i \sigma_i x} = \prod_i e^{-i\beta_j \sigma_i^x} = R_x(-2\beta_j)^{\otimes n}, \tag{1.12}$$

where R_x is a single-qubit rotation gate around the x-axis. This is a standard gate in all quantum hardware. For the classical term we have that

$$e^{-i\gamma_j H_C} = e^{-i\gamma_j \sum_{k,l\in E} \sigma_k^z \sigma_j^z} = \prod_{k,l\in E} e^{-i\gamma_j \sigma_k^z \sigma_l^z}. \tag{1.13}$$

On superconducting hardware, for example, the term $e^{-i\gamma_j \sigma_k^z \sigma_l^z}$ can be implemented as



(1.14)

using two CNOT gates and one R_z rotational gate.

The QAOA has been successfully implemented on superconducting hardware as well as ion trap quantum computers [A+20b, PBB+20], however on both platforms an implementation beyond around 20 qubits has so far been out of reach and circuit depth was limited. Notably, in a recent work a neutral atom quantum processor was used to implement the algorithm on 289 qubits [EKC+22] for a hardware-native graphs, which is by far the largest implementation of the algorithm so far.

1.4 Variational Quantum Eigensolver

The second widely studied Variational Quantum Algorithm (VQA), that we will discuss in this work, is the Variational Quantum Eigensolver (VQE) [PMS+14b]. Introduced by Peruzzo et al. [PMS+14b] in 2014, the primary aim of VQE is to approximate the ground state $|\psi_{\text{GS}}\rangle$ of a Hamiltonian H with a variational wave function $|\psi(\boldsymbol{\theta})\rangle$. Unlike the QAOA, the Hamiltonian in VQE is typically non-diagonal. A quantum computer prepares a variational function using a set of unitary gates, $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\,|\psi_0\rangle$, where $|\psi_0\rangle$ is the initial state, typically assumed to be a product state. The variational parameters are then iteratively updated to minimize the expectation value of the Hamiltonian, also referred to as the cost function $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,H\,|\psi(\boldsymbol{\theta})\rangle$.

There is a large number of different variational ansätze that have been proposed to perform this task, with varying levels of complexity and computational efficiency. One example is the Unitary Coupled Cluster (UCC) ansatz, which builds upon the classical coupled-cluster method by using a unitary version of the cluster operators [RBM+17]. The UCC ansatz has been demonstrated to provide accurate results for molecular systems [BNR+18]. Another example, and the one we will be focusing on in this work, is the Hardware-Efficient Ansatz (HEA) [KMT+17a]. This ansatz leverages the native gate set of a quantum processor to construct a variational wave function, reducing the overall circuit depth and error rate. The HEA is given by the following unitary

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{p} W_l \left(\prod_{i=1}^{N} R_l^i(\theta_l^i) \right), \tag{1.15}$$

where $\theta_l^i \in [-\pi, \pi)$ are pN variational angles, concisely denoted as $\boldsymbol{\theta}$. We will be using this variational circuit in this work. 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 $G_{l,i} \in \{\sigma^x, \sigma^y, \sigma^y\}$. W_l is an entangling layer that consists of two qubit entangling gates, they are typically either CNOT or CZ gates. Often, the two qubit gates are arranged such that they follow the connectivity of the underlying hardware (giving rise to the name). Since the emergence of ion trap quantum computers, which offer all-to-all connectivity, this has become less of a requirement.

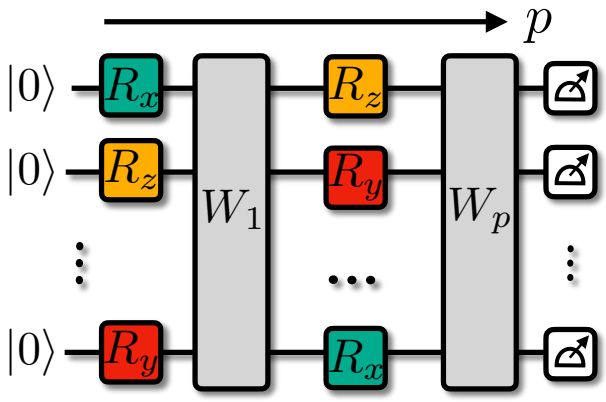


Figure 1.3: Illustration of a HEA circuit, R_x, R_y and R_z indicate randomly chosen single-qubit rotation gates, W_l is the entangling layer.

The VQE has been demonstrated on current hardware using different quantum platforms. Peruzzo et al. [PMS+14b] used a photonic quantum processor to estimate the ground-state energy of hydrogen (H_2) at different internuclear distances. O'Malley et al. [OBK+16] employed

superconducting qubits for simulating the hydrogen molecule (H_2) and the helium hydride cation (HeH^+), achieving energy estimates with chemical accuracy. Furthermore, Shen et al. [SZZ+17] implemented the VQE for simulating the dissociation profile of the hydrogen molecule (H_2) using a trapped ion quantum processor with energy estimates within chemical accuracy. In recent years, the VQE has been extended to solve more complex problems, such as electronic structure calculations of larger molecules [AAA+20] and simulations of strongly correlated systems [CRO+21].

Whilst these experiments are promising, the VQE is still far away from outperforming any sophisticated classical simulation, in addition, the algorithm is even more heuristic in nature than the QAOA and offers no known analytical performance guarantees. This is due to the fact that there exists a plethora of different variational ansätze as well as different problem Hamiltonians. Each ansatz and problem Hamiltonian can lead to a vastly different optimization landscape.

1.5 Ansatz Representability and Generalization to Higher Dimensional Systems

For VQAs the choice of the quantum circuit, or ansatz, is crucial. The ansatz allows exploration of the Hilbert space states explored during the optimization process. The ansatz representability, i.e. what quantum states it can effectively prepare, is thus a crucial question.

Numerous studies have ventured to tackle the problem of ansatz design, with the aim of understanding the conditions that allow an ansatz to efficiently represent the solution to a given problem. These investigations have shown that the suitability of an ansatz can depend on a multitude of factors, including the problem instance, the gate set, and qubit connectivity [HMM+20, SJAG19]. Despite these advancements, the realm of ansatz expressibility remains an active and open field of research.

While the central focus of this work is on the efficient optimization of variational parameters, the subject of ansatz representability remains significant. A noteworthy avenue for potentially expanding the representability of variational ansatz is the utilization of higher-dimensional quantum systems, termed as qudits, where d local Hilbert space dimension of the system. Qudits offer several advantages over qubits, including more efficient quantum error correction, improved fault-tolerant quantum computation, and the potential for more compact quantum circuits [Zhu17b, NLR20]. Accordingly, we will touch upon this active field of research in Chapter 4 of this work by discussing a generalization of the QAOA to qudit systems.

1.6 Parameter Optimization

In VQAs, once the circuit ansatz is chosen, the central aim is to find optimal parameters such that

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}), \tag{1.16}$$

where $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle$ is the cost function or energy expectation value. In practice this is carried out by an optimization algorithm, starting from some initial value $\boldsymbol{\theta}_{\text{init.}}$, that iteratively updates the parameters until convergence to a local minimum [MRBAG16], see Fig. 1.4 for an illustration. In general, there are two types of optimization algorithms: gradient-free and gradient based optimization algorithms.

1.6.1 Gradient-free optimization algorithms

In gradient-free algorithms, the cost function is evaluated at a number of points in the optimization landscape to determine the optimal direction for descent. Prominent examples for this are Nelder-Mead and COBYLA [NM65, Pow94]. Both construct a so-called simplex, which is a polytope in the parameter space, to approximate the local landscape of the cost function. These methods adjust the simplex based on the function values at its vertices to iteratively move towards a local minimum. The main advantage of gradient-free methods is that they do not require the computation of gradients, making them suitable for problems where the gradients are difficult or expensive to calculate. However, these methods can be less efficient than gradient-based methods, especially for high-dimensional problems.

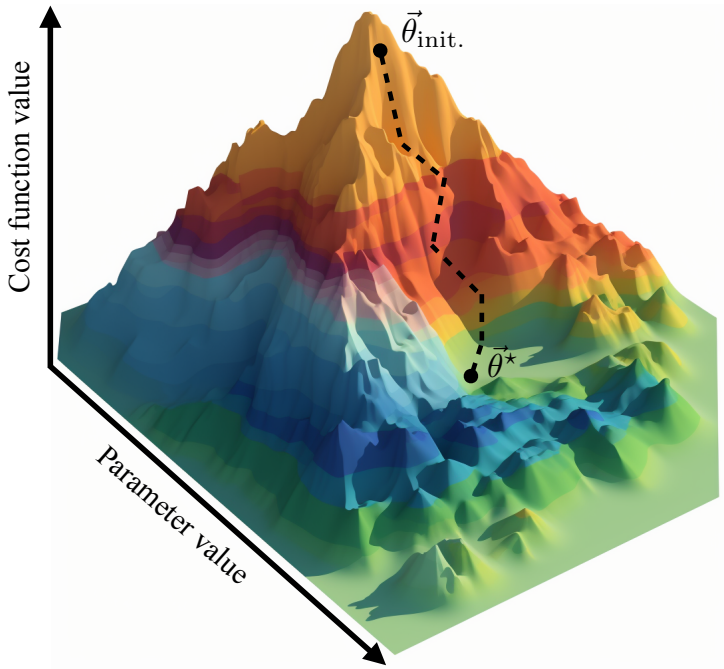


Figure 1.4: Illustration of an optimization landscape with multiple local minima. The goal of variational algorithms is to find parameter values corresponding to a point with the lowest possible cost function value. The shape of the landscape depends on the problem instance as well as the variational ansatz that is used. $\boldsymbol{\theta}_{\text{init.}}$ indicates initial parameter values, $\boldsymbol{\theta}^{\star}$ converges final parameters values. The two points are connected by dashed lines that indicate a path that an optimization algorithm might take. Figure generated using Midjourney's generative AI.

1.6.2 Gradient-based Optimization Algorithms

Gradient-based optimization algorithms, on the other hand, leverage the gradient of the cost function to guide the search for the optimal parameters. Examples of gradient-based methods include gradient descent,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \eta \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}), \tag{1.17}$$

where η is the learning rate which controls the step size. Other examples include conjugate gradient, and the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [NW06]. These methods typically provide faster convergence compared to gradient-free methods, but they require the computation of gradients, which can be challenging in some cases.

There are two primary methods for gradient evaluation on a quantum computer: finite-difference techniques and the parameter-shift rule. Finite-difference methods involve perturbing the variational parameters and computing the differences, given by:

$$\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_j} \approx \frac{E(\boldsymbol{\theta} + s\boldsymbol{e}_j) - E(\boldsymbol{\theta} - s\boldsymbol{e}_j)}{2s}, \tag{1.18}$$

where s is a small perturbation, and $\boldsymbol{e}_j$ is the unit vector in the direction of the j-th parameter. In contrast, the parameter-shift rule leverages the periodic structure of variational single-qubit gates, such as the rotation gate $R(\theta) = e^{-i\frac{\theta}{2}G}$, where θ is the angle of rotation and G is the generator of the rotation, typically represented by Pauli matrices. This periodicity enables the exact computation of the gradient with a constant number of additional circuit evaluations:

$$\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_j} = \frac{1}{2}\left[E(\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_j) - E(\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_j)\right]. \tag{1.19}$$

The parameter-shift rule has been shown to be particularly effective for optimizing variational circuits in quantum machine learning and quantum chemistry applications. It enables more efficient and accurate optimization of quantum circuits compared to finite-difference methods.

1.7 Obstacles for Variational Quantum Algorithms

While there have been numerous successful implementations of VQAs, they have so far they have not been able to outperform any classical algorithms. Generally, VQAs do not offer any rigorous performance guarantee, unlike typical fault-tolerant algorithms. This is primarily due to the lack of information about the high-dimensional energy landscape, or optimization landscape, in which one attempts to find a high-quality minimum. The main obstacles for VQAs are:

(I.) Finding good parameter initializations,

(II.) Enhancing ansatz expressibilty, and

(III.) Avoiding flat regions in parameter space, so-called "barren plateaus".

The variational landscape is often characterized by a large number of local minima, where the number of local minima usually scales exponentially with the parameter dimension. Good algorithm performance thus hinges on a suitable parameter initialization to avoid convergence in one of the many poor-quality local minima. In addition, the optimization landscape can have large flat regions with vanishing gradients, so-called barren plateaus which can be encountered both in the parameter initialization and during parameter optimization. The encounter of barren plateaus thus completely prevents parameter optimization. Both the issue of local minima, as well as barren plateaus, pose significant challenges for the successful implementation of VQAs and a potential quantum advantage. In this work, we present the substantial progress that we made on these issues that have led to a significant improvement as well as a better understanding of the capabilities and limitations of VQAs. Lastly, there is circuit ansatz expressibility which determines if the ansatz is in principle capable of expressing the solution to the problem or not.

1.7.1 Contents of Chapter 2

In Chapter 2 we will address obstacle (I.) and present effective strategies for initializing the QAOA. By carefully selecting suitable initial parameters, the algorithm's performance can be significantly improved. In particular, we present a simple strategy that allows to achieve the same performance as the best out of an exponentially scaling number of random initializations. The presented algorithm requires few computational resources compared to other currently existing techniques and thus makes it highly appealing to be used as a standard initialization technique.

1.7.2 Contents of Chapter 3

In Chapter 3 we present an extension to the results obtained in Chaper 2 where we introduce a novel tool from energy landscape theory to systematically study the emergence of local minima in the energy landscape. This leads us to propose an optimization strategy with guaranteed performance improvement that successfully avoids poor-quality local minima. This is the first analytical study of the QAOA landscape in the regime of large circuit depths, while previous studies were purely limited to numerical observations. We believe that this work may be extended in the future to potentially answer the question if the QAOA could outperform classical methods under certain conditions, which has only been out of reach from an analytical perspective.

1.7.3 Contents of Chapter 4

In Chapter 4 of this work we consider obstacle (II.) and present a generalization of the QAOA to qudit systems. Furthermore, we propose a representation of the quantum circuit in terms of native gates for ion trap quantum computers. This chapter contains the theory required for experimental implementation of the algorithm. In particular, we will discuss the effect of noise on the performance of the algorithm and realistic settings for a future experiment. A successful implementation of the algorithm on qudits has the potential to become the first implementation of a quantum algorithm on a qudit system. This could open up a new exciting avenue for quantum computation in the NISQ era and beyond.

1.7.4 Contents of Chapter 5

In Chapter 5 we address obstacle (III.) and resolve the issue of barren plateaus by employing a novel approach based on entanglement entropy applied to the VQE. Furthermore, we suggest to use classical shadow tomography, a recently introduced scheme for efficient partial state tomography, to efficiently estimate the entanglement entropy. This approach allows us to identify and circumvent regions in the optimization landscape that exhibit vanishing gradients, thereby improving the convergence properties of VQAs. In particular, our work establishes the physical connection between barren plateaus and typical entanglement entropy thus providing a new point of view on the problem.

1.7.5 Contents of Chapter 6

Finally, in Chapter 6 we summarize our results, comment on the open questions and future directions as well as a broader outlook for quantum computing into the future.

CHAPTER 2

Trotterized quantum annealing initialization of the quantum approximate optimization algorithm

2.1 Introduction

Recent technological advances have led to a large number of implementations [A+20b, AAB+20, A+20a, W+19] of so-called Noisy Intermediate-Scale Quantum (NISQ) devices [Pre18]. These machines, which allow to manipulate a small number of imperfect qubits with limited coherence time, inspired the search for practical quantum algorithms. The quantum approximate optimization algorithm (QAOA) [FGG14b] has emerged as a promising candidate for such NISQ devices [ZWC+20, Cro18, WWJ+20].

The QAOA is a variational hybrid quantum algorithm where the classical computer operates a NISQ device. The computer is responsible for the optimization of the cost function over a set of variational parameters. The cost function is calculated using a NISQ device that prepares a quantum state corresponding to chosen parameters and performs quantum measurements. In QAOA of depth p the wave function is prepared by a unitary circuit parametrized by $2p$ parameters, see Fig. 2.1(a). Each of the p layers consist of two unitaries: the first is generated by a classical Hamiltonian H_C that encodes the cost function of a combinatorial optimization problem, and the second is generated by the mixing quantum Hamiltonian, H_B.

While the $p = 1$ limit of QAOA allows for analytic considerations and derivation of performance guarantees [FGG14b], subsequent work suggested that higher depth p may be required in order to achieve a quantum advantage [BKKT19, Cro18]. However, increasing p leads to a progressively more complex optimization landscape, that is characterized by a large number of local suboptimal minima [ZWC+20, WWJ+20, GM19, SSL19], see Fig. 2.1(c). The convergence of classical optimization algorithms into such sub-optimal solutions was demonstrated to be a potential bottleneck of QAOA performance as finding a nearly optimal minimum usually requires exponential in p number of initializations of the classical optimization algorithm [FGG14b, ZWC+20].

The complexity of the energy landscape of large-p QAOA motivated the search for heuristic ways of improving the convergence to a (nearly) optimal minimum values of variational parameters. The recent work demonstrated concentration of QAOA landscape for typical

problem instances [BBF+18b], which implies existence of typical landscape and hints that same variational parameter choice may work between different problem instances or sizes. A particular example of such heuristic was proposed in Ref. [ZWC+20] which constructs good initialization for the QAOA at level $p+1$ using solution at level p, thus requiring a polynomial in p number of optimization runs. Other approaches, such as reusing parameters from similar graphs [SSL19], using an initial state that encodes the solution of a relaxed problem [EMW20], or utilizing machine learning techniques to predict QAOA parameters [AAG20, KSC+19] were also proposed.

In this work we propose a different approach to the QAOA initialization, based on the relation between QAOA and the quantum annealing algorithm. Quantum annealing uses adiabatic time evolution to find the lowest energy state of H_C, but often requires unfeasible evolution time T [AL18]. We explore the observation that Trotterization of unitary evolution in quantum annealing provides a particular choice of parameters for the QAOA [FGG14b]. This leads us to introduce a one-parameter family of Trotterized quantum annealing (TQA) initializations for QAOA, controlled by the time step or, equivalently, total time used in adiabatic evolution.

The central result of our work is the demonstration that TQA initialization for QAOA gives comparable performance to the search over an exponentially scaling number of random initializations. To this end, we establish that TQA initialization leads to convergence of QAOA to a nearly optimal minimum for a certain range of time steps, see Fig. 2.1(c) for visualization. Furthermore, we identify the optimal time step of TQA initialization and suggest a purely experimental way of fixing this parameter of TQA initialization.

Our work reveals a connection between intermediate-p QAOA and short-time quantum annealing. Previous studies [FGG14b, Cro18, ZWC+20] established correspondence between quantum annealing with long annealing times and the QAOA protocol with large p (potentially increasing exponentially with the problem size). More recent work proposed quantum annealing inspired initialization strategies for the so-called 'bang-bang' modification of QAOA [LLL20] that however also correspond to high circuit depths. Our work is different from this context, since we establish that the best performance is achieved for a very *coarse discretization* of quantum annealing, resulting in a realistic circuit depth. We show the existence of an optimal step for TQA discretization that does not depend on problem size and QAOA depth. This suggests an intimate relation between QAOA and TQA, since the optimal value of the time step is in close correspondence to the point where proliferation of Trotter error occurs in TQA [HHZ19].

2.2 Optimization landscape of the QAOA

2.2.1 Visualizing optimization landscape

The performance of the classical optimization in Eq. (1.10) strongly depends on the properties of the optimization landscape. While this landscape can be readily visualized for $p=1$, the dependence of approximation ratio $r_{\vec{\gamma},\vec{\beta}}$ on $2p$ angles parametrizing QAOA was suggested to become progressively more complex for larger values of p. In order to visualize the properties of this high-dimensional landscape, we focus below on properties of points where $r_{\vec{\gamma},\vec{\beta}}$ achieves (local) minima.

We quantify properties of minima using two different characteristics. First, we measure the difference between the approximation ratio of the given minimum characterized by angles $\vec{\gamma},\vec{\beta}$

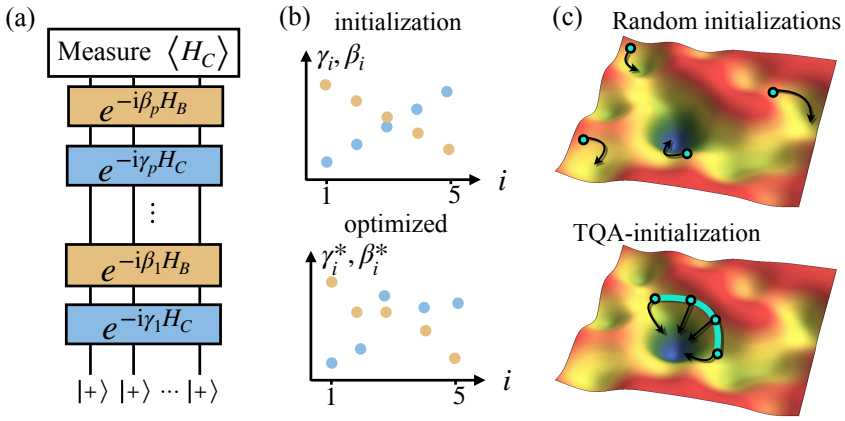


Figure 2.1: (a) The circuit that prepares a quantum state in the QAOA is parametrized by a set of $2p$ angles γ_i, β_i. (b) The optimization of $\langle H_C \rangle$ is launched from a certain guess of parameters and state preparation and measurements are iterated until the algorithm converges to a set of optimized angles γ_i^*, β_i^*. (c) The cartoon of the cost function $\langle H_C \rangle$ landscape as a function of variational parameters shows that random initializations are prone to converge to sub-optimal local minima. In contrast, the family of TQA initializations proposed in this work converges to the (nearly) optimal minimum.

and the global minimum characterized by angles $\vec{\gamma}^*, \vec{\beta}^*$, $\Delta r_{\vec{\gamma},\vec{\beta}} = r_{\vec{\gamma}^*,\vec{\beta}^*} - r_{\vec{\gamma},\vec{\beta}}$. This definition implies that the smallest possible value of $\Delta r_{\vec{\gamma},\vec{\beta}}$ is 0, and larger values of $\Delta r_{\vec{\gamma},\vec{\beta}}$ corresponds to local minima with poor performance (i.e. much larger value of cost function) compared to the global minimum. The second characteristics measures the distance between minima in parameter space,

$$d_{\vec{\gamma},\vec{\beta}} = \sum_{i=1}^{p} \left(|\beta_i - \beta_i^*|_{\frac{\pi}{2}} + |\gamma_i - \gamma_i^*|_{\pi} \right), \tag{2.1}$$

where $|\ldots|_\alpha$ denotes the absolute value modulo α which takes into account symmetries, see Appendix A.1.

We calculate values of $\Delta r_{\vec{\gamma},\vec{\beta}}$ and $d_{\vec{\gamma},\vec{\beta}}$ numerically. For a given graph realization we use 2^p different random initializations of variational parameters $\vec{\gamma}, \vec{\beta}$ and optimize them using the iterative BFGS algorithm [BRO70, Fle70, Gol70, Sha70]. The algorithm is accessed via the `scipy.optimize` python module with default parameters [VGO+20]. Convergence is achieved when the norm of the gradient is less than 10^{-5}, maximum number of iterations is set to $400p$, where p is QAOA depth. In our simulations the routine typically converged before using up the maximum number of allowed iterations. We use the converged angles with lowest value of $r_{\vec{\gamma},\vec{\beta}}$ as an estimate for the global minimum γ_i^*, β_i^*.

Figure 2.2 visualizes the structure of local minima via the joint probability distribution of $\Delta r_{\vec{\gamma},\vec{\beta}}$ and $d_{\vec{\gamma},\vec{\beta}}$ for 50 different graphs using Kernel Density Estimation [Ros56, Par62]. We observe that for QAOA with $p = 5$ the most typical local minima reached from random initialization are far away from the best minimum (corresponding to $\Delta r_{\vec{\gamma}^*,\vec{\beta}^*} = 0$ and $d_{\vec{\gamma}^*,\vec{\beta}^*} = 0$) both in terms of quality of approximation ratio and parameter values. While this figure illustrates a particular choice of system size and QAOA depth, a similar trend is observed for different N, p, and other graph ensembles, see Appendix A.1.

The tendency of random initialization to converge to suboptimal solutions highlights the importance of better initialization methods. In the next section we investigate a family

of initializations inspired by quantum annealing and demonstrate that it achieves a good approximation ratio with a suitable choice of parameters.

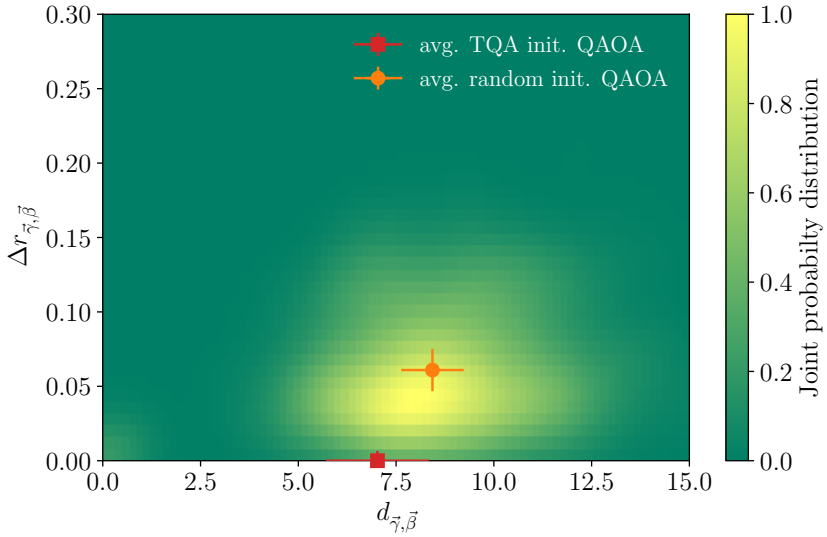


Figure 2.2: Joint probability distribution of distance to the global minimum in parameter space $d_{\vec{\gamma},\vec{\beta}}$ and in terms of approximation ratio $\Delta r_{\vec{\gamma},\vec{\beta}}$ reveals that the most probable outcome of random initialization is a convergence to sub-optimal local minima (yellow region). The orange dot corresponds to average values of $d_{\vec{\gamma},\vec{\beta}}, \Delta r_{\vec{\gamma},\vec{\beta}}$ for random initialization. In contrast, TQA initialization leads to a local minima with a better approximation ratio that occasionally outperforms the best random initialization (red dot, shifted from slightly negative values to $\Delta r_{\vec{\gamma},\vec{\beta}} = 0$ for improved visibility). The data is averaged over 50 random unweighted 3-regular graphs with $N = 12$ vertices and QAOA at level $p = 5$.

2.3 Trotterized quantum annealing as initialization

2.3.1 Optimal time for TQA

Quantum annealing [KN98, BBRA99] was among the first algorithms proposed for quantum computing [FGG+01, FGGS00], and was demonstrated to be universal for $T \to \infty$ and equivalent to digital quantum computing [AvK+08]. The general idea of quantum annealing is to prepare the ground state $|0\rangle_C$ of a classical Hamiltonian H_C starting from the ground state $|0\rangle_B$ of the mixing Hamiltonian H_B using adiabatic time evolution under Hamiltonian $H(t) = (t/T)H_C + (1 - t/T)H_B$. Practical execution of quantum annealing on NISQ devices requires discretization to represent such unitary evolution via a sequence of gates, resulting in the TQA algorithm. The first order Suzuki-Trotter decomposition allows to approximate the time evolution with $H(t)$ over time interval Δt as $e^{-\mathrm{i}\Delta t H(t)} \approx e^{-\mathrm{i}\beta H_B} e^{-\mathrm{i}\gamma H_C} + \mathcal{O}(\Delta t^2)$ with $\beta = (1 - t/T)\Delta t$ and $\gamma = (t/T)\Delta t$.

Applying such decomposition to the quantum annealing protocol that is uniformly discretized over p steps, so that $\Delta t = T/p$ and $t_i = i\Delta t$ we obtain the unitary circuit equivalent to the depth-p QAOA ansatz (1.8) with angles being

$$\gamma_i = \frac{i}{p}\Delta t, \quad \beta_i = \left(1 - \frac{i}{p}\right)\Delta t. \tag{2.2}$$

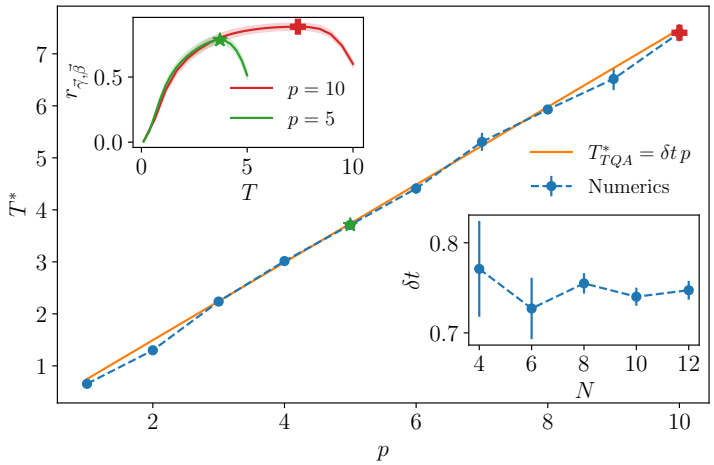


Figure 2.3: Optimal time of TQA evolution T^* increases linearly with number of discretization steps p. Top inset illustrates that optimal performance of TQA at time T^* is followed by the rapid decrease in approximation ratio at longer times T^*. Data is shown for $N = 12$. Bottom inset shows finite size scaling of the time step δt, determined by the slope of the T^* vs p dependence, that assumes approximately constant value with the graph size. All averaging is performed over 50 random instances of unweighted 3-regular graphs.

In what follows we refer to such choice of angles as TQA initialization, controlled by the time step Δt at a fixed depth p.

The mapping between TQA and QAOA along with the universality of quantum annealing for $T \to \infty$ was previously used as an argument for the existence of good QAOA protocols at depths $p \to \infty$ [FGG14b]. Typically the required evolution time of quantum annealing is inversely proportional to the square of the minimal energy gap $T \propto \Delta^{-2}$ encountered in the Hamiltonian $H(t)$ over the time evolution. Numerous studies established that the time required for a good performance often blows up exponentially due to encounter of exponentially small gaps in N [AL18].

In contrast to previous studies, we investigate TQA performance in a different setting that is motivated by its subsequent usage as QAOA initialization. The QAOA is characterized by a fixed circuit depth, p. Therefore, we fix p and study the performance of TQA as a function of total time or, equivalently, time step Δt, related as $T = p\,\Delta t$. Generally the performance of quantum annealing tends to increase with the total annealing time. However in case of fixed p, longer annealing time corresponds to a coarser discretization, which leads to larger Trotter errors that scale proportionally to $\mathcal{O}(\Delta t^2)$ at small values of Δt. It is the interplay between increased efficiency and Trotter errors that leads to the existence of an optimal annealing time in the present setting. This is illustrated in Fig. 2.3 (top inset), where the approximation ratio for the TQA protocol increases with T for small times, reaching a maximum at time T^* followed by a sharp downturn. The sharp decrease of QA performance after T^* was reported by [HHZ19], who attributed a phase transition caused by proliferation of Trotter errors.

Main panel of Fig. 2.3 reveals a linear scaling of the optimal time T^* with the number of time steps p. This is equivalent to the existence of an *optimal time step* δt, that determines T^* as

$$T^*_{\mathsf{TQA}} = \delta t\, p. \tag{2.3}$$

The bottom inset in Fig. 2.3 shows that the time step δt defined as a slope of a linear fit

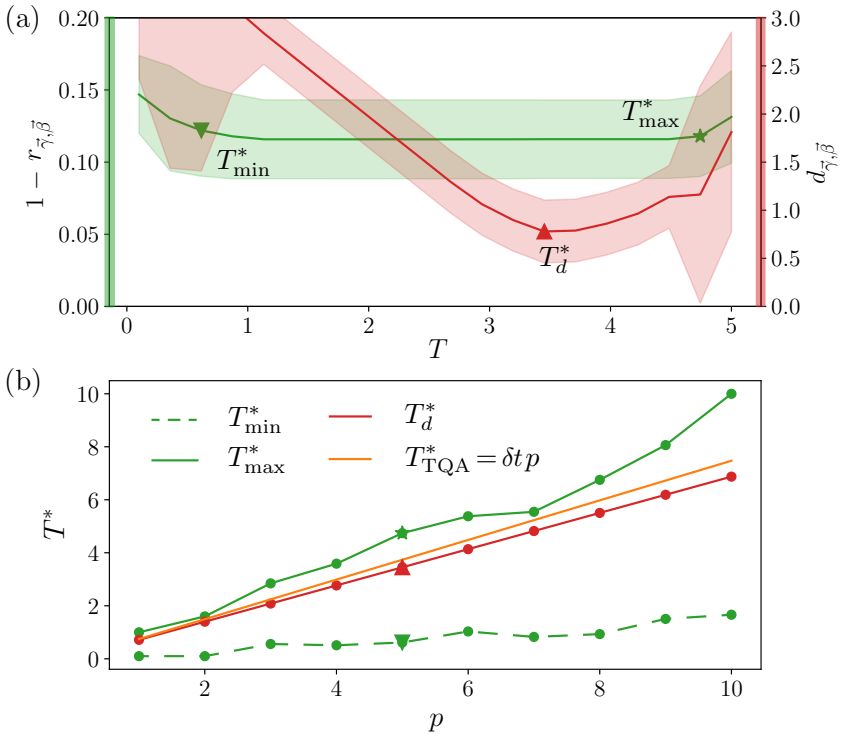


Figure 2.4: (a) Approximation ratio of the $p = 5$ QAOA as a function of TQA initialization time T reveals that a range of initialization times $[T^*_{\mathsf{min}}, T^*_{\mathsf{max}}]$ (green triangle and star) yield the performance within 1% of the minimal $1 - r_{\vec{\gamma},\vec{\beta}}$. On the other hand, the study of the distance between TQA initialization and converged value of angles reveals the existence of a time T^*_d where the QAOA performs the smallest parameter updates. (b) All three times T^*_{min}, T^*_{max}, and T^*_d defined in panel (a) increase linearly with QAOA circuit depth p. Moreover, the T^*_d is very close to the time where TQA protocol itself achieves optimal performance, T^*_{TQA}, see Fig. 2.2. Data was obtained for $N = 12$ and averaged over 50 random graphs.

of T^* with p converges with the problem size N. This gives a strong evidence that δt is a well-defined quantity in the thermodynamic limit $N \to \infty$. For the family of the 3-regular graphs considered here we observe that the optimal time step tends to value $\delta t \approx 0.75$. The existence of an optimal time step that is of order one holds for three other graph ensembles, considered in Appendix A.2, although the numerical value of this time step depends on the specific graph ensemble.

We use the TQA initialization in Eq. (2.2) with time step $\Delta t = 0.75$ for the QAOA and observe in Fig. 2.2 that it allows to avoid the local minima and helps the QAOA to converge to a minima that is very close to the global minima in terms of approximation ratio. This result motivates the systematic analysis of the performance of TQA initialization.

2.3.2 TQA initialization of QAOA

We continue with a detailed study of the TQA initialization defined in Eq. (2.2) as a function of time T at fixed p. The green line in Fig. 2.4(a) reveals that approximation ratio remains constant for a range of times, denoted as $[T^*_{\mathsf{min}}, T^*_{\mathsf{max}}]$. This figure shows results for $p = 5$ QAOA applied to graphs with $N = 12$ vertices, but a similar trend holds for other values of depth, problem sizes, and graph ensembles. The constant approximation ratio in a range of T is naturally explained by the convergence of parameter optimization routine to the same

minimum for $T \in [T^*_{\mathsf{min}}, T^*_{\mathsf{max}}]$, see cartoon in Fig. 2.1(c). In order to discriminate between different times in the above range, we study the distance between initialization parameters and optimized values of $\vec{\gamma}, \vec{\beta}$. The red line Fig. 2.4(a) shows that this distance has a well-pronounced minimum at a time denoted as T^*_d that is contained within the same interval $[T^*_{\mathsf{min}}, T^*_{\mathsf{max}}]$. The TQA initialization with time T^*_d is closest to the local minimum achieved from it in a sense of distance defined in Eq. (2.1).

All three times T^*_{min}, T^*_{max}, and T^*_d were defined above using the QAOA with fixed depth p. Figure 2.4(b) reveals that all three times scale approximately linearly with p. This allows to define a range of time steps for TQA initialization that yield the same performance of optimized QAOA, $\Delta t \in [0.16, 0.92]$ for the present graph ensemble. Moreover, the time T^*_d nearly coincides with the optimal TQA time $T^*_{\mathsf{TQA}} = \delta t\, p$ obtained in the previous section, implying that $\Delta t = \delta t = 0.75$ is the optimal value of time step. This result also holds for the MaxCut problem on other graph families, see Appendix.

The similarity between the optimal time of the TQA protocol to the time where the angular distance $d_{\vec{\gamma},\vec{\beta}}$ between the initial and final protocol is minimized, suggests that the performance of the QAOA is bounded by the same phase transition that occurs in TQA [HHZ19]. However, the QAOA is able to provide a significant improvement over TQA by doing additional optimizations of variational parameters. Recent work [ZWC+20] suggested that such performance improvement may be due to utilization of "diabatic pumps" that allow to return the population from excited states back to the ground state. This could potentially explain the systematic deviation of the QA protocol from TQA initialization as seen in Fig. A.3 in Appendix A.3.

Finally, we compare the performance of QAOA that used 2^p random initializations to the QAOA launched from TQA initialization with optimal time step δt. Surprisingly, Fig. 2.5 shows that TQA initialization yields the *same* performance as the best result for random initialization even for QAOA protocols with depth comparable to the problem size, N. Moreover, the inset of Fig. 2.5 illustrates that the excellent performance of TQA initialization holds true for a broad range of system sizes N, while Appendix A.4 presents equally encouraging results for other graph ensembles. Note that the QAOA performance for fixed p decreases with system size N, which was attributed to the fact that the QAOA with fixed p cannot "probe" the whole graph. In order for the QAOA to achieve constant performance for increasing problem size N, the depth of QAOA should increase at least as $\log N$ [ZWC+20].

2.4 Summary and Discussion

Our central result is the establishment of a family of TQA initializations for QAOA parametrized by a time step Δt. We find that TQA initialization allows the QAOA to find a solution close to the global optima for a broad range of parameter Δt. In this range our initialization scheme achieves results similar to the best outcome of 2^p random initializations, with a single optimization run. Moreover we establish a heuristic way to identify the optimal Δt for the TQA initialization from the performance of the TQA protocol.

Our results open the door to more time-efficient practical implementations of the QAOA on NISQ devices. To this end, we propose a two-step practical NISQ algorithm that capitalizes on the success of TQA initialization and uses the heuristic results to establish an optimal value of the time step. The first two steps of Algorithm 1 simulate the TQA protocol on a NISQ device, thus obtaining an estimate for the optimal time in the TQA initialization. This can be readily carried out on today's NISQ devices [SKPK19]. The second part of the algorithm consists

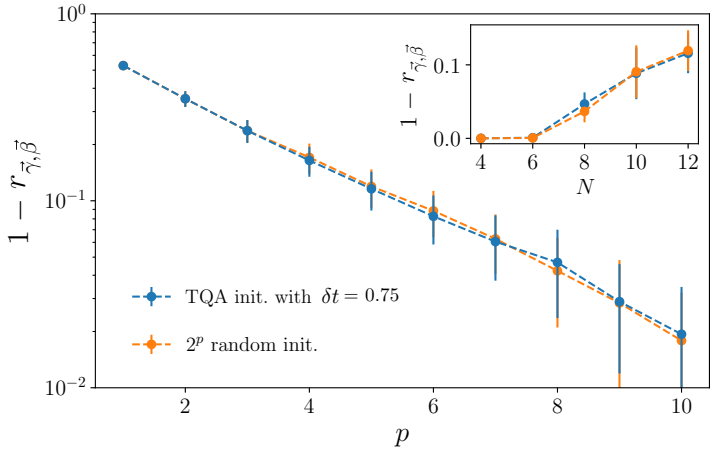


Figure 2.5: A single optimization run of the QAOA with TQA initialization with time $\delta t\, p$ yields equivalent performance to the best out of 2^p random initializations. System size is $N = 12$. Inset reveals that the comparable performance persists over the entire range of considered system sizes. Averaging was performed over 50 random graphs.

of running the QAOA optimization loop using values of variational parameters according to Eq. (2.2).

Algorithm 1 QAOA with TQA initialization

1: Choose p_1 and p_2 such that $p_1 < p_2$.
2: Estimate the slope δt using the TQA algorithm:
find optimal times $T^*_{1,2} \leftarrow \arg\min_{T_{1,2}} \left\langle H_C \right\rangle_{p_{1,2}}$
and set the value of time step $\delta t \leftarrow \frac{T^*_2 - T^*_1}{p_2 - p_1}$, see Fig. 2.3.
3: Use TQA initialization $\gamma_i \leftarrow \frac{i}{p}\delta t$ and $\beta_i \leftarrow (1 - \frac{i}{p})\delta t$.
4: Run the QAOA parameter optimization, see Fig. 2.1.

Numerical simulations presented above suggest good performance of the above algorithm in the idealized case when presence of noise, gate errors, and other imperfections are neglected. Moreover, the fact that TQA initialization converges to a good minimum for the range of times (equivalently, time steps) $T \in [T^*_{\min}, T^*_{\max}]$, see Fig. 2.4, suggests that this algorithm has a high tolerance towards imperfections in determining the value of δt. Determining the performance of this algorithm on a real NISQ device or incorporating some of the imperfections into the numerical simulation remains an interesting open problem.

In our studies we restricted our attention to the MaxCut problem and demonstrated success of our approach for three different random graph ensembles. We expect that this results hold for other graph ensembles, provided that concentration of QAOA landscape is true [BBF+18b]. It is also interesting to check if our findings hold true beyond the MaxCut problem. Furthermore, it will be interesting to study the finite size scaling for problem sizes $N > 12$ considered here using Matrix Product States (MPS) [Sch11b] or neural-network quantum states [CT17, MC20].

In addition to practical NISQ algorithms, our finding suggest a previously unknown connection between the QAOA at relatively small circuit depth and quantum annealing. The fact that quantum annealing inspired initializations belong to a basin of attraction of a high-quality minimum in the QAOA landscape, see Fig. 2.1(c), invites a more comprehensive study of the QAOA landscape from this perspective. How many good quality minima typically exist in such

landscape? How different are they from each other and what are their basins of attraction? Can one use other information measures such as entanglement or Fisher information [ASZ+20] to characterize the QAOA landscape? Finding answers to such questions may lead to other prospective families of QAOA initializations.

While TQA provides a good initialization, the subsequent QAOA optimization is able to significantly improve the performance. Understanding the underlying mechanisms of such performance improvement is an outstanding challenge. In particular, there remains an intriguing possibility that QAOA optimization routine implements some of the techniques, developed to improve the annealing fidelity, such as diabatic pumps [ZWC+20], shortcuts to adiabaticity [GORK+19], and counterdiabatic driving [SP17, CPSP19]. The fact that the optimal time step coincides with the point of proliferation of Trotter errors [HHZ19], thus effectively taking maximal possible value suggests interesting parallels to the Pontryagin's minimum principle considered in context of variational quantum algorithms [YRS+17].

To conclude, we hope that TQA initialization of the QAOA established in this work will help to achieve practical quantum advantage by executing the QAOA on available devices and inspire future research that could lead to better understanding of what happens under the hood of QAOA optimization.

CHAPTER 3

Recursive greedy initialization of the quantum approximate optimization algorithm with guaranteed improvement

3.1 Introduction

The quantum approximate optimization algorithm (QAOA) [FGG14b] 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, SS21b].

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$^+$20], or to use the parallels between the QAOA and quantum annealing [SS21b]. Recently the success of the interpolation strategies that appeal to annealing was attributed to the ability of the QAOA to effectively speed up adiabatic evolution via the so-called counterdiabatic mechanism [WL22]. This result was used to explain cost function concentration for typical instances and concentration of optimal, typically smoothly varying, parameters, which was previously introduced on Ref. [BBF$^+$18a] and [ARCB21] respectively.

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

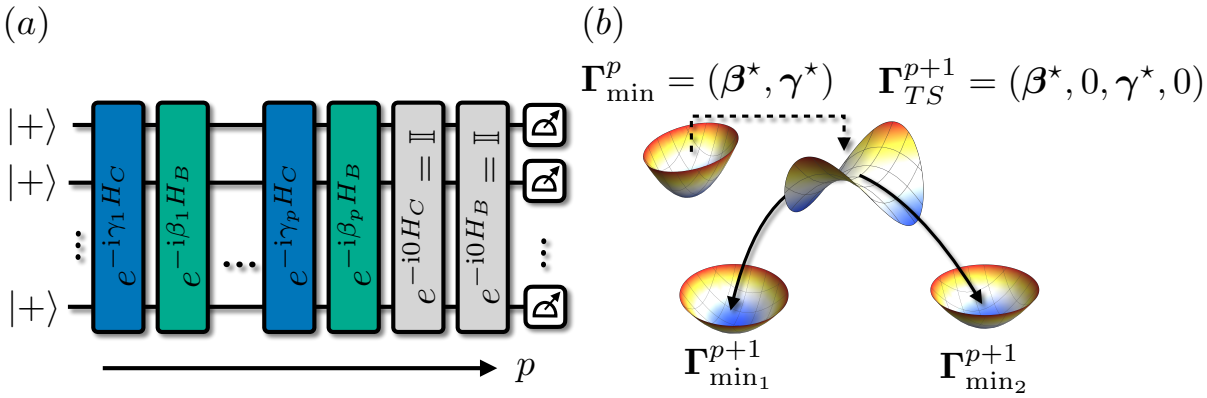


Figure 3.1: (a) Circuit diagram that implements the QAOA ansatz state with circuit depth p, see Eq. (1.8). Gray boxes indicate the identity gates that are inserted when constructing a TS, as indicated in Theorem 1. (b) Local minima $\mathbf{\Gamma}^{p}_{\min}$ of QAOA$_p$ generate a TS $\mathbf{\Gamma}^{p+1}_{\text{TS}}$ for QAOA$_{p+1}$ that connects to two *new local minima*, $\mathbf{\Gamma}^{p+1}_{\min_{1,2}}$ with lower energy.

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. 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 chapter is organized as follows. In Section 3.2 we review the QAOA, present newly found symmetries, and introduce the analytical construction of TS. In Section 3.3 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 3.4 we discuss our results and potential future extensions of our work. Appendices B.1-B.6 present detailed proofs of our analytical results, as well as supporting numerical simulations.

3.2 QAOA optimization landscape

3.2.1 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 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.

3.2.2 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 1 (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*

$$\begin{aligned}\mathbf{\Gamma}^{p+1}_{TS}(i,j) = (&\beta^\star_1, ..., \beta^\star_{j-1}, 0, \beta^\star_j, ..., \beta^\star_p,\\ &\gamma^\star_1, ..., \gamma^\star_{i-1}, 0, \gamma^\star_i, ..., \gamma^\star_p)\end{aligned} \tag{3.1}$$

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. (3.1) 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

$$\begin{aligned}\partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l)} &= \partial_{\beta_{l-1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}},\\ \partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l+1)} &= \partial_{\beta_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}},\\ \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l)} &= \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}},\\ \partial_{\gamma_{l+1}}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,l+1)} &= \partial_{\gamma_l}|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle\big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}}.\end{aligned} \tag{3.2}$$

Since $\nabla E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\mathbf{\Gamma}^{p}_{\mathsf{min}}} = 0$, it directly follows that the TS constructed using Theorem 1 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 B.2 we show that we can always write the Hessian at the TS in the following form

$$H\left[\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}(l,k)\right] = \begin{pmatrix} H(\mathbf{\Gamma}^{p}_{\mathsf{min}}) & v(l,k) \\ v^T(l,k) & h(l,k)\end{pmatrix}, \tag{3.3}$$

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

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

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 taking over the old parameters and the second derivative corresponding 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 Theorem 5) 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 B.2.

3.3 From transition states to QAOA intialization

3.3.1 Initialization graph

For each local minimum of QAOA$_p$, Theorem 1 provides $p+1$ symmetric TS where zeros are padded at the same position, $i=j$, like in Fig. 3.1(a), and additionally p non-symmetric TS with $j=i+1$, where zeros are padded in adjacent layers of the QAOA circuit. Fig. 3.1(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 1 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. 3.2. We find a unique minimum for QAOA$_1$ using grid search (see Appendix B.5) in the fundamental region defined in Eq. (B.52) from which we construct two symmetric TS according to Eq. (3.1), 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. 3.2. 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.

3.3.2 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 1.1 (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*

[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.

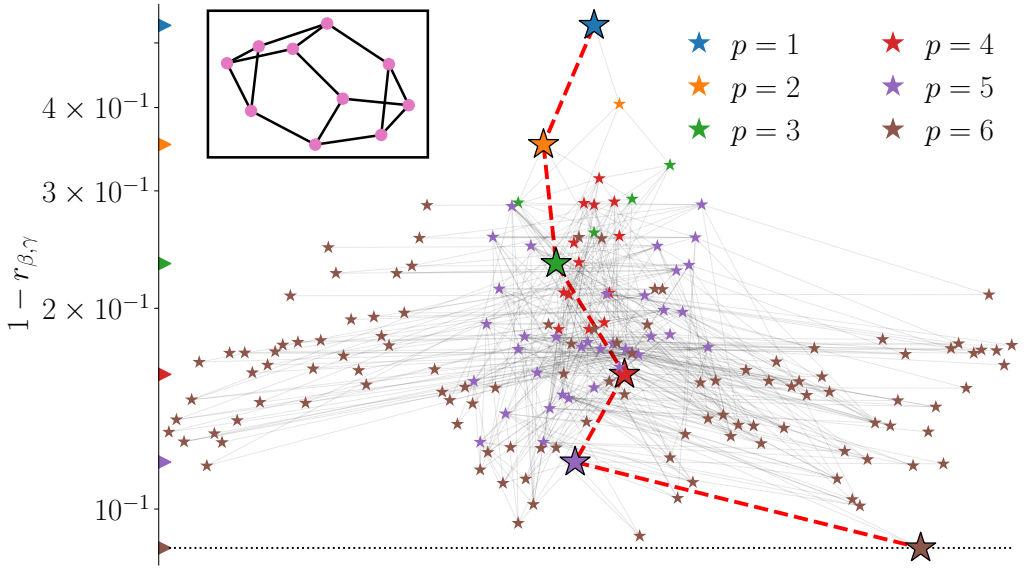


Figure 3.2: Initialization graph for the QAOA for MaxCut problem on a particular instance of RRG3 with $n = 10$ vertices (inset). For each local minima of QAOA$_p$ we generate $p+1$ TS for QAOA$_{p+1}$, find corresponding minima as in Fig. 3.1(b), and show them on the plot connected by an edge to the original minima of QAOA$_{p+1}$. Position along the vertical axis quantifies the performance of QAOA via the approximation ratio, and 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.

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 B.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. 3.2 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 B.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 [SS21b], that initializes QAOA$_p$ using $\gamma_j = (1 - \frac{j}{p})\Delta t$ 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. 3.3 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.

3.3.3 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

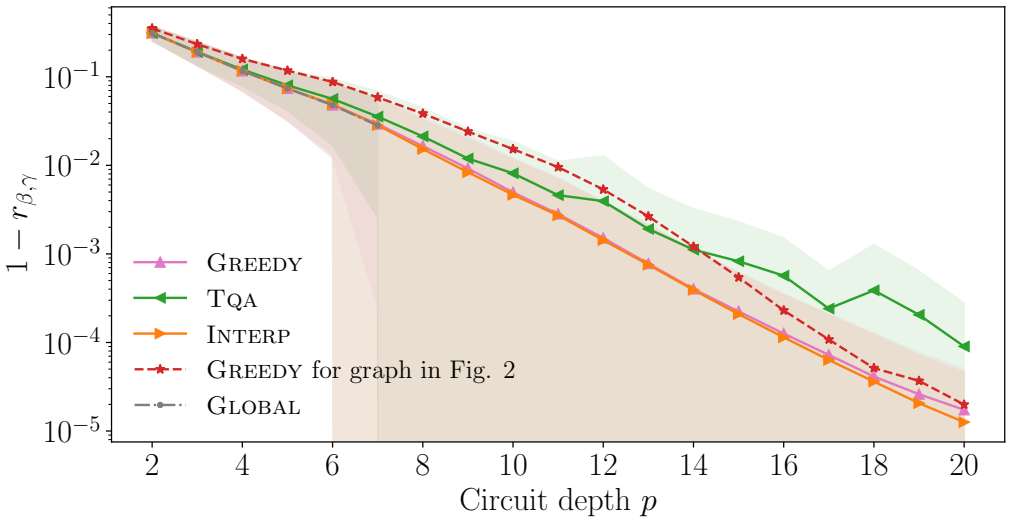


Figure 3.3: Performance comparison between different QAOA initialization strategies used for avoiding low-quality local minima. Greedy approach proposed in this work yields the same performance as Interp [ZWC+20] and slightly outperforms TQA [SS21b] at large p. Global refers to the best minima found 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 B.6.

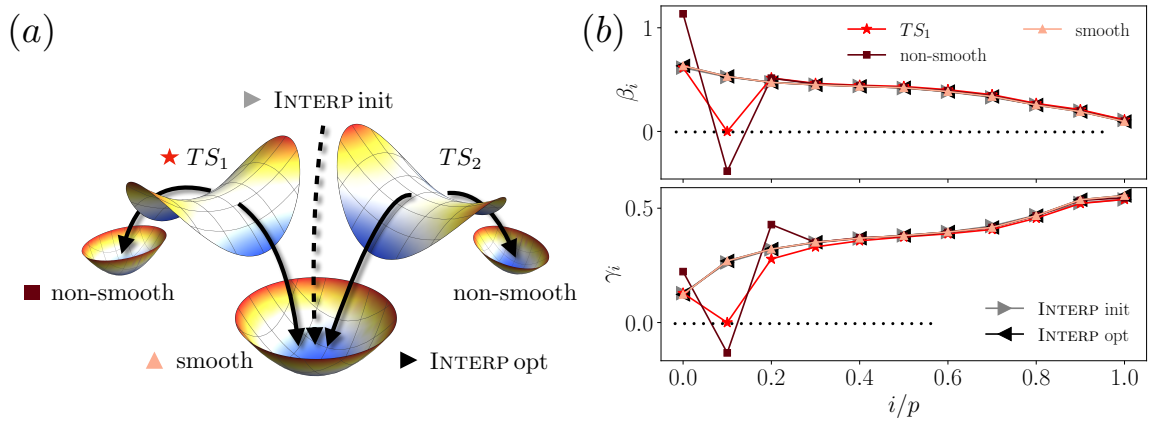


Figure 3.4: (a) Cartoon of descent from two different TS at of QAOA$_{p+1}$ generated from a QAOA$_p$ minimum with a smooth pattern leads to the same new smooth pattern minima of QAOA$_{p+1}$, also reached from the INTERP [ZWC+20] initialization. Two additional non-smooth local minima typically have higher energy. (b) shows the corresponding initial and convergent parameter patterns for the RRG3 graph shown in Fig. 3.2 for $p = 10$.

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 1 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. 3.4(a), finding two new minima. Remarkably, the eigenvector corresponding to the index-1 direction of the Hessian has dominant weight on the variational angles with initially zero value, see B.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. 3.4(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 in 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 [FGG14b, SS21b, 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. 3.4(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. 3.4(b). Geometrically, the Interp initialization can be obtained from the symmetric TS constructed by Theorem 1 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 $T \sim p$ [SS21b], 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.

3.4 Summary and 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. 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 [FGG14b, WL21, FGG20] 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+17b, PMS+14b] and quantum machine learning [BLSF19a].

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, particularly 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 4

Generalization of the quantum approximate optimization algorithm to qudits

4.1 Introduction

The field of quantum computing has seen tremendous advancements in recent years, with substantial progress in hardware across various platforms [KEA+23, EBK+23] as well as first promising demonstrations of quantum algorithms executed on hardware [EKC+22]. Despite great advancements in hardware and algorithms, quantum computers have not yet achieved quantum advantage, i.e., outperformed a classical computer for a practically relevant task. This is largely due to noise and limited qubit connectivity, which restricts each qubit's ability to interact directly with other qubits in the system. These factors limit the expressibility and the range of quantum circuits that can be implemented.

To address these challenges, we propose an alternative approach that involves qudits, quantum systems with more than two levels, as the core computational units instead of qubits. Despite being natively available on various hardware architectures, such as superconducting, ion trap, and neutral atom quantum computers, the use of qudits for computation has so far been limited. Notable experiments include IBM's implementation of a generalized measurement (so-called POVM) using a four-level qudit system [FMT+22] and a team in Innsbruck demonstrating a universal gate set with up to seven energy levels [RMP+22] on an ion trap quantum computer. These initial implementations highlight the potential of qudits, yet much of this potential remains unexplored, in particular for the application of quantum algorithms.

Qudits allow for a richer exploration of the computational space due to the larger dimension of the Hilbert space (d^n where n is the number of qudits and d is the number of energy levels), as compared to qubits ($d = 2$). This increased dimension utilized by qudits could be a promising path to enhance the expressibility of near-term quantum algorithms.

To explore this, we propose a generalization of the quantum approximate optimization algorithm (QAOA) to qudit systems. We apply the Qudit-QAOA to the problem of graph coloring, a well-known NP-hard problem in computer science [Kar72]. There, the qudits' d energy levels can conveniently represent the colors of nodes in a graph, and the QAOA ansatz can evolve the system to a configuration where no nodes connected by an edge have the same color,

see Fig. 4.2 for an illustration. We will show, that in this setting, a correct coloring can be interpreted as the ground state of the antiferromagnetic Potts model.

Next, we will study the issue of noise and explore how it affects the performance of the Qudit-QAOA. In particular, we propose an efficient technique that allows for a highly computationally efficient simulation of single qudit noise. Lastly, we explore how the algorithm can be implemented on an ion trap quantum computer using hardware native gates.

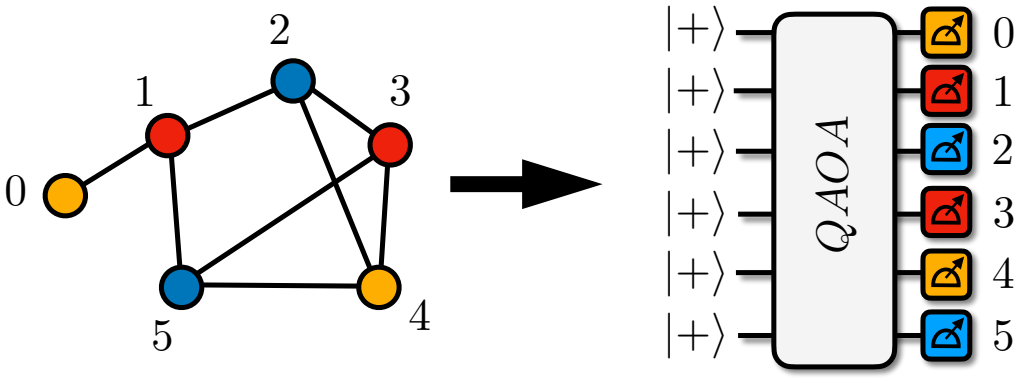


Figure 4.1: Illustration of a graph with six vertices that is colorable with three colors (yellow, red, and blue). In the Qudit-QAOA each qudit is assigned to one vertex in the graph and each qudit state represents one color. Here the initial state $|+\rangle^{\otimes n}$ is an equal superposition of all possible colorings.

Note that this work is based on work by [BKKT20] who originally proposed an extension of the QAOA to qudit systems. Our work extends on this in two main avenues: first, we provide an alternative formulation of the Qudit-QAOA ansatz that allows for an expression in terms of quantum gates. Secondly, we analyze the effect of noise on the Qudit-QAOA, a previously entirely unexplored subject.

The remainder of the chapter is organized as follows: Sec. 4.2 discusses the basics of qudits and how Pauli matrices can be generalized to higher dimensions. In Sec. 4.3 we discuss the graph coloring problem and propose the Qudits-QAOA circuit. Furthermore, we show how the effect of noise can be simulated efficiently and present numerical results for two different graphs. In addition, we show that the Qudits-QAOA can be readily implemented on an ion trap quantum computer using hardware native gates. In Sec. 4.4 we numerically study the performance of the Qudit-QAOA under the effect of noise. Lastly, in Sec. 4.5 we summarize the results and discuss their implications.

4.2 Qudits - Beyond Two-Level Systems

4.2.1 Qudits as generalizations of qubits

Qudits are d-level quantum systems that generalize the two-level systems commonly known as qubits. A qudit, with $d > 2$, can exist in any superposition of its d basis states. These basis states, also known as computational basis states, are often denoted as $\{|0\rangle, |1\rangle, ..., |d-1\rangle\}$, similar to the qubit basis states $\{|0\rangle, |1\rangle\}$. Each qudit state can then be expressed as a linear combination of these basis states $|\psi\rangle = \sum_{j=0}^{d-1} c_j |j\rangle$ where c_j are complex coefficients such that $\sum_{j=0}^{d-1} |c_j|^2 = 1$ for normalization.

The set of local operations on a qudit is described by the group $SU(d)$. In the case of a qutrit, the Lie algebra of $SU(3)$ is spanned by Gell–Mann matrices, which are a natural generalization of the Pauli matrices from $SU(2)$ to $SU(3)$, with further generalizations towards $SU(d)$ [GM62]. For our purposes we require generalizations of the standard Pauli matrices σ^z (phase flip), σ^x (bit-flip) and the Hadamard matrix H which implements a basis transformation between the computational basis and the X-basis.

4.2.2 Generalization of Pauli matrices

In particular we have have the generalized X_d operator, which implements a cyclic shift

$$X_d = \sum_{j=0}^{d-1} |j \oplus 1\rangle\langle j|, \tag{4.1}$$

where $\oplus$ denotes addition modulo d. For the generalized Z_d we have

$$Z_d = \sum_{j=0}^{d-1} \omega^j |j\rangle\langle j|, \tag{4.2}$$

where $\omega = e^{2\pi i/d}$, a dth root of unity, the operator thus introduces a relative phase to the basis states. These generalized Pauli operations maintain many of the same basic properties as their qubit counterparts, such as unitarity and tracelessness.

4.2.3 Generalization of the Hadamard gate

Finally, we turn our attention to the generalized Hadamard gate. The generalization of the Hadamard gate to qudits can be expressed as

$$H_d = \frac{1}{\sqrt{d}} \sum_{j,k=0}^{d-1} \omega^{jk} |j\rangle\langle k|. \tag{4.3}$$

Note that this is not yet a universal gate set, in particular we have not yet defined an entangling gate. We will discuss the entangling gate used in this work in detail in Sec. 4.3.3.

4.3 Generalization of the QAOA to qudits

4.3.1 Graph coloring and Potts model

While the QAOA can be implemented for any binary optimization problem it was originally suggested for the problem of MaxCut [FGG14a]. There the goal is to partition a graph $\mathcal{G} = \{E, V\}$, consisting of vertices V and edges E, such that the partition cuts through a maximal number of edges, see Sec. 1.3 for a detailed discussion. The assignment of the vertices is thus binary. In the QAOA the two energy levels of the qubits are use to represent the two binary choices. However, when working with qudit systems which offer multiple energy levels, graph coloring becomes a suitable problem to exploit the additional dimensions. In the graph coloring problem, the task is to assign colors to the vertices of a graph such that no two adjacent vertices have the same color. The goal is to find the smallest number of colors required to properly color a graph, known as the chromatic number χ, this problem is known to be NP-hard. In this context we can understand MaxCut as a special case of graph coloring with just two colors or 2-coloring.

We can frame the problem of finding a correct coloring as a minimization problem for the following cost function

$$C = \sum_{i,j \in E} \delta_{c_i c_j}, \tag{4.4}$$

where c_i is the color of the i-th qudit, this is known as the antiferromagnetic Potts model in condensed matter physics (see [Wu82] for a review). The aim is thus to find a set of colors $\{c_i\}$ such that the cost function is minimized. There is an energy contribution if two colors are the same, on two vertices connected by an edge, and no energy contribution if two colors are different. A valid coloring where no two vertices connected by an edge have the same color, thus has cost function value zero. In the complete graph coloring problem there is the additional complexity of finding the minimal number of colors χ for which the cost function has a ground state with zero energy. The problem can thus be formulated as

$$\chi = \min_{d, \{c_i\}_{i=1}^N} d : C(\{c_i\}) = 0, c_i \in 0, 1, \ldots, d-1 \ \forall i, \tag{4.5}$$

which means that we find aim to find a set of colors $\{c_i\}$ such that the cost function C is zero with the least number of colors d. If this problem is solvable, we find the chromatic number χ, if not we simply find a low energy coloring which has a non-zero cost function value.

With this established, we can make a connection between the concept of colors in the graph and quantum states of the qudits. To do so, we can express our cost function in terms of quantum projectors. Each color c_i corresponds to a specific quantum state $|a\rangle_i$ of the i-th qudit. The Kronecker delta in the cost function, $\delta_{c_i c_j}$, is then equivalent to a projector in the i-th and j-th qudit state spaces

$$\delta_{c_i c_j} = \sum_{a=0}^{d-1} |a\rangle_i \langle a|_i \otimes |a\rangle_j \langle a|_j . \tag{4.6}$$

This equivalence is a natural one: the projector returns the value 1 (analogous to the Kronecker delta) when the states $|a\rangle_i$ and $|a\rangle_j$ are the same, thus indicating that the vertices i and j share the same color and contributing to the cost function. Conversely, if the states are different, the projector returns 0 and the cost function does not increase. This way, the quantum representation of the graph coloring problem naturally mirrors its classical counterpart, and this is what allows us to use qudits to solve the problem.

4.3.2 Qudit-QAOA ansatz circuit

Similar to the QAOA ansatz for qubits in Eq. (1.8) the Qudit-QAOA requires two alternating unitaries that drive the evolution from the equal superposition $|+\rangle^{\otimes n}$ to the ground state of C. One unitary is the classical term that involves the diagonal Hamiltonian C and generates entanglement between qudits, the second unitary is the non-diagonal quantum term that generates a global single qudit rotation that allows to shift the weight of the wave function between qudit states. Inspired by Ref. [BKKT20] we propose the following Qudit-QAOA ansatz circuit

$$|\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle = \prod_{t=1}^{p} U_B(\boldsymbol{\beta}_t)^{\otimes n} e^{-i\gamma_t C} |+\rangle^{\otimes n}, \tag{4.7}$$

where similar to the qubit case $|+\rangle^{\otimes n}$ is an equal superposition of all basis states, which is equivalent to an equal superposition of all colors. The classical unitary $e^{-i\gamma_t C}$ entangles pairs of qudits, we have that

$$e^{-i\gamma_t C} = \prod_{i,j \in E} e^{-i\gamma_t \delta_{c_i c_j}}. \tag{4.8}$$

$U_B(\boldsymbol{\beta}_t)$ is the non-diagonal quantum unitary, given by

$$U_B(\boldsymbol{\beta}) = H^\dagger \left(\sum_{a \neq 0} e^{i\beta_a} |a\rangle \langle a| \right) H, \tag{4.9}$$

where H is the generalized Hadamard gates as defined in Eq. (4.3). The unitary implements a phase shift of phase $e^{i\beta_a}$ in the X-basis, where the basis transformation is implemented by the Hadamard gates. For the variational parameters we have that $\boldsymbol{\beta} \in \mathbb{R}^{p(d-1)}$ and $\gamma \in \mathbb{R}^p$ which implies that there is total of pd variational parameters, see Appx. C.1 for a more detailed discussion of the ansatz and its relation to the proposal in Ref. [BKKT20].

4.3.3 Hardware native qudit entangling gate for ion tap quantum computer

Lastly we will discuss how the Qudit-QAOA can be run on an ion trap quantum computer. In order to implement the ansatz from Eq. (4.7) on a quantum hardware we need to find a representation of the classical and quantum unitary in terms of available gates. In recent work by [HWG+23] a novel entangling gate was suggested for qudit entanglement on an ion trap quantum computer. They experimentally demonstrate the implementation of the native two-qudit entangling gate up to dimension $d = 5$ in a trapped-ion system. This is achieved by generalizing a recently proposed light-shift gate mechanism to generate genuine qudit entanglement in a single application of the gate. The action of the gate can be expressed as

$$G(\gamma) : \begin{cases} |jj\rangle \to |jj\rangle \\ |jk\rangle \to e^{i\gamma} |jk\rangle \quad j \neq k. \end{cases} \tag{4.10}$$

The gate thus implements a phase shift of $e^{i\gamma}$ on two qudits that are not in the same state, or equivalently the same color. If we consider again the classical unitary from Eq. (4.8) we find that the term $e^{-i\gamma_t \delta_{c_i c_j}}$ is equivalent to the entangling gate $G(\gamma)$ up to a global phase, see Appx. C.2 for details. This allows us to implement the classical unitary as

$$e^{-i\gamma_t C} \sim \prod_{i,j \in E} G_{i,j}(\gamma_t), \tag{4.11}$$

or in words the native qudit entangling gate acts on pairs of qudits that are elements of the edges E. This is a highly compact representation since it only requires a total of $|E|$ entangling gates. Typically entangling gates would only target pairs of qudit states rather then all of them at once, which would require multiple gates per qudit pair which vastly increases errors.

4.3.4 Representation of quantum unitary using phase shifts

Next, we require a compact representation of the quantum unitary $U_B(\boldsymbol{\beta})$. Eq. (4.9) is straight forward to implement using the phase shift gate presented in Ref. [RMP+22] given by

$$Z_i(\beta) |j\rangle = \begin{cases} e^{-i\beta} |j\rangle & \text{if } i = j \\ |j\rangle & \text{else,} \end{cases} \tag{4.12}$$

which allows to induce a phase shift on an arbitrary level. We can use this to implement the term $\sum_{a \neq 0} e^{i\beta_a} |a\rangle \langle a|$ as

$$\sum_{a \neq 0} e^{i\beta_a} |a\rangle \langle a| = \prod_{a \neq 0} Z_a(-\beta_a). \tag{4.13}$$

To implement the full quantum unitary $U_B(\boldsymbol{\beta})$ we also require generalized Hadamard gates H which are also readily available [RMP+22]. This representation is highly optimal since the phase shift gates are in fact implemented "in software" and does not induce any error. The Hadamard gates have a constant error, which makes the errors of this gate independent of the angles $\boldsymbol{\beta}$ which is highly desired since one aims at finding optimal parameters that minimize the cost function.

4.3.5 Qudit-QAOA circuit using native ion trap gates

We can thus write down the full Qudit-QAOA circuit for ion traps using a hardware native qudit entangling gate and virtual phase shift gates

$$|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle = \prod_{t=1}^{p}\left[\Big[H^\dagger\Big(\prod_{a\neq 0} Z_a(-\beta_a^t)\Big)H\Big]^{\otimes n}\prod_{i,j\in E} G_{i,j}(\gamma_t)\right]|+\rangle^{\otimes n}, \tag{4.14}$$

we illustrate the circuit in Fig. 4.2. This circuit is highly compact and ideal for an implementation on an ion trap qudit device. Ion traps offer all-to-all connectivity which implies that the term $\prod_{i,j\in E} G_{i,j}(\gamma_t)$ can be directly implemented without requiring an swaps with just $|E|$ two qudit entangling gates. For qutrits ($d=3$) $G_{i,j}$ has an error of roughly $\sim 1\%$ while for higher d is may be larger. The single qudit Hadamard gates have an error of $\sim 2\times 10^{-4}\frac{d^2}{2}$ while the phase shift gates are implemented in software at the end of the measurement and thus do have any error. From previous experiments on the ion trap hardware we know that in order to achieve a circuit fidelity of over 50% we can roughly implement total of $p|E| < 50$ gates.

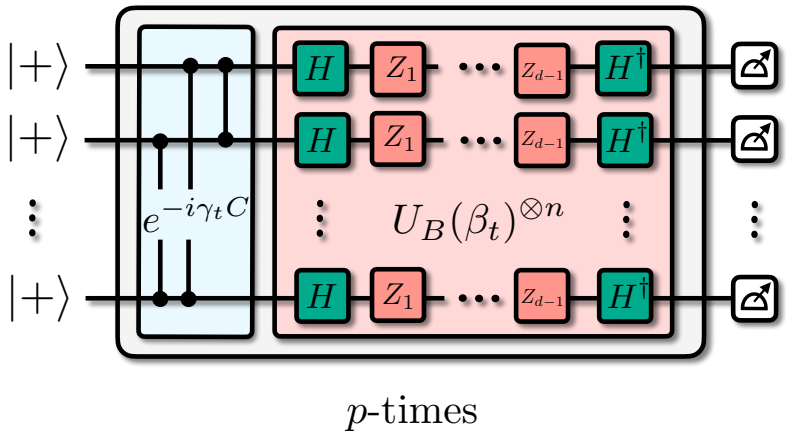


Figure 4.2: Circuit diagram of a Qudit-QAOA with ion trap native gates. The initial state $|+\rangle^{\otimes n}$ is the equal superposition of all qudit levels. Vertical lines indicate the native qudit entangling gates $G(\gamma)$ as defined in Eq. (4.10), they are applied to pairs of qudits $i,j\in E$ to implement $e^{-i\gamma_t C}$ (light blue box). Generalized Hadamard gates H are used to implement a basis transformation into the X-basis (green boxes) where phase shift gates $Z_a(-\beta_a^t)$ are applied consecutively to implement $\prod_{a\neq 0} Z_a(-\beta_a^t)$ (dark red boxes). $H^\dagger$ gates are used to transform back into the computational basis , which completes the implementation of $U_B(\boldsymbol{\beta}_t)^{\otimes n}$ (light red box). This pattern (gray box) is repeated p-times to implement a QAOA of circuit depth p. We omit the angle in phase shift gates and qudit entangling gates in the cartoon for simplicity.

4.3.6 Efficient noise simulation for qudit systems

In the NISQ era noise is one of the main limiting factors for quantum computation. For any algorithm it is thus important to take into account the effect of noise in the computation. For qudit systems the effect of noise has so far been largely unexplored. Mathematically noise can be described as a completely positive (CP) map (also called a quantum channel) $\mathcal{E}(\rho)$ which in Kraus representation can be expressed as

$$\mathcal{E}(\rho) = \sum_{\alpha} E_\alpha \rho E_\alpha^\dagger \text{ with } \sum_{\alpha} E_\alpha^\dagger E_\alpha \leq \mathbb{I}, \tag{4.15}$$

where E_α are non-unique Kraus operators [NC02]. To simulate a general quantum noise channel it is thus required to construct the full $d^N \times d^N$ density matrix and act with Kraus operators on the density matrix. For qudits this approach becomes rapidly intractable due to the increased Hilbert space dimension compared to qubits. In order to numerically study the effect of noise on qudit quantum computation we thus have to resort to alternative means that allow to closely approximate a full density matrix simulation with reduced computational costs.

To this end we will use a single qudit depolarizing noise channel which is given by

$$\mathcal{E}(\rho) = (1 - p_{\text{err}})\rho + \frac{p_{\text{err}}}{d}\mathbb{I}, \tag{4.16}$$

with error probability p_{err}. This implies that with probability $1 - p_{\text{err}}$ the qudit is unchanged and with probability p_{err} it is mapped to the maximally mixed state. The maximally mixed state can be expressed as an average over all generalized Pauli unitaries applied to ρ

$$\frac{1}{d}\mathbb{I} = \frac{1}{d^2}\sum_{p,q=0}^{d-1} Z^p X^q \rho (X^q)^\dagger (Z^p)^\dagger, \tag{4.17}$$

where X^p and Z^q are powers of generalized Pauli matrices as defined in Eq. (4.1) and Eq. (4.2) respectively, see Appx. C.3 for a proof of this equivalence. This allows us to approximate the full depolarizing channel, Eq. (4.16), as a Monte Carlo sampling experiment where the state is unchanged with probability $1 - p_{\text{err}}$ and with probability p_{err} we apply $Z^p X^q$ with integers p, q randomly sampled between 0 and $d - 1$.

So far we have shown that we can approximate a single qudit depolarizing channel as a sampling experiment for density matrices. Therefore, we have not yet gained any efficiency in terms of the computational complexity of the simulation. The reduction in computational complexity is achieved by sampling state vectors rather than density matrices. Using similar arguments as in the derivation of Eq. (4.17) we can show that the sampling experiment can be directly carried out using state vectors as illustrated in Fig. 4.3.

We can thus simulate single qudit depolarization noise using only state vectors with dimension d^N rather than the full density matrix, which has dimension $d^N \times d^N$. In addition, the full noisy Qudit-QAOA simulation can be carried out using only fast vector-vector multiplications. This is since C is diagonal in the computational basis, the term $e^{-i\gamma_t C}$ is also diagonal and can thus be implemented as a simple vector-vector multiplication, this holds also true for the error term Z^p. To implement the non-diagonal quantum unitary $U_B(\boldsymbol{\beta}_t)$, as well as X^q, as a vector-vector multiplication, rather than a matrix-vector multiplication, we can use the Fast-Walsh-Hadamard Transform (FWHT) to transform into the X-basis where the operators become diagonal. This allows to carry out the full noisy simulation using only vector-vector

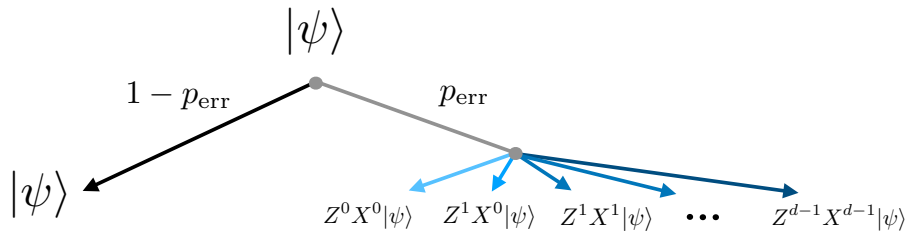


Figure 4.3: Decision diagram for approximating a single-qudit depolarizing channel as a random choice experiment. State vectors sampled according to the diagram approximate the single-qudit depolarizing channel provided a large enough number of samples are used.

multiplications. These techniques allow us to simulate substantially larger qudit systems than a naive approach. For details on the FWHT and a discussion on the total computational complexity see Appx. C.4.

4.4 Performance of Qudit-QAOA for graph coloring

4.4.1 Overlap of optimized Qudit-QAOA with valid graph colorings

To study the performance of the Qudit-QAOA and the effect of noise we consider two different graphs: a triangular graph and a 4-regular graph with six vertices, see inset in Fig. 4.4 and Fig. 4.5 respectively. Both graphs are three-colorable and the ground state energy is thus zero.

For optimizing the variational parameters we use COBYLA, a popular gradient-free algorithm, we find that for the noisy simulation it performs significantly better than the gradient based BFGS algorithm that is typically used in the noise-free settings (see Sec. 1.6 for a discussion on optimization algorithms in VQAs).

We use single qudit depolarizing noise, as discussed in Sec. 4.3.6, for the error probability we use 1%, such that is closely approximates the values of the ion trap quantum computer discussed in Sec. 4.3.5. The depolarizing error that acts on each qudit independently once after applying the entangling layer (i.e. $e^{-i\gamma_t C}$). We omit errors induces by the single qubit gates, since they are an order of magnitude smaller than the errors of the entangling gates.

Fig. 4.4 shows the overlap, of the optimized Qudit-QAOA state with the eigenstates of the Potts model, Eq. (4.4), given by

$$\text{Overlap} = |\langle \boldsymbol{\gamma}^\star, \boldsymbol{\beta}^\star | E_i \rangle|^2, \tag{4.18}$$

where $(\boldsymbol{\gamma}^\star, \boldsymbol{\beta}^\star)$ are optimized parameters and $|E_i\rangle$ are the eigenstates or equivalently colorings of the graph. We evaluate the overlap for circuit depth $p = 1$ and $p = 2$ for a noise-free and noisy simulation. We can see that in both cases a circuit depth of $p = 1$ is not sufficient to prepare the ground state exactly, this is because the entangling gates act on pairs on qudits and the QAOA has thus not "seen" the full graph. For $p = 2$ the noise-free QAOA is able to express the ground state exactly and there are no more contributions of higher energy excited states. In the noisy case, this is not the case and we see that we still observe higher energy contributions due to the noise. In both instances we can observe the sixfold degeneracy of the ground state, the degeneracy arises from the fact that the coloring can be permuted and yields the same energy.

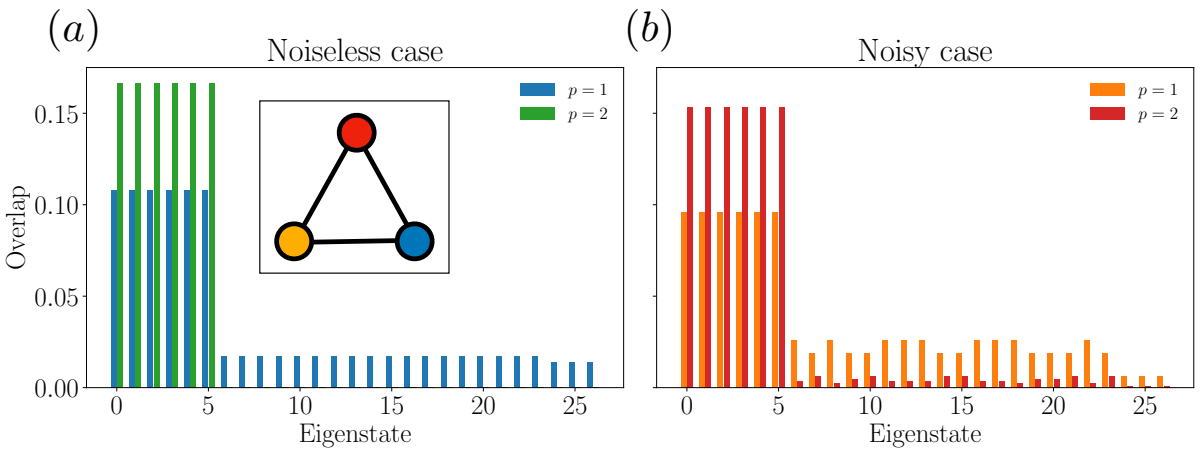


Figure 4.4: Inset shows the triangular graph used for graph coloring. The graph is three colorable, we use red, blue, and yellow as an example. The coloring is sixfold degenerate since the coloring can be permuted $3!$ times. (a) The overlap of the optimized QAOA-Qudit state with the eigenstates of the Potts model, Eq. (4.4), for a noise-free simulation with circuit depths $p = 1$ (blue) and $p = 2$ (green). We can see that for $p = 1$ the ground state cannot be expressed exactly and higher energy excitations are present. For $p = 2$ an exact superposition of the degenerate ground states is prepared and no higher energy contributions can be observed. (b) Under the presence of noise, this is no longer the case and both $p = 1$ (orange) and $p = 2$ (red) have higher energy contributions that reduce the overlap with the ground state. For the noise, we consider a single qudit depolarization error with $p_{\text{err}} = 0.01$ acting on each qudit after the entangling layer. The result is obtained using the Monte Carlo technique described in Sec. 4.3.6. We use 100 random samples.

4.4.2 Optimization performance for cost function

Next, we consider a 4-regular graph with six vertices that is also three colorable. Due to the larger size, a deeper circuit depth is required to prepare the ground state and we will thus be able to study the effects of noise induced by the larger circuit depth. Fig. 4.5 illustrates the change of the cost function, Eq. (4.4), during the parameter optimization. We compare the noise-free with the noisy instance. As anticipated, in the noise-free case we are able to obtain a lower cost function value compared to the noisy case. We can also see that a deeper circuit generally leads to a lower final cost function value, this is not the case in the noisy simulation. There we obtain the lowest cost function value for $p = 2$. The reason for this is twofold: first, the noise can make the parameter optimization more challenging since it alters the optimization landscape in a non-deterministic way. Secondly, the increased circuit depth leads to a proliferation of noise until we are essentially implementing a random sampling. The cost function expectation value for random sampling with equal probability is $|E|/d = 12/3 = 4$ (see Appx.C.5 for details), which is just slightly higher than what is reached for $p = 5$, this is however only an expectation value and does not reveal any information about the underlying probability distribution.

4.4.3 Optimization performance for sampling probability

In fact, a different metric to evaluate the performance of the algorithm might be better suited to better reflect the goal of the algorithm, sampling a valid coloring. We thus consider the

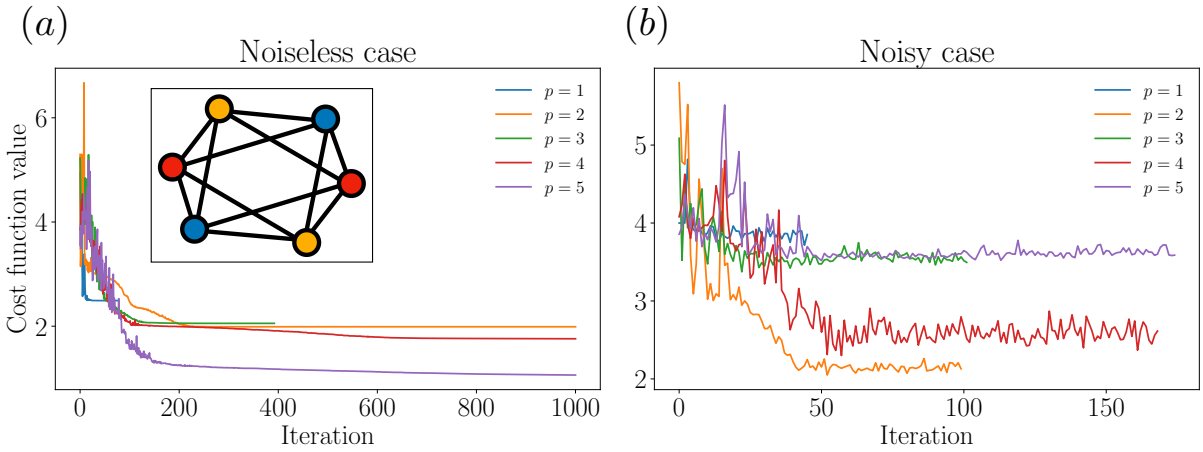


Figure 4.5: Inset shows a 4-regular graph with six vertices that is three colorable illustrated using yellow, blue, and red color. (a) Cost function value during the parameter optimization of the Qudit-QAOA for different circuit depths. (b) Result for a noisy simulation. We use a single qudit depolarizing error $p_{\text{err}} = 0.01$ applied to each qudit after the entangling layer and 100 random samples.

probability for sampling a valid coloring, defined as

$$\text{Prob. of valid coloring} = \sum_{|E_i\rangle \text{ with } E_i=0} |\langle \boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star | E_i \rangle|^2, \tag{4.19}$$

where $|E_i\rangle$ are the eigenstates of the Potts model (Eq. (4.4)). Fig. 4.6 reveals that the probability of sampling a valid coloring in the noise-free simulation increases with circuit depth p and is close to 1 beyond $p = 5$. This shows that the cost function expectation value can be misleading since few higher energy contributions can already shift the value away from the ideal zero value while in fact, the probability of sampling a valid coloring is well above 0.9. For the noisy simulation, we can see that we reach a maximum probability of around 0.5 for $p = 5$, beyond that the probability decreases again due to the noise. This implies that every second sample would be a valid coloring. In contrast, the probability for a valid coloring by randomly coloring the vertices is $6/729 \approx 0.0082$ (the ground state is six-fold degenerate and there is a total of 729 possible ways to color the graph).

In comparison, a commonly used classical algorithm for graph coloring is greedy coloring where the algorithm starts with the first vertex, assigns it the first color, and then moves on to the next vertex, checking if the assignment of the first color would lead to a non-valid assignment given the colored neighbors. If it does, it assigns the next available color. The algorithm continues to do so for each uncolored vertex in the graph. The algorithm is not guaranteed to find a valid coloring or a minimum number of colors for a given graph. The computational complexity of greedy graph coloring is $\mathcal{O}(|V| + |E|)$ and is thus highly efficient. For the small graph sizes that we used in the simulation, the algorithm directly finds a valid coloring.

It is unclear if there are instances where the Qudit-QAOA can outperform greedy coloring, numerical explorations are very limited due to the rapidly scaling Hilbert space and only very small system sizes can be simulated. For a thorough comparison, we will thus have to resort to real hardware which will allow us to study the performance for large graphs, beyond what can be simulated classically. Note that an equivalent implementation of graph coloring on qubits would require additional ancillary qubits to represent the degrees of freedom beyond $d > 2$ as well as additional entangling gates. The Qudits-QAOA is thus ideal for studying this problem on current hardware due to its compact quantum circuit.

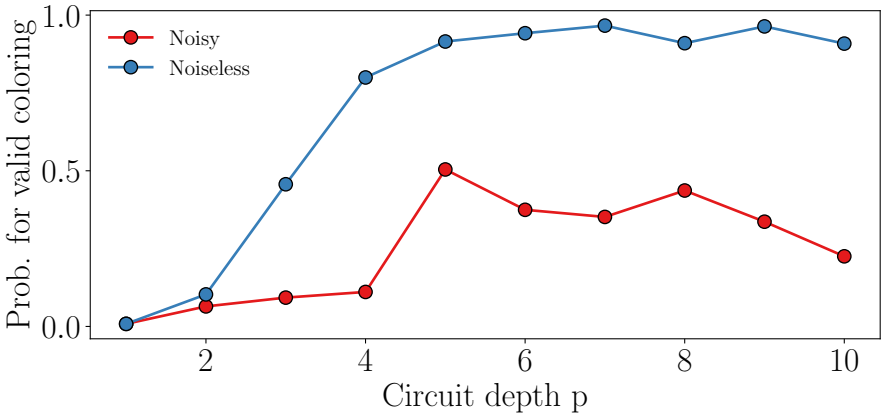


Figure 4.6: Probability of sampling a valid coloring for different circuit depths p, see Eq. (4.19). We can see that the noiseless simulation achieves a significantly higher value compared to the noisy simulation. The noiseless simulation converges to a probability above 0.9 beyond depth $p = 5$, while the noisy simulation reaches a maximum at $p = 5$ of around 0.5. We use the graph shown in Fig. 4.5 (a). For the noisy simulation, we use 100 random samples and an error probability of $p_{\text{err}} = 0.01$ for a single-qudit depolarizing noise.

4.5 Summary and Discussion

The main result of this work is that we propose a generalization of the QAOA to qudit quantum information systems. Qudits are quantum systems with local Hilbert space dimension $d > 2$, which is naturally available on most quantum computing platforms. However, the additional energy levels have so far typically not been utilized for quantum algorithms. We show that a generalization of the QAOA to qudits naturally implements a ground state search for the Potts model on a graph. Furthermore, we show that this can be mapped to graph coloring, a well-known NP-hard problem from classical computer science. There, the aim is to color the nodes of a graph such that no two nodes connected by an edge are the same color.

Even though an implementation of graph coloring on qubits would also be possible, it would require additional auxiliary qubits to encode the colors and a large number of non-local entangling gates. While the number of qubits is somewhat limited on current quantum computers, the required large number of entangling gates for qubits makes this approach entirely unfeasible. However, on qudit quantum information systems this restriction is no longer the case. The energy levels of the qudits can be naturally used to encode the colors without requiring auxiliary qudits as well as a large number of entangling gates. In fact, we show that the number of entangling gates for one QAOA layer is equivalent to the number of edges in the graph. This thus allows for a highly compact circuit that is significantly more expressible than a qubit circuit of the same size.

Next, we show that the unitaries used in the Qudit-QAOA can be implemented on an ion trap quantum computer using native gates. In fact, the entangling gate that is required for the instance of graph coloring is equivalent to a recently proposed native qudit entangling gate [HWG$^+$23]. The proposed circuit is straightforward to implement on ion traps and we believe that it is ideal for testing the capabilities of qudit-based quantum algorithms. We hope that the proposed circuit will be implemented by experimental physicists to explore the capabilities of qudit-based quantum computing beyond what can be simulated classically. Due to the rapidly scaling Hilbert space dimension, a small number of qudits is already enough to

go well beyond what can be simulated. It will be very interesting to explore these realms in the experiment and compare the performance of the Qudit-QAOA to classical algorithms. We believe that this could be an exciting new path for exploring quantum advantage.

To study the anticipated performance of this algorithm on quantum hardware under the effect of noise, we develop a novel Monte Carlo sampling scheme that allows us to study system sizes beyond what would be possible with a naive brute force approach using the density matrix. Our numerical results suggest that using the energy expectation value or cost function value for tracking the performance of the Qudit-QAOA might be misleading. A valid coloring yields a zero cost function value. However, due to the presence of noise, higher energy contributions become unavoidable, leading to a cost function value that deviates significantly from the ideal zero value. This leads us to propose a different figure of merit, namely the probability of sampling a valid coloring. This quantity reveals that the Qudit-QAOA, even in the presence of noise, is able to prepare a valid coloring with a high probability for modest circuit depths of a few QAOA layers. This gives hope that the proposed Qudit-QAOA can be successfully implemented by experimental physicists to explore the capabilities of qudit-based quantum computing. In particular, it will be interesting to explore systems sizes beyond what can be simulated classically. Due to the rapidly scaling Hilbertspace already a small number of qudits will be enough. We believe that this could be an exciting new path for exploring a potential quantum advantage.

CHAPTER 5

Avoiding barren plateaus using classical shadows

5.1 Introduction

Despite the large number of suggested applications, the variational approach encountered also a number of 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 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 lead 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), in order 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 in order 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 by rigorous results and numerical simulations. The structure of the chapter is as follows.

In Sec. 5.2 we introduce the theoretical framework and present our main results. Sec. 5.2.2 introduces the phenomenon of BPs, which dramatically hinders the performance of VQEs. In Sec. 5.2.3 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 algorithm, which allows prevention of the occurrence of BPs by avoiding WBPs.

In Sec. 5.3 we present a bound for the expectation value of the second Rényi entropy in quantum circuits drawn from a unitary ensembles 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. 5.3.1 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 D.1) for the measurement budget require in order to estimate a WBP using classical shadows. Finally, in Sec. 5.3.2 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. 5.4, we explore how BPs and WBPs emerge at different stages in the VQE optimization and perform a systematic performance analysis. Next, in Sec. 5.4.1 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. 5.4.2 we explore the performance of the WBP-free VQE algorithm in different settings for the Heisenberg model on a chain. Finally, in Sec. 5.4.3, 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 D.5 which exhibits a nearly maximally entangled ground state.

Finally, in Sec. 5.5 we summarize our results, discuss their implications, and outline open questions.

5.2 Avoiding barren plateaus in variational quantum optimization

5.2.1 Variational quantum eigensolver

We focus our study on k-local Hamiltonians H, defined as 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 a 2-local Hamiltonian from the many-body physics is provided by the

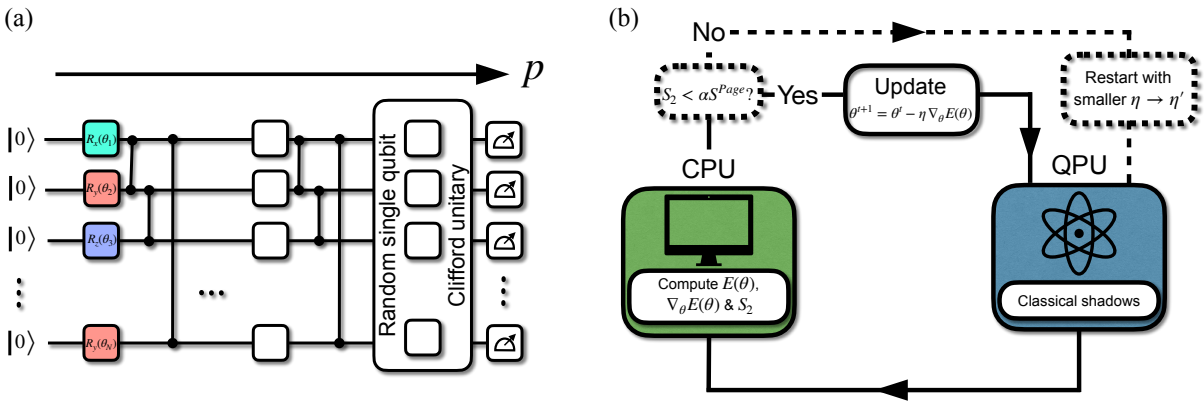


Figure 5.1: (a) Illustration of the variational quantum circuit $U(\boldsymbol{\theta})\,|0\rangle$ that is considered in the main text followed by the shadow tomography scheme [HKP20]. The variational circuit consists of alternating layers of single-qubit rotations represented as boxes and entangling 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 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 D.1.

Heisenberg (XXX) model subject to a magnetic field

$$H_{XXX} = \sum_{i,j\in V_{\mathcal{G}}} J\left(\sigma_i^z\sigma_j^z + \sigma_i^y\sigma_j^y + \sigma_i^x\sigma_j^x\right) + h_z\sum_{i=1}^{N}\sigma_i^z, \tag{5.1}$$

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 one-dimensional (1D) chain with periodic boundary condition, 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 condition, 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. (1.15).

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. 5.1 (a). For our example of a 2-local Hamiltonian on the 1D chain, the required measurements include the value of the σ^z operator on all sites along with the $\sigma_i^a\sigma_{i+1}^a$ expectation values of all $i = 1,\dots N$ (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

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\nabla_{\boldsymbol{\theta}}E(\boldsymbol{\theta}), \tag{5.2}$$

where η is the *learning rate*, which controls the update magnitude (see Sec. 1.6 for details). 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. 5.1 (b) by solid lines. Following the initialization of the variational angles $\boldsymbol{\theta}$, that 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.

5.2.2 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]:

$$\mathrm{Var}[\partial_{i,l} E(\boldsymbol{\theta})] \sim \mathcal{O}\left(\frac{1}{2^{2N}}\right). \tag{5.3}$$

[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$^+$18] (see Appendix D.2 for details and the definition of a t-design). To understand the different circuit depth at which BPs are encountered, the authors in Ref. [CSV$^+$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 the 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, that 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 $p \sim \mathcal{O}(\mathrm{poly}(N))$ 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 initializations strategies that avoid BPs. Until now, several

BP-free initializations were considered in the literature. Ref. [GWOB19] suggests to initialize 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 in fact one and the same. Avoiding entanglement-induced BPs is equivalent to avoiding BPs for local cost functions, the details are presented in Sec. 5.3.

Experimentally probing a BP is a hard task. The estimation of the exponentially small gradient in Eq. (5.3) 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 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. 5.4, 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 is discussed in the algorithm proposed below.

5.2.3 Weak barren plateaus and improved algorithm

In order 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,

$$S^{\text{Page}}(k, N) \simeq k \ln 2 - \frac{1}{2^{N-2k+1}}, \tag{5.4}$$

where we explicitly used that the Hilbert space dimension of region 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 $|\text{GS}\rangle$, the choice can however be more informed to help avoid large entanglement local minima, see Sec. 5.3.

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, the shadow tomography replaces the measurements performed in the computational basis with a more general measurements, that turns out to be sufficient for reconstructing linear and non-linear function 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 5.1 (b). In essence, we suggest to use 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 D.1. Access to the second Rényi entropy allows prevention of the appearance of WBPs not only at the initialization step, but throughout the optimization cycle. The explicit algorithm works as follows.

Algorithm 2 WBP-free optimization with classical shadows

Choose α, default is $\alpha = 1$ ▷ see Sec. 5.3.1 for details
Choose $\boldsymbol{\theta}$ such that $S_2 < \alpha S^{\text{Page}}(k, N)$
Choose learning rate η
repeat ▷ see Appendix D.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. 5.3. 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 Sec. 5.4 we discuss the optimization process in greater details. In particular, we 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.

5.3 Weak barren plateaus and initialization of VQE

5.3.1 Definition and relation to barren plateaus

As mentioned in the 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 the following.

Theorem 2. *(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*

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

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 a proof in Appendix D.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. In fact, entanglement provides an intuitive picture for the emergence of BPs and its 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 as follows.

Definition 3. *(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 $S_2 \geq \alpha S^{\text{Page}}(k,N)$ 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 subregion with k spins contains all 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 actually interested only in avoiding a BP. For this, WBPs provide a powerful tool, since the following holds.

Corollary 3.1. *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 2. □

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

$$T \geq \frac{4^{k+1} \operatorname{tr} \rho_A^2}{\epsilon^2 \delta}, \tag{5.5}$$

where ϵ is a 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 D.1, see Eq. (D.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., extend of a subregion with entropy close to the maximal value) in the circuit, see Fig. 5.2 (a). Such a scrambled region grows with every consecutive entangling layer W_l (see Eq. (1.15)). 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. 5.2 provides a numerical illustration for the Corollary 3.1 stated above. We use the hardware-efficient circuit, presented in Eq. (1.15), 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.

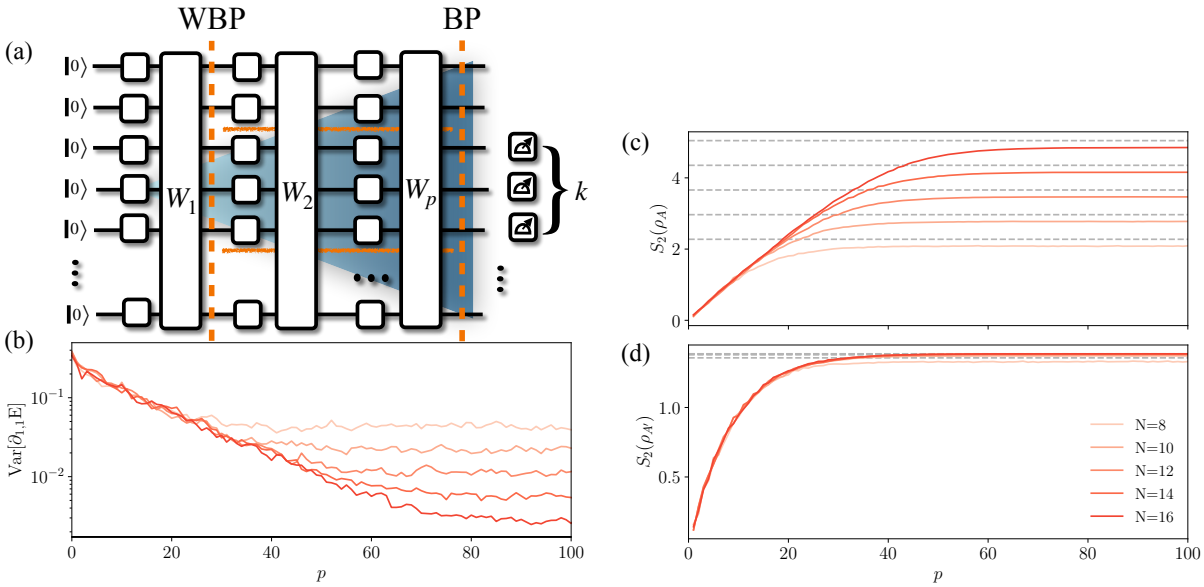


Figure 5.2: (a) Sketch of the circuit, where the blue color shows the scrambling lightcone. The lightcone first extends over k qubits, where the WBP occurs, and for larger circuit depths extends to the full system size where the BP occurs. (b) The saturation of the gradient variance $\mathrm{Var}[\partial_{1,1}E]$ 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 is 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. (5.1), looks similar.

Finally, we address the effects of the control parameter α, that enters in Definition 3 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 $|\mathrm{GS}\rangle$ is available, it can be used to set a smaller value of α. If, for instance, the ground state is only weakly entangled, a choice of $\alpha \ll 1$ may be appropriate. In this way Algorithm 1 in Sec. 5.2.3 can also help in avoiding convergence to highly entangled local minima. We discuss this in more detail in Sec. 5.4.2.

5.3.2 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. (1.15) as $\theta_l^i \in \epsilon_\theta[-\pi,\pi)$, where $\epsilon_\theta \in [0,1)$ is the control parameter. This strategy allows the onset of the BP to be delayed to arbitrary circuit depths by tuning ϵ_θ accordingly.

Similarly, it allows the onset of WBPs to be delayed. 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

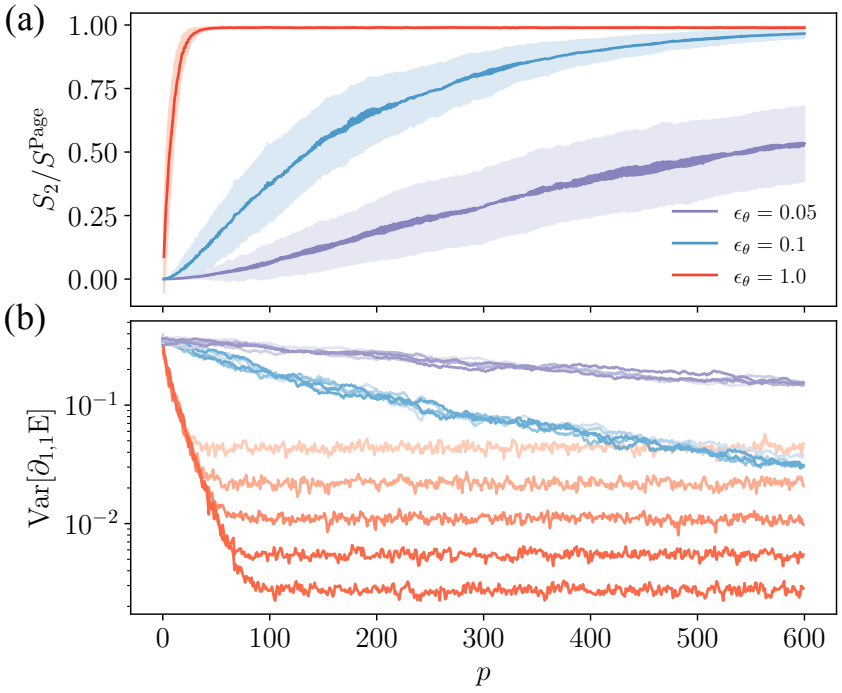


Figure 5.3: (a) Decreasing parameter ϵ_θ from 1 slows down the growth of the second Rényi entropy with the circuit depth p. The chosen region contains two qubits. (b) The encounter of BP in the variance of the gradient of the cost function is visible only for the case $\epsilon_\theta = 1$, and it is 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. 5.2. Data is averaged over 100 random instances, variance is for the local term $\sigma_1^z \sigma_2^z$.

entangling gates in W_l do not create any entanglement [since the CZ gates used in Eq. (1.15) are diagonal in the computational basis], leaving $|0\rangle$ invariant. Note that, for example, the *identity block* initialization, proposed by [GWOB19] works in a similar way in that the unitary is constructed such that it also implements the identity and one is equally left with the zero state.

In Fig. 5.3 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.

5.4 Entanglement control during optimization

5.4.1 Bounding entanglement increase at a single optimization step

In Sec. 5.2 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. 5.1 (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

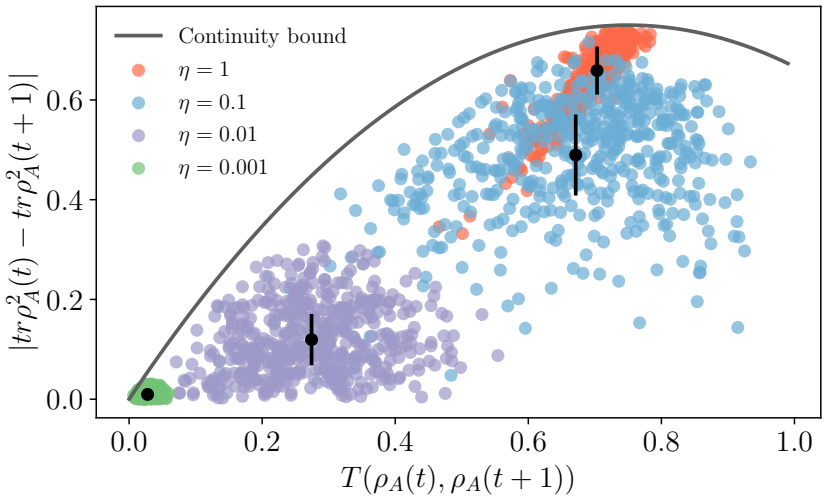


Figure 5.4: We numerically illustrate the continuity bound Eq. (5.6) and its relation to the learning rate η for $t = 0$, i.e. at the beginning of the optimization schedule. This shows that one should be careful with the choice of the learning rate since a large learning rate leads to a big change in the trace distance and 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 is averaged over 500 random instances.

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]

$$\left|\operatorname{tr}\rho_A^2(t+1) - \operatorname{tr}\rho_A^2(t)\right| \leq 1 - (1 - T_A(t))^2 - \frac{T_A^2(t)}{2^k - 1}, \tag{5.6}$$

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 pertubatively close (see Appendix D.4 for details), as measured by the trace distance, one can show that

$$T(\rho_A(t+1), \rho_A(t)) \lesssim \sqrt{\frac{\eta^2}{4}(\nabla_{\boldsymbol{\theta}} E)^T \mathcal{F}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} E}, \tag{5.7}$$

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 (5.6)-(5.7) imply that the learning rate η can be used to limit the maximal possible change of the purity. 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. 5.4. 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. (5.6) is not proven to be tight, we observe that points corresponding to the extreme learning rates closely approach the theoretical line.

Using Eq. (5.7), 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 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.

5.4.2 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. 5.5 (a)-(c), we see that the energy expectation value in Fig. 5.5 (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 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. 3 being smaller than one may be beneficial. For instance, setting $\alpha = 0.5$ could help avoiding the suboptimal local minima characterized by large entanglement, see gray dashed line in Fig. 5.5 (b). Note that the large gradient persistent after many iterations for the blue line in Fig. 5.5 (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 ground state energy (violet line in Fig. 5.5 (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 within the considered number of iterations and 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.

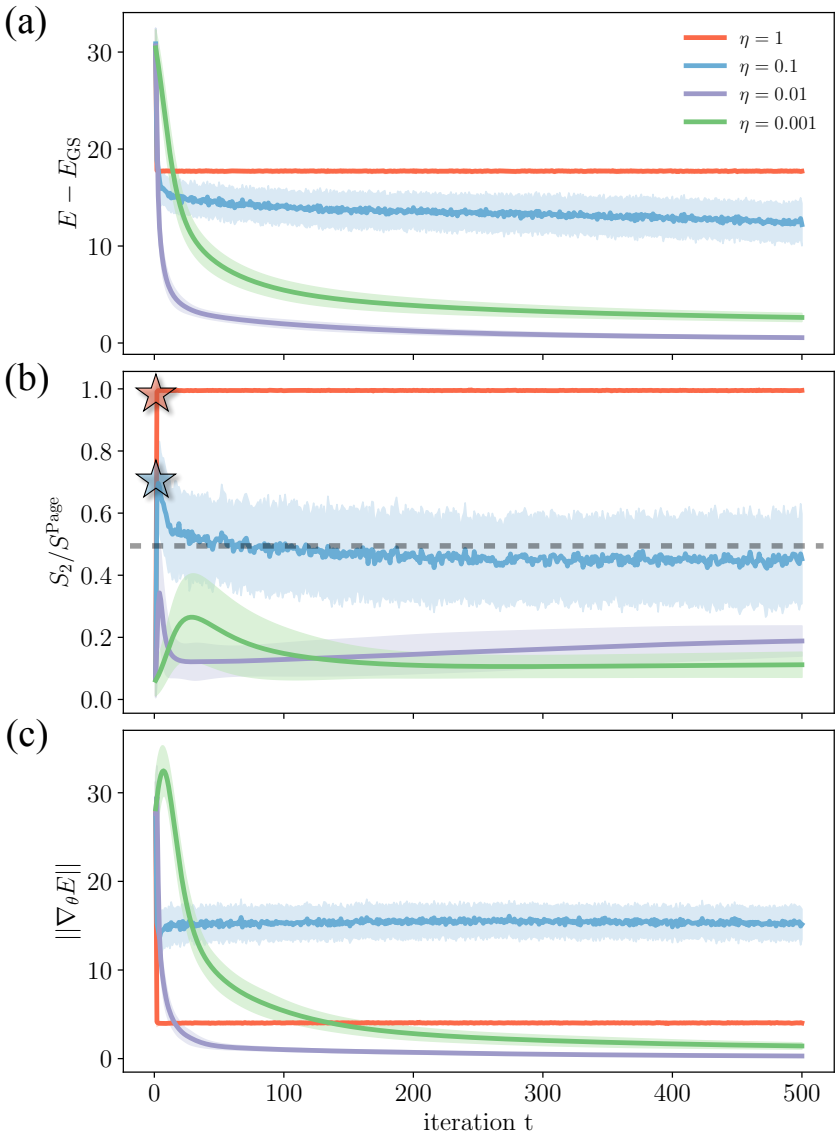


Figure 5.5: (a-c) The application of the proposed algorithm to the problem of finding the ground state of the Heisenberg model. For large learning rates $\eta = 1$ and 0.1 (red and blue lines) the optimization gets into a large entanglement region as is shown in (b), indicated by colored stars, forcing the restart of the optimization with smaller value of η. For $\eta = 0.01$ the algorithm avoids large entanglement region 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 $S_2/S^{\mathsf{Page}}(k, N) \approx 0.246$. (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. (1.15)) 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 gray dashed line, see the last paragraph of Sec. 5.3.1 for a discussion on the choice of α. Data is averaged over 100 random instances.

5.4.3 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 $\alpha = 1$. While this restriction does place a limitation on the entanglement generated by the circuit for a region of k qubits, it does not imply classical simulatabilty 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. 3 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 [Sch11a]. The 1D Heisenberg model, considered in the previous subsection, is an example for such a system.

The Heisenberg model, however, can be made hard to simulate classically by considering a random-graph geometry illustrated in Fig. 5.6 (a), instead of a 1D chain. This leads to nonlocal 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. 5.6 (b)-(c) reveal 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. D.1, 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 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 SYK model [Kit15] that features a volume-law entangled ground state [HG19]. In Appendix D.5 we illustrate that our method is also successful in preventing the BP occurrence and results in finding the SYK ground state.

5.5 Summary and Discussion

The main result of this work is the introduction of the concept of WBPs, which in essence provides an efficiently detectable version of BPs. In particular, we propose to use the classical shadows protocol to estimate the second Rényi entropy of small subregions that are independent of system size. If these subregions avoid nearly maximal entanglement – a condition sufficient

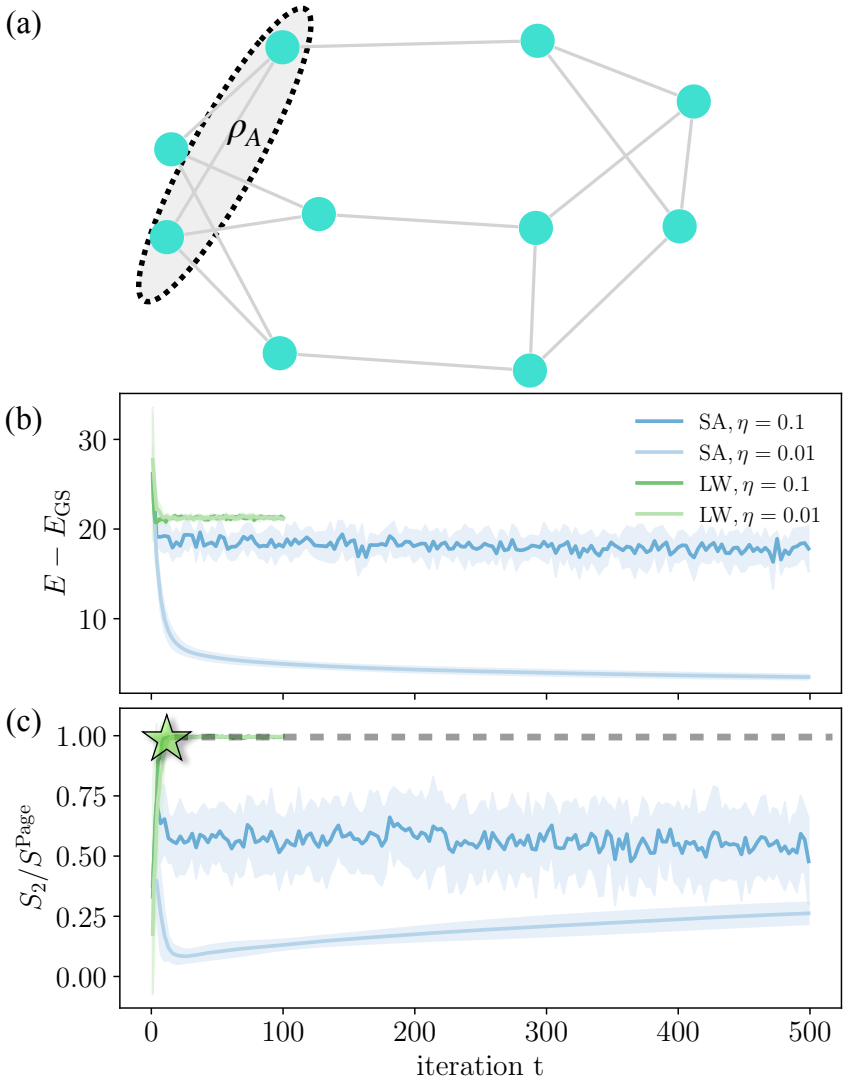


Figure 5.6: Application of our algorithm to the problem of finding the ground state for the Heisenberg model on a 3-regular random graph depicted in (a). Panel (b) shows the energy as a function of GD iterations t and panel (c) illustrates the second Rényi entropy of two-spin region A with $k = 2$ shown in panel (a). Since the interactions are now nonlocal 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 consider (green star). In contrast, SA avoids the WBP for both learning rates. Good performance and further convergence in the local minimum is only achieved through a smaller learning rate of $\eta = 0.01$. We use a system size of $N = 10$ and a random circuit from Eq. (1.15) with circuit depth $p = 100$. Data is averaged over 100 random instances.

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.

In order 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 VQEs, they readily extend to other variational hybrid algorithms, such as quantum machine learning [BLSF19b, HCT+19b, SBSW20], quantum optimization [FGG14b, SS21c, Har21], 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 [LJG+21]. Our results indicate that entanglement, in addition to playing a crucial role for circumventing BPs at the launch of the VQE, is also important for achieving a good optimization performance. In addition, our heuristic results in Sec. 5.4 suggest that postselection 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 postselection 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 a 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 on current NISQ machines using the proposed shadows protocol.

CHAPTER 6

Summary and Outlook

6.1 Summary of thesis content and open research questions

Quantum computation is promising revolutionary advancements for computation but it still requires significant hardware and algorithmic development to make useful quantum computation a reality. Variational quantum algorithms (VQAs), are well suited for dealing with noise in the current hardware generation due to their reduced circuit depth. However, VQAs present their own challenges, namely the complexities of the optimization landscape and the variational nature, resulting in limited performance guarantees. In this thesis, we have presented the substantial progress that we have made towards resolving these obstacles.

Chapter 2 introduces an initialization technique for the Quantum Annealing Optimized Algorithm (QAOA), called Trotterized Quantum Annealing (TQA). This counters the common issue of convergence to high energy local minima in the optimization landscape that limits algorithm performance. TQA matches the best performance achieved from exponentially many random initializations. Its success in experiments and simulations suggests it may become a standard QAOA initialization technique. In the future, it will be particularly interesting to further explore the connection between Quantum Annealing and the QAOA at finite circuit depth from an analytical perspective which may lead to further improved initialization techniques.

Chapter 3 further deepens our understanding of the optimization landscape's properties and efficient exploration strategies. We introduce Transition States (TS), as a tool for systematic landscape exploration, and a greedy optimization strategy that guarantees performance improvement. These insights could shape future work on performance bounds and QAOA's potential to outperform classical computation.

Chapter 4 explores another critical aspect of VQAs: circuit expressibility. In particular, extending VQAs to higher-dimensional quantum systems or "qudits" to utilize a higher dimensional Hilbert space for computation. We illustrate how the QAOA can be extended to qudits and used for graph coloring, an NP-hard problem from classical computer science. A comparable implementation on a qubit based quantum computer would require significantly more qubits and entangling gates. Furthermore, we show that the Qudits-QAOA can be implemented on an ion trap quantum computer using novel qudit entangling gates. This research may serve as a foundation for a first quantum algorithm implementation on qudits, offering exciting opportunities to probe system sizes beyond classical tractability.

The concluding Chapter 5 addresses barren plateaus, regions in parameter spaces with vanishing gradients that prevent parameter optimization for VQAs in general (not only the QAOA). We demonstrate that measuring reduced density matrices with the help of classical shadows can diagnose barren plateau regions. An algorithm summarizing our findings allows initialization in barren plateau-free regions and avoids encountering such regions by tracking entanglement entropy. The current generation of quantum hardware can now leverage this technique, offering a standard method for barren plateau avoidance. A particularly interesting topic for future extension of this work is to explore the connection between circuit universality and the entanglement structure generated by the quantum circuit.

6.2 Outlook for the future of quantum computing

Quantum computing is a rapidly evolving field, new technological and algorithmic advances are announced every year. For example, while the early stages of quantum computing hardware experiments were dominated by superconducting quantum hardware, recently ion trap and neutral atom quantum computers have emerged as serious contenders. As these devices are further scaled up and improved we will soon soon routinely operate in a regime where the quantum circuits implemented by these machines can no longer be simulated classically.

An early example of this is a recent work by IBM where they used their superconducting device to simulate the Ising model on a 2D hardware native grid. This allowed them to utilize the full 127 qubits, despite the limited connectivity [KEA+23]. In this work, they explore a system size that may be on the edge of what can still be simulated using state-of-the-art techniques such as tensor networks. A claim that was however subsequently disproved. While this model is still too simple to be practically relevant it gives hope that in future experiments more complicated models can be explored. Furthermore, this result highlights that the field is making great progress and we are rapidly approaching system sizes with the potential of a quantum advantage.

While at this time it remains unclear if near-term quantum computing without error correction will be able to achieve a quantum advantage, fault-tolerant quantum computing (FTQC) stands on a thorough theoretical foundation that promises to achieve this goal. In particular, we are not aware of any physical laws prohibiting us from building FTQC and implementing algorithms with proven speed-ups. In this direction we have seen first promising results, for example, Google implemented a surface code experiment where they showed that they were able to successfully suppress quantum errors by increasing the code distance [AAA+23]. This marks a first important step in making FTQC a reality.

We believe that this rapid pace of development will continue into the future. A defining question to answer will be if NISQ machines are able to achieve a quantum advantage. If it becomes clear that they are not sufficient, we can expect to see a so-called "quantum winter", a time period of greatly decreased funding and interest in the field. Many technologies have seen their own respective "winter", a recent example is machine learning which only after decades of relative quietness saw its recent burst after great algorithmic and hardware improvements.

Regardless of whether the NISQ era sees a quantum advantage or a quantum winter, the long-term future of quantum computing remains bright. We stand on the cusp of a new era where the boundaries of computation can be pushed far beyond current limits. With continuous improvements in quantum algorithms, quantum error correction, and quantum

hardware, we may see quantum computers tackling complex problems in fields ranging from materials science to cryptography. As quantum computing matures we expect a profound impact across a broad range of academic disciplines and industry applications. The future may hold challenges, but with perseverance and continuous innovation, the quantum age may just be over the horizon.

APPENDIX A

Further numerical results for different graph ensembles and discussion on optimal TQA time

A.1 Optimization landscape for different graph ensembles

We start by reviewing all graph ensembles used in the main text and Appendices. In particular, we focus on symmetries that allow to reduce the space of QAOA parameters.

3-regular unweighted graphs represent the graph ensemble considered in the main text. Each vertex is connected exactly to three other vertices chosen at random. In order to sample graphs from this ensemble we use the `networkx` Python package [HSSC08]. For 3-regular unweighted graphs the space of variational parameters can be restricted using the fact that the classical Hamiltonian has integer eigenvalues (thus γ_i are defined modulo π) and that shifting any of angles β_i by $\pi/2$ is equivalent to a spin flip of H_C that has no effect [ZWC+20]. This allows to restrict $\beta_i \in [-\frac{\pi}{4}, \frac{\pi}{4})$ and $\gamma_i \in [-\frac{\pi}{2}, \frac{\pi}{2})$, and is reflected in the definition of distance in Eq. (2.1) in the main text.

3-regular weighted graphs are characterized by presence of random weights w_{ij} assigned to each edge $\langle i, j\rangle$. These weights are chosen to be $w_{ij} \in [0, 1)$. Presence of random weights does not allow to restrict the domain of γ_i angles as before, though restriction $\beta_i \in [-\frac{\pi}{4}, \frac{\pi}{4})$ still works. Therefore the analogue of Eq. (2.1) for this and other weighted ensembles reads $d^{(w)}_{\vec{\gamma},\vec{\beta}} = \sum_{i=1}^{p}(|\beta_i - \beta_i^*|_{\frac{\pi}{2}} + |\gamma_i - \gamma_i^*|)$.

*E*rdős-Rényi graphs represent a random graph ensemble where two edges are connected on random with a fixed probability, chosen to be $q = 0.5$. In contrast to above examples, the fixed value of q implies that edge connectivity increases with number of vertices as qN. Erdős-Rényi graphs exhibit the same symmetries as 3-regular unweighted graphs.

The presence of an unbounded region of parameters γ_i in the weighted graph ensemble represents an additional challenge in visualizing the QAOA optimization landscape and choice of initialization parameter. In order to explore the importance of large values of $|\gamma_i|$, we consider the sequence of enlarged intervals $\gamma_i \in [-k\frac{\pi}{2}, k\frac{\pi}{2})$ with $k = 1, 2$. Figure A.1 shows

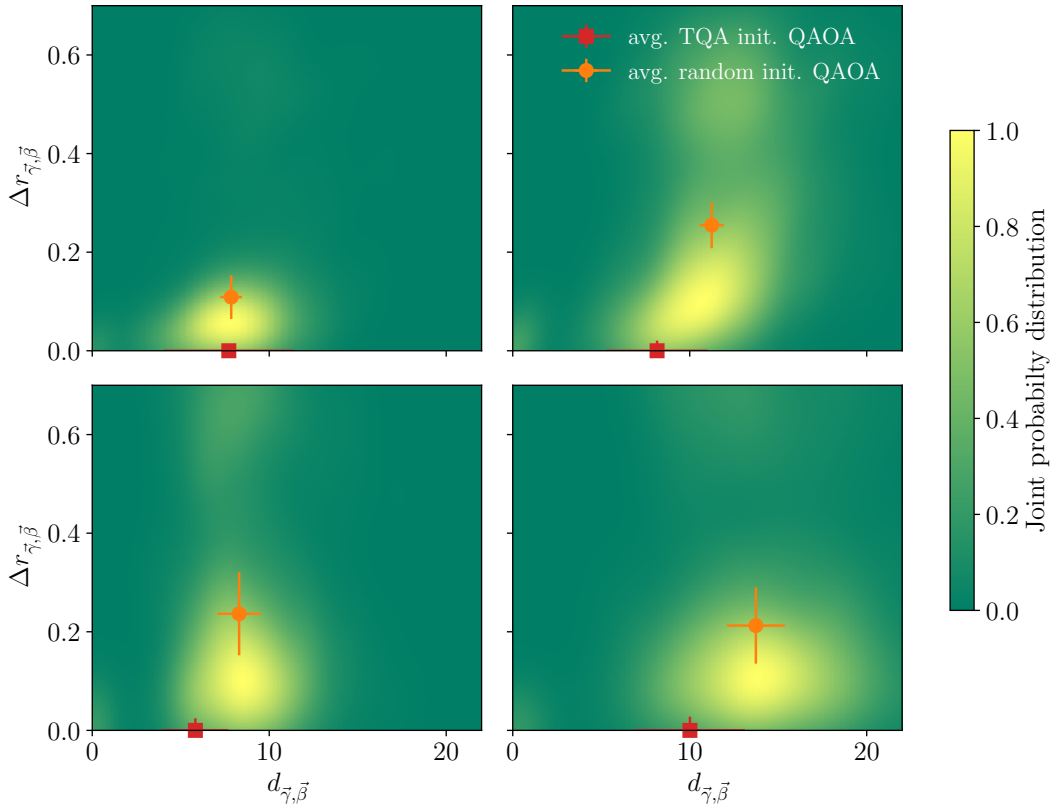


Figure A.1: Comparing the joint probability distribution of the distance to the global minimum in parameter space $d_{\vec{\gamma},\vec{\beta}}$ and in terms of approximation ratio $\Delta r_{\vec{\gamma},\vec{\beta}}$ for weighted 3-regular (top) and Erdős-Rényi graphs with edge probability 0.5 (bottom) reveals that the distribution is dependent on the initialization interval for weighted 3-regular graphs. We initialize the parameters for $k = 1$ (left) and $k = 2$ (right) and observe that for weighted 3-regular graphs the enlarged interval leads to an increased spread of the local optimas in $\Delta r_{\vec{\gamma},\vec{\beta}}$ (yellow region). The spread in $\Delta r_{\vec{\gamma},\vec{\beta}}$ for Erdős-Rényi graphs remains largely unaffected, as expected from the symmetry considerations. Similarly to Fig. 2.2, red squares correspond to the QAOA minimum achieved from TQA initialization (shifted from small negative values of $\Delta r_{\vec{\gamma},\vec{\beta}}$ to zero for improved visibility), orange dots correspond to the average performance of random initialization. Data is for 50 random graphs with $N = 10$ and $p = 5$.

the joint probability distributions similar to Fig. 2.2. We see that for 3-regular weighted graphs the enlarged initialization interval $k = 2$ leads to a concentration of local optima further away from the global solution compared to the $k = 1$ interval. When we repeat the same analysis for Erdős-Rényi graphs, we observe that $\Delta r_{\vec{\gamma},\vec{\beta}}$ is unaffected by the enlarged $k = 2$ interval. This numerically confirms the symmetry considerations from above and allows us to restrict $\vec{\gamma}$ to the $k = 1$ interval in all further analysis. For unweighted graphs such restriction relies on symmetry, and for weighted graphs this is motivated by the fact that an extended region of γ_i worsens the performance of random initialization in the QAOA.

A.2 Optimal time for TQA

Below we discuss the dependence of the optimal time step δt of the TQA algorithm on the graph ensemble. An analytical *upper bound* on the number of Trotter steps p needed to approximate the time evolution with precision ϵ in terms of operator trace distance was obtained

in Ref. [BACS07]. Translating this bound into the scaling of δt we obtain $\delta t \propto 1/(||H_C||_F N)$, where $||H_C||_F$ is the Frobenius norm of the classical Hamiltonian. This norm exponentially diverges with N, suggesting very small values of δt at large system sizes. This is not surprising, since the bound of Ref. [BACS07] operates on the distance between two many-body unitary operators. In contrast, the performance of the TQA algorithm is studied using the approximation ratio that quantifies how close the expectation value of the local observable H_C, is to the ground state energy.

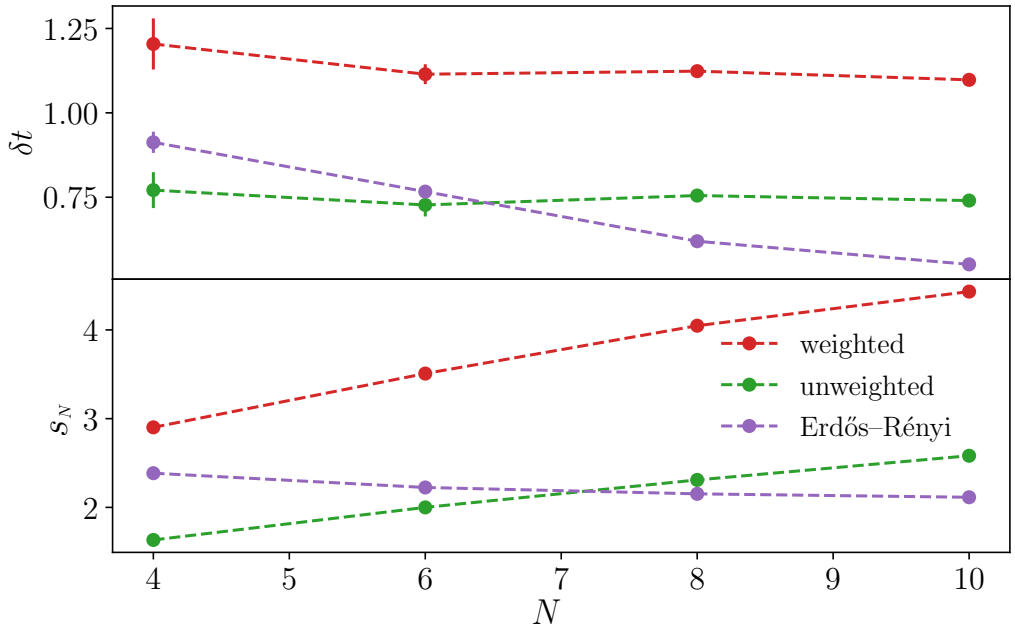


Figure A.2: (Top) Optimal time time step of TQA evolution δt is largely independent of system size and scales qualitatively similar to Eq. (A.1) shown in the bottom panel.

The effect of Trotterization on local observables was considered in Ref. [HHZ19]. This work conjectured the existence of a finite value of the time step of order one, at which the discretization of time evolution fails to approximate the local observables. This value of the time step may be related to the convergence radius of the Baker-Campbell-Hausdorff series expansion, which is governed by the norm of the classical Hamiltonian and its commutator with H_B. Phenomenologically, the Frobenius norm divided by the square root of Hilbert space dimension and problem size N,

$$s_N = \frac{N 2^{N/2}}{||H_C||_F}, \tag{A.1}$$

is expected to be N-independent in the thermodynamic limit.

Figure A.2 compares the dependence of δt on the system size with the phenomenological scaling s_N defined in Eq. (A.1). We observe that the expression s_N qualitatively matches the numerical scaling that we observe for δt between different graph ensembles. In particular, the value of the time step is largest for weighted 3-regular graphs that are expected to have the smallest norm of the classical Hamiltonian. However, s_N fails to capture δt quantitatively, highlighting the need to develop a better analytical understanding of the point that governs the phase transition from localization to quantum chaos for local observables according to Ref. [HHZ19].

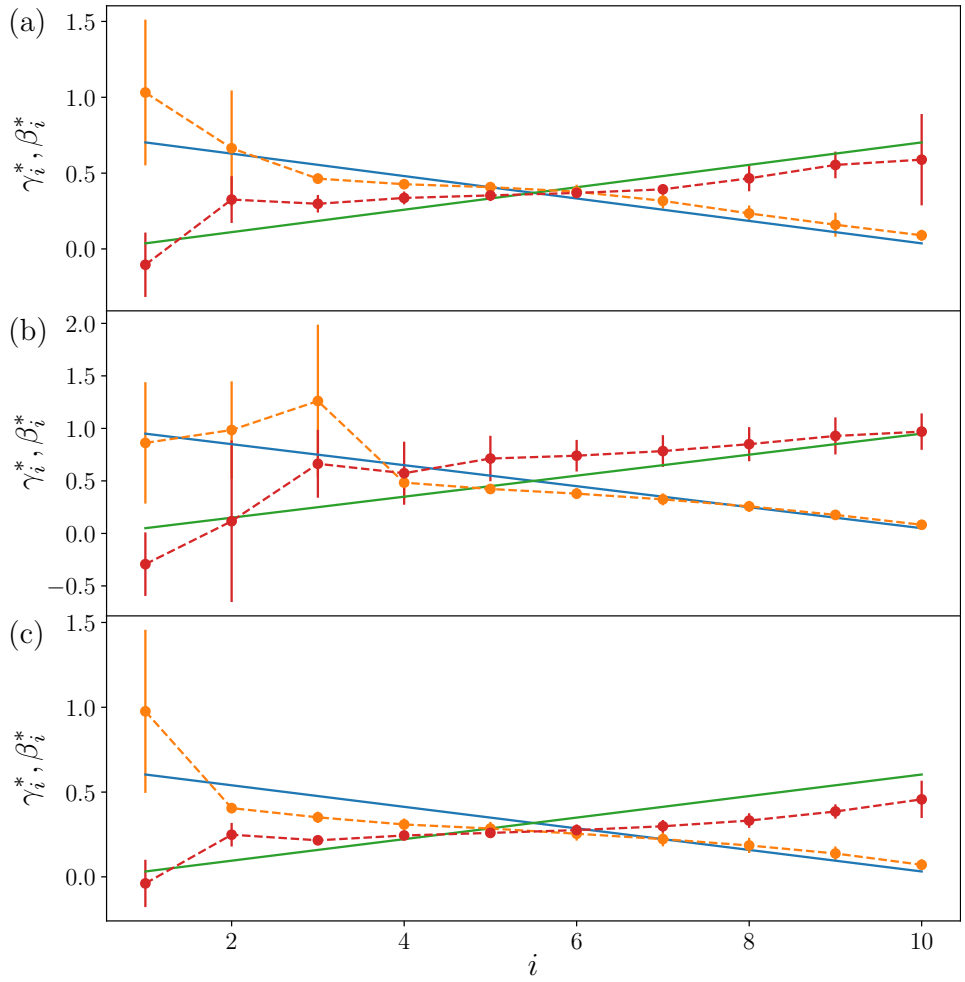


Figure A.3: Converged parameters $\vec{\gamma}^*$ (red) and $\vec{\beta}^*$ (orange) show only slight alterations from the TQA initialization indicated by the green and blue lines respectively. The QAOA optimization modifies parameters at small i, while they remain TQA-like in the rest of the protocol. The results were averaged over 50 random unweighted 3-regular graphs (a), weighted 3-regular graphs (b) and Erdős-Rényi graphs (c), all data is for $p = 10$ and $N = 10$.

A.3 Patterns in optimized parameters

The QAOA is inspired by TQA and is thus universal for $p \to \infty$. However, for finite p the converged QAOA parameters also display stark similarity to a QA protocol which was noticed in some earlier works [ZWC+20, Cro18]. In Fig. A.3 we compare the TQA initialization and final QAOA parameters. The QAOA parameters show only slight alterations at the beginning of the protocol and remain close to their original values throughout the rest of the protocol. This holds true for the three graph types that we considered in our analysis. In addition, the small variation between optimal parameters for different graph instances is in line with the concentration of the QAOA landscape demonstrated analytically at low p in Ref. [BBF+18b].

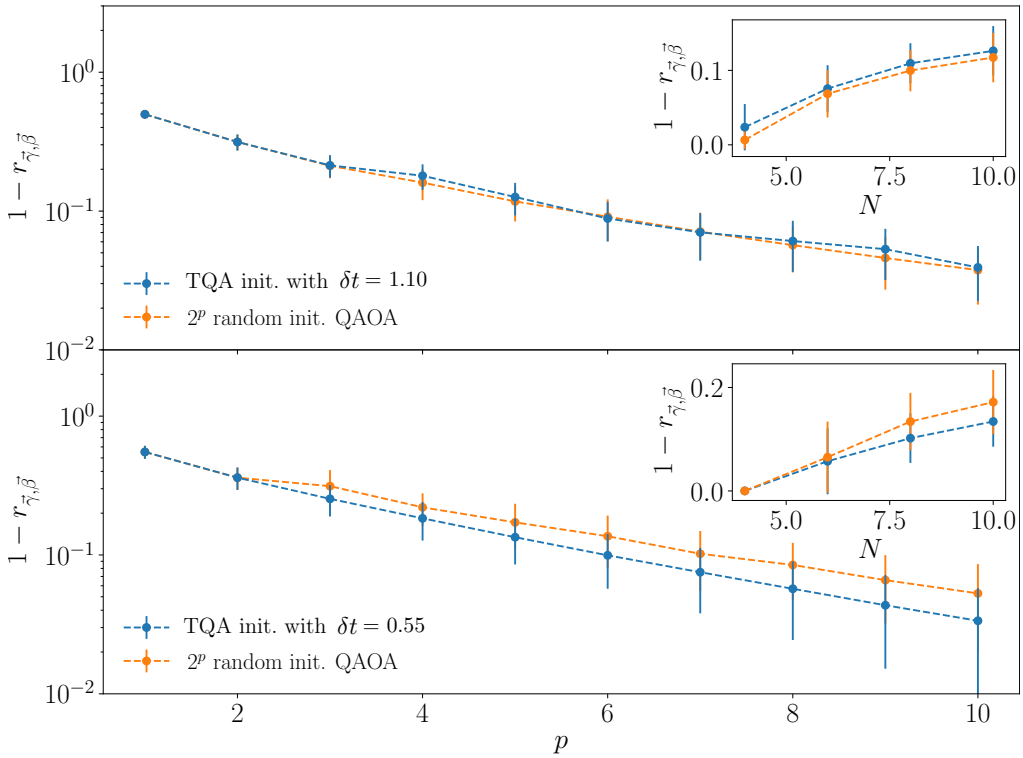


Figure A.4: TQA initialization leads to the same QAOA performance as the best of 2^p random initializations for both weighted 3-regular graphs (top) and Erdős-Rényi graphs (bottom). We average the results over 50 graph realizations, the main plot was obtained for system size $N = 10$, inset is for circuit depth $p = 10$.

A.4 Random vs TQA initialization for other graph ensembles

In addition to the unweighted 3-regular graphs, discussed in the main text, we also test TQA initialization on weighted 3-regular graphs and Erdős-Rényi graphs. We find that TQA initialization yields the same performance as the best of random initializations for weighted 3-regular graphs, see Fig. A.4. For Erdős-Rényi, TQA initialization even outperforms the best of 2^p random initializations.

APPENDIX B

Proof of transition state properties and details for greedy algorithm

B.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:

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

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

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,

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

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

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

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, SS21b]. 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:

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

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

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

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

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

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

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

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:

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

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

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

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

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

B.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 4 (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:*

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

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 1 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 [LJG$^+$21] (if there exists some combination of variational angles, such that its changes do not influence the quantum state, it leads to vanishing eigenvalue of Hessian). Analogously, in the case of the Hessian at the TS of QAOA$_{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.

B.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:

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

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. (B.15). They read

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

Our goal is to prove that given a local minimum $\boldsymbol{\Gamma}_{\mathsf{min}}^p = (\beta_1^\star, \ldots, \beta_p^\star, \gamma_1^\star, \ldots, \gamma_p^\star)$ for a QAOA$_p$ the set of $2p+1$ points

$$\begin{aligned} \boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}(l,k) = (&\beta_1^\star, ..., \beta_{l-1}^\star, 0, \beta_l^\star, ..., \beta_p^\star, \\ &\gamma_1^\star, ..., \gamma_{k-1}^\star, 0, \gamma_k^\star, ..., \gamma_p^\star), \end{aligned} \tag{B.17}$$

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 $\boldsymbol{\Gamma}_{\mathsf{min}}^p$ being a local minimum (i.e. derivatives are vanishing). For the derivatives over newly introduced angles using Eq. (B.15), we see that

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

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

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

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

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

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

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

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

as desired. For case **(ii)** we have that

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

which handles the second special case:

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

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

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

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.

B.2.2 Cost function Hessian

We now proceed with the study of the Hessian for each of the stationary states in the set $\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}(l,k)$ 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:

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

where $H(\boldsymbol{\Gamma}_{\mathsf{min}}^{p}) \in \mathbb{R}^{2p\times 2p}$, $v(l,k) \in \mathbb{R}^{2p\times 2}$ and, $h(l,k) \in \mathbb{R}^{2\times 2}$. It is important to note that the determinant of the Hessian at the point $\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}(l,k)$ 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|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\beta_1}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\beta_{l-1}}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\beta_{l-1}}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \partial_{\beta_{l+1}}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\beta_{l+1}}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\beta_{p+1}}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\beta_{p+1}}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \partial_{\gamma_1}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\gamma_1}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\gamma_{k-1}}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\gamma_{k-1}}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \partial_{\gamma_{k+1}}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\gamma_{k+1}}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \\ \vdots & \vdots \\ \partial_{\gamma_{p+1}}\partial_{\beta_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} & \partial_{\gamma_{p+1}}\partial_{\gamma_k}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}} \end{pmatrix},$$

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

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

Our goal is to restrict the properties of the Hessian (B.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 $\boldsymbol{\Gamma}_{\mathsf{min}}^{p}$. To this end, we use a powerful theorem from matrix analysis.

Theorem 5 (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*

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

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:

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

Using the fact that $H_p(\boldsymbol{\Gamma}_{\mathsf{min}}^{p})$ has $\lambda_k^p > 0$ for all k, we see that the Hessian of QAOA$_{p+1}$ at point $\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}(l,k)$ 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,

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

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 (B.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

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

This matrix element will play a special role in the calculation of $\det H(\boldsymbol{\Gamma}_{\mathsf{TS}}^{p+1}(l,k))$. It is important to note, that while the specific expression of b differs for all the stationary points in the set given by Eq. (B.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 $\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}$:

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

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

$$\begin{aligned}&\partial_{\gamma_j}\partial_{\beta_{p+1}}E(\boldsymbol{\beta},\boldsymbol{\gamma})\Big|_{\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}} = \\ &\qquad\qquad \langle +|U^\dagger_{<j}[U^\dagger_{\geq j}[H_B,H_C]U_{\geq j},H_C]U_{<j}|+\rangle = a_{p+1+j}.\end{aligned} \tag{B.31}$$

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

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

Note that above we used $U_{>p+1} = \mathbb{I}$. This is because when the derivative is taken with respect to 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.(B.15) is equivalent to $U_{>p+1} = \mathbb{I}$. Doing the same on Eq. (B.31) gives

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

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 $\partial_{\gamma_{p+1}}E(\vec{\beta},\vec{\gamma})$:

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

When evaluated at point $\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}$, we obtain that $[H_C, U^\dagger_{p+1}H_CU_{p+1}] = 0$ 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. (B.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:

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

with

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

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

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

we rewrite the determinant of the full Hessian as follows

$$\begin{aligned} \det\Big[H(\Gamma_{TS}^{p+1})\Big] = & \\ & \det \begin{pmatrix} a_{p+1} & b \\ b & 0 \end{pmatrix} \det\Big[H(\Gamma_{\mathsf{min}}^{p}) - vh^{-1}v^T\Big] \\ & = -b^2 \det\Big[H(\Gamma_{\mathsf{min}}^{p})\Big]. \end{aligned} \tag{B.38}$$

We used that $vh^{-1}v^T = 0$ in the last line. We then see that as long as $b \neq 0$ the determinant of the Hessian at the TS is negative, $\det[H(\Gamma_{TS}^{p+1})] < 0$. The explicit expression (B.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 1 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

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

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

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

where for better readability we introduced the short-hand notation c_j with $j \in [2, p]$. Finally, evaluating the above expressions Eq. (B.39) and Eq. (B.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

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

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

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

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

Using that the point $\mathbf{\Gamma}^{p}_{\mathsf{min}}$ 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. (B.41) is non-vanishing can be inferred from the similar argument to the one used at the end of Appendix B.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 $\mathbf{\Gamma}^{p+1}_{\mathsf{TS}}$ 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 $\mathbf{\Gamma}^{p}_{\mathsf{min}}$. 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 $H(\mathbf{\Gamma}^{p}_{\mathsf{min}})$. 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

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

where

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

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

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

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

$$[A,[B,C]]+[B,[C,A]]+[C,[A,B]]=0, \quad \text{(B.46)}$$

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

$$\begin{aligned} b-\partial_{\beta_{l-1}}\partial_{\gamma_l}E(\boldsymbol{\beta},\boldsymbol{\gamma})\big|_{\boldsymbol{\Gamma}^{p}_{\mathsf{min}}} &= \\ \langle +|U^\dagger_{\leq l-1}[U^\dagger_{>l-1}H_CU_{>l-1},[H_B,H_C]]U_{\leq l-1}|+\rangle &= \bar{b}. \end{aligned} \quad \text{(B.47)}$$

Considering now the matrix elements of v we get

$$\begin{aligned} 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}. \end{aligned} \quad \text{(B.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

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

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

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

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

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

concluding our proof for the general TS.

B.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. 5.1 and 5.2. 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 $N_{\mathsf{min}}(p)$ can be approximated as $N_{\mathsf{min}}(p) \approx 0.19e^{0.98p}$]. However, the limited range of p does not allow us to completely rule out factorial growth, see Fig. B.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. 5.4.

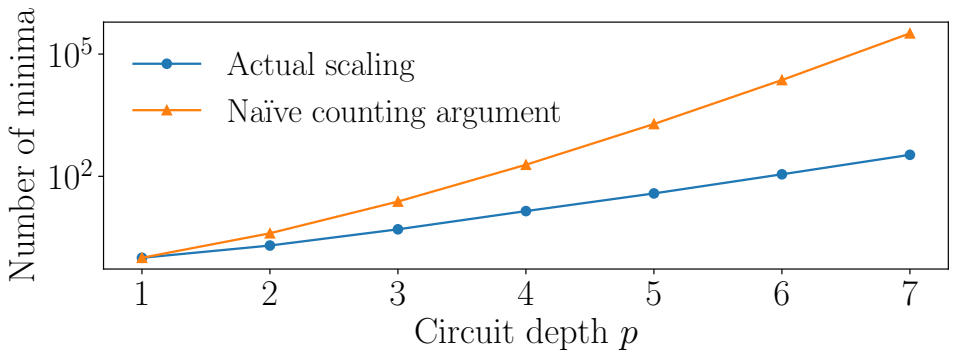


Figure B.1: Number of minima found in the initialization graph in Fig. 5.2 with system size $n = 10$. The orange line describes a naïve counting argument ($2^{p-1}p!$) while the blue line lists the actual number of distinct minima that can be approximated as $0.19\,e^{0.98p}$.

B.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. 5.2(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. 5.2(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$_p$.

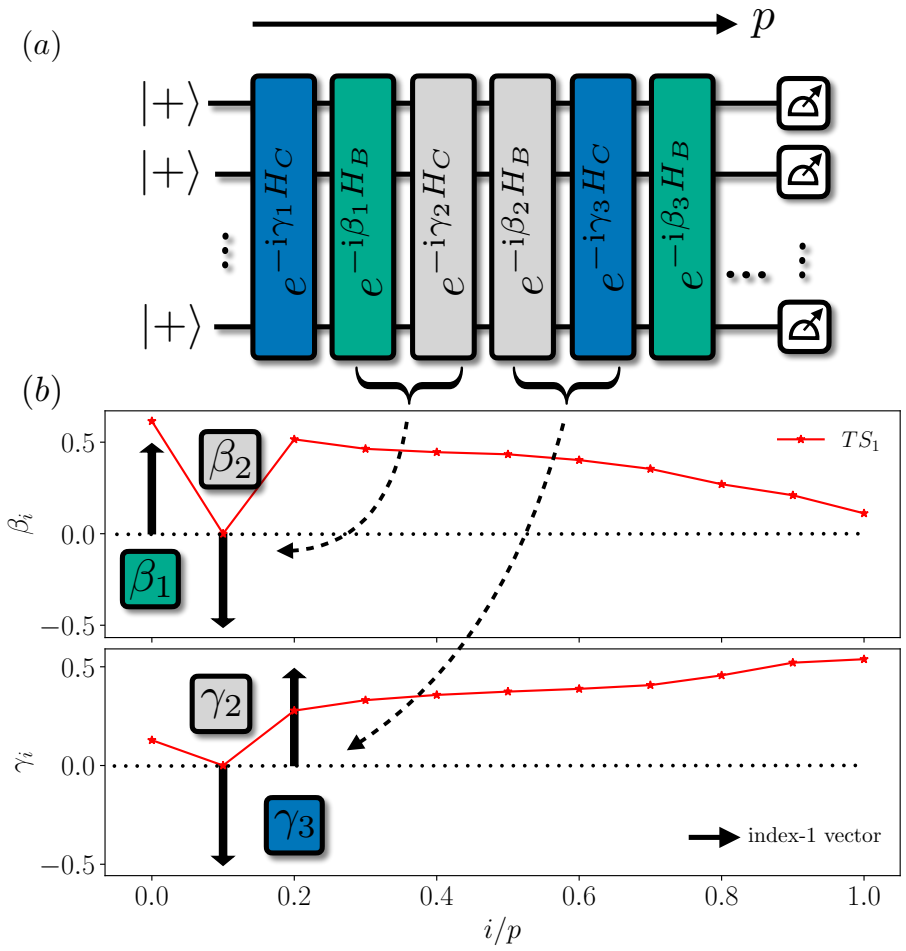


Figure B.2: (a) Illustration of the circuit implementing the QAOA at a TS. Gray gates correspond to the zero insertion. The index-1 direction has mainly weight at the position of the zeros as well as the two adjacent gates. (b) Numerical example of the index-1 vector and the QAOA parameter pattern at 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.

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. B.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.

B.5 Description of the Greedy algorithm

In the following, we provide a detailed description for the GREEDY QAOA initialization, as well as the sub-routines required to implement the algorithm. To this end, we first provide 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 3 QAOA sub-routine

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 $\mathbf{\Gamma}^p_{\min}$

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 4 Grid search sub-routine

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

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

with $i \in [1, p]$ and $j \in [2, p]$
2: QAOA sub-routine initialized from each point in grid
3: Return local minimum with the lowest energy $\mathbf{\Gamma}^p_{\min}$

Using the two sub-routines presented above we can provide a detailed pseudo-code for the GREEDY QAOA algorithm, see Fig. B.3 for a visualization.

Algorithm 5 GREEDY QAOA

Choose maximum circuit depth $p_{\max}$
Choose small offset $\epsilon \ll 1$
Grid search for $p = 1$ to find $\boldsymbol{\Gamma}_{\min}^{p=1}$ ▷ See grid search sub-routine
repeat
    Construct $p+1$ symmetric TS $\boldsymbol{\Gamma}_{TS}^{i,p+1}$ from $\boldsymbol{\Gamma}_{\min}^{p}$
    Compute or approximate the index-1 unit vector $\hat{\boldsymbol{v}}$ for each TS
    Construct points $\boldsymbol{\Gamma}_{\pm}^{i,p+1} = \boldsymbol{\Gamma}_{TS}^{i,p+1} \pm \epsilon \hat{\boldsymbol{v}}_i$ for each TS
    Run QAOA init. from $\boldsymbol{\Gamma}_{\pm}^{i,p+1}$ ▷ See QAOA sub-routine
    Keep local minimum with the lowest energy $\boldsymbol{\Gamma}_{\min}^{p+1}$
    $p \longleftarrow p+1$
until $p = p_{\max}$
Return minimum $\boldsymbol{\Gamma}_{\min}^{p=p_{\max}}$

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, approximation to index-1 direction is expected to give similar performance of QAOA sub-routine at a lower classical computational cost.

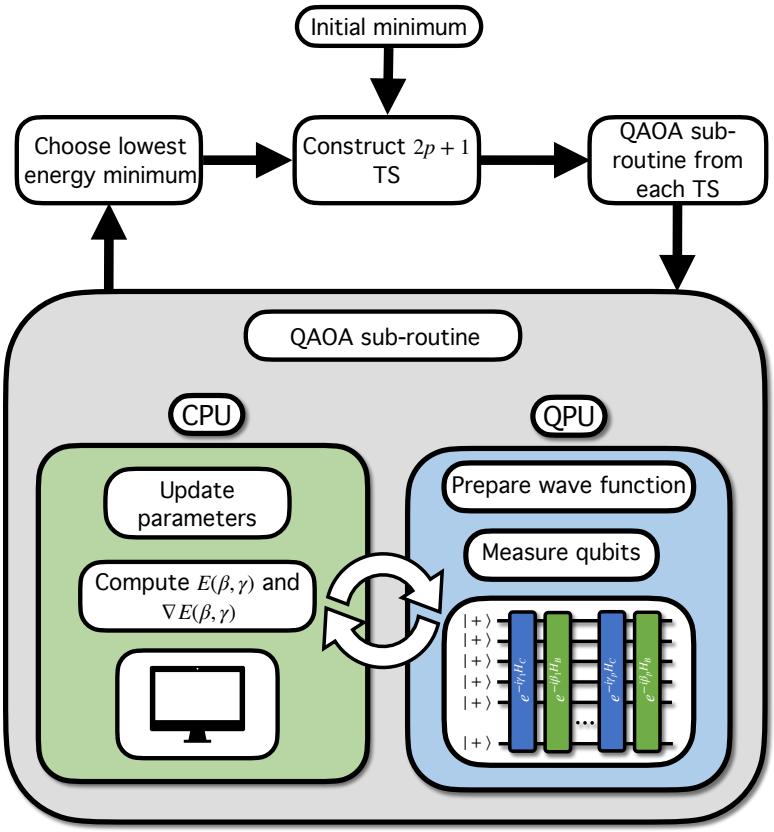


Figure B.3: Flow diagram to visualize the GREEDY QAOA initialization algorithm presented in Algorithm 5.

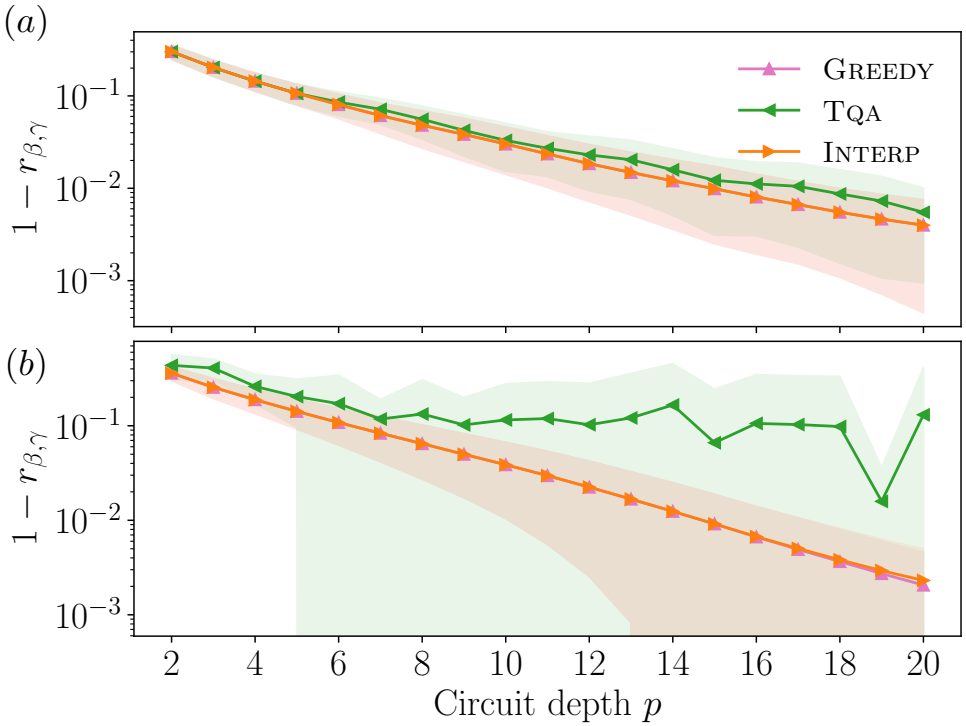


Figure B.4: Performance comparison on (a) RWRG3 and (b) RERG with system size $n = 10$. Data is averaged over 19 non-isomorphic graphs.

B.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 $H_C = \sum_{\langle i,j\rangle \in E} w_{ij}\sigma_i^z\sigma_j^z$ and w_{ij} are random weights $w_{ij} \in [0, 1)$, as well as random Erdős-Rényi graphs (RERG) with edge probability $p_E = 0.5$.

Fig. B.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 that 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. B.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 it yields 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 gain in performance from every additional layer is becoming less for bigger system sizes. This is due to the fact that in order for the QAOA to "see" the whole graph, a circuit depth p scaling as $p \sim \log n$ is required [FGG20].

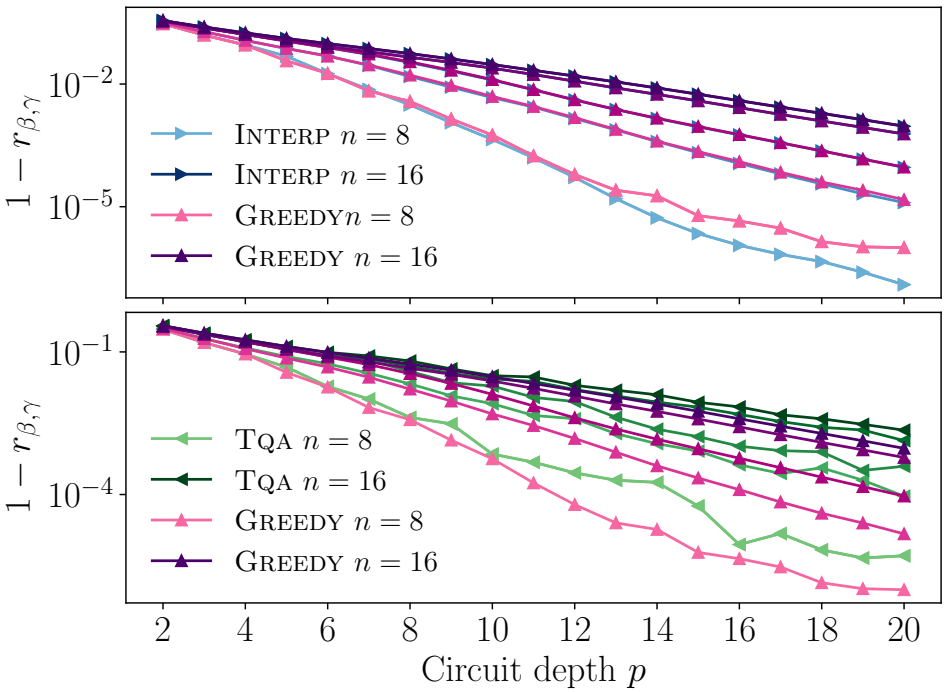


Figure B.5: System size scaling for performance comparison on RRG3. Color shade indicates system size, light color is $n = 8$ and dark color is $n = 16$. System size changes in steps of two between those values. Data is averaged over 19 non-isomorphic RRG3 graphs.

APPENDIX C

Mathematical details for Qudit-QAOA ansatz and fast numerical simulation of qudit noise

C.1 Comparison with previous formulations of the Qudit-QAOA ansatz

The Qudit-QAOA ansatz used in this work is based on the original work in Ref. [BKKT20]. However, in our work we use different formulations of both the cost function C as well as the quantum unitary U_B. Our formulations allow for a direct representation in terms of gates available on ion trap quantum computers. In the following we will discuss the differences and similarities between the two formulations.

Ref. [BKKT20] uses the following classical cost function

$$C = -\sum_{i,j\in E}\sum_{a,b}(1-\delta_{b,0})\,|a\rangle_i \otimes |a\oplus b\rangle_j \langle a|_i \otimes \langle a\oplus b|_j\,, \tag{C.1}$$

which is directly formulated in terms of projectors. We can unfold this formula to obtain the following

$$= -\sum_{i,j\in E}\sum_{a,b}(1-\delta_{0,b})(|a\rangle_i \langle a|_i) \otimes (|a\oplus b\rangle_j \langle a\oplus b|_j) \tag{C.2}$$

$$= -\sum_{i,j\in E}\sum_{a,b}|a\rangle_i \langle a|_i \otimes |a\oplus b\rangle_j \langle a\oplus b|_j - \sum_{i,j\in E}\sum_{a,b}\delta_{0,b}\,|a\rangle_i \langle a|_i \otimes |a\oplus b\rangle_i \langle a\oplus b|_j \tag{C.3}$$

$$= -\sum_{i,j\in E}\Big(1_i\otimes 1_j - \sum_a |a\rangle_i \langle a|_j \otimes \sum_b \delta_{0,b}\,|a\oplus b\rangle_j \langle a\oplus b|_j\Big) \tag{C.4}$$

$$= -\sum_{i,j\in E}\Big(1_i\otimes 1_j - \sum_a |a\rangle_i \langle a|_i \otimes |a\rangle_j \langle a|_j\Big) \tag{C.5}$$

Here we have that $\sum_{i,j\in E} 1_i\otimes 1_j = |E|\times 1^{\otimes n}$ where $|E|$ is the number of edges in the graph.

$$= -|E|\times 1^{\otimes n} + \sum_{i,j\in E}\sum_a |a\rangle_i \langle a|_i \otimes |a\rangle_j \langle a|_j \tag{C.6}$$

The projector in the second term (as seen in Eq. (4.6))is diagonal and only returns 1 if qudit i is in the the same state as qudit j, i.e. it has the same color, which is precisely the desired δ_{c_i,c_j}. We thus find

$$= -|E| \times 1^{\otimes n} + \sum_{i,j\in E} \delta_{c_i,c_j}. \tag{C.7}$$

This shows that our formulation of the cost function in terms of the antiferromagnetic Potts model, Eq. (4.4), is equivalent to the formulation in Ref. [BKKT20] up to a constant shift in energy.

Next, we have the classical unitary $U_B(\boldsymbol{\beta}_t)$ for which [BKKT20] defines the following unitary that generates mixing on a single site

$$U_B(\boldsymbol{\beta}) = \sum_a e^{i\beta_a} |\phi_a\rangle\langle\phi_a|, \tag{C.8}$$

where $|\phi_a\rangle$ are defined as $|\phi_a\rangle = Z^a |+\rangle$. The generalized plus state is given by $|+\rangle = d^{-1/2} \sum_a |a\rangle$ and the $+1$ eigenstate of the generalized X operator. In this definition we can in fact pull out a global phase

$$U_B(\boldsymbol{\beta}) = e^{i\beta_0} \sum_a e^{i(\beta_a - \beta_0)} |\phi_a\rangle \langle\phi_a| , \tag{C.9}$$

which is irrelevant (since probabilities are given by absolute squares and global phase thus does not change any physically observable quantities). The parameter space of $\boldsymbol{\beta}$ can thus be reduced to $\boldsymbol{\beta} \in (\mathbb{R}^{d-1})^p$. Let us now transform into the X-basis using a Hadamard transformation. We can thus write

$$U_B = \sum_{a\neq 0} e^{i\beta_a} Z^a |+\rangle \langle+| (Z^a)^\dagger = \sum_{a\neq 0} e^{i\beta_a} H^\dagger X^a H |+\rangle \langle+| H^\dagger (X^a)^\dagger H, \tag{C.10}$$

note that $H |+\rangle \langle+| H^\dagger = |0\rangle \langle 0|$. This allows us to write

$$U_B = \sum_{a\neq 0} e^{i\beta_a} H^\dagger X^a |0\rangle \langle 0| (X^a)^\dagger H, \tag{C.11}$$

since $X^a |0\rangle = |a\rangle$ we can finally write

$$U_B = H^\dagger \left(\sum_{a\neq 0} e^{i\beta_a} |a\rangle \langle a| \right) H, \tag{C.12}$$

which is the definition used in this work, see Eq. (4.9).

C.2 Details on representation of Qudit-QAOA in terms of ion trap native gates

In this section we discuss the details of how the Qudit-QAOA ansatz can be expressed in terms of gates that are native to ion trap quantum computers.

First, we consider the unitary $e^{-i\gamma_t C}$. In particular, we will show that, up to a global phase, it is equivalent to the all-level entangling gate given in Ref. [HWG+22]

$$G(\gamma) : \begin{cases} |jj\rangle \to |jj\rangle \\ |jk\rangle \to e^{i\gamma} |jk\rangle \quad j \neq k \end{cases} \tag{C.13}$$

that has to be applied to the pairs of qudits associated with the edges in the graph. To see this we first write $G(\gamma)$ out in its matrix form for qutrits (without loss of generality)

$$G(\gamma) = \begin{bmatrix} 1 & & & & & & & & \\ & e^{i\gamma} & & & & & & & \\ & & e^{i\gamma} & & & & & & \\ & & & 1 & & & & & \\ & & & & e^{i\gamma} & & & & \\ & & & & & e^{i\gamma} & & & \\ & & & & & & 1 & & \\ & & & & & & & e^{i\gamma} & \\ & & & & & & & & e^{i\gamma} \end{bmatrix} = e^{i\gamma} \begin{bmatrix} e^{-i\gamma} & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & & & \\ & & & e^{-i\gamma} & & & & & \\ & & & & 1 & & & & \\ & & & & & 1 & & & \\ & & & & & & e^{-i\gamma} & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{bmatrix}$$
$$= e^{i\gamma} e^{-i\gamma\delta_{c_i,c_j}} \tag{C.14}$$

Recalling the target unitary $e^{-i\gamma C} = \prod_{i,j\in E} e^{-i\gamma\delta_{c_i,c_j}}$, we see that this is exactly a product of $G(\gamma)$ acting on pairs of qutrits that are edges of the graph, up to a global phase $e^{i\gamma}$ which we can ignore (probabilities are given by absolute squares and a global phase does therefore not change any physically observable quantities). Explicitly we thus have

$$e^{-i\gamma_t C} \sim \prod_{i,j\in E} G_{i,j}(\gamma_t), \tag{C.15}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^p$.

Next, we consider the quantum gate $U_B = H^\dagger\Big(\sum_{a\neq 0} e^{i\beta_a} |a\rangle \langle a|\Big)H$. The Hadamard gate H is a standard gate that is available on all quantum computing platforms. To implement the sum $\sum_{a\neq 0} e^{i\beta_a} |a\rangle \langle a|$ we can use a phase shift gate, given in [RMP+22] as

$$Z_i(\beta)|j\rangle = \begin{cases} e^{-i\beta}|j\rangle & \text{if } i = j \\ |j\rangle & \text{else.} \end{cases} \tag{C.16}$$

To see this, let us apply the phase shift gate on two qutrits levels (without loss of generality). We have that

$$Z_1(\beta_1)Z_2(\beta_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-i\beta_1} & 0 \\ 0 & 0 & e^{-i\beta_2} \end{bmatrix}, \tag{C.17}$$

which we can quickly convince ourselves that this realizes the desired sum term. So we can finally write

$$U_B = H^\dagger\left(\prod_{a\neq 0} Z_a(-\beta_a)\right)H, \tag{C.18}$$

which is a single qudit unitary that introduces relative phases between the basis states. Note that the phase shift gate is applied "in software" and thus does not have an error.

C.3 Generalized Pauli representation of depolarizing channel

In this section, we prove the equivalence in Eq. (4.17) used for the Monte Carlo strategy for simulating noise with random sampling. To this end, consider the Pauli group in d dimensions

and let $|0\rangle, \ldots, |d-1\rangle$ be the computational basis in $\mathbb{C}^d$. Recall, that the generalized X and Z operators act on these basis states like

$$X|k\rangle = |k \oplus 1\rangle \quad \text{and} \quad Z|k\rangle = \omega_d^p |z\rangle \quad \text{for all } k = 0, \ldots, d-1.$$

Here, $\oplus$ denotes addition modulo d (e.g. $|(d-1) \oplus 1\rangle = |0\rangle$) and $\omega_d = \exp(2\pi \mathrm{i}/d)$ is a d-th root of unity. These transformation rules readily extend to powers of X and Z. For $p, q \in \{0, \ldots, d-1\}$, we obtain

$$X^q|k\rangle = |k \oplus q\rangle \quad \text{and} \quad Z^p|k\rangle = \omega_d^{pk}|k\rangle.$$

Up to global phases (which do not matter here), the generalized Pauli matrices encompass the following d^2 unitary operators: $\mathcal{P}_d = \{Z^p X^q : \ p, q \in \{0, \ldots, d-1\}\}$.

Lemma 6. *Averaging over all generalized Pauli unitaries produces the completely depolarizing channel*

$$\mathcal{T}(A) = \frac{1}{d^2} \sum_{p,q=0}^{d-1} Z^p X^q A (X^q)^\dagger (Z^p)^\dagger = \frac{\mathrm{tr}(A)}{d} \mathbb{I} = \mathcal{D}(A) \quad \textit{for all } d \times d \textit{ matrices } A \tag{C.19}$$

Proof. Let us start with computing the desired expression for $A = |k\rangle\langle l|$ with $k, l = 0, \ldots, d-1$:

$$\begin{aligned}
\mathcal{T}(|k\rangle\langle l|) =& \frac{1}{d^2} \sum_{p,q=0}^{d-1} Z^p X^q |k\rangle\langle l| (X^q)^\dagger (Z^p)^\dagger \\
=& \frac{1}{d^2} \sum_{p,q=0}^{d-1} Z^p |k \oplus q\rangle\langle k \oplus q| (Z^p)^\dagger \\
=& \frac{1}{d^2} \sum_{p,q=0}^{d-1} \omega^{p(k\oplus q)} |k \oplus q\rangle\langle k \oplus q| \omega^{-p(k\oplus q)} \\
=& \frac{1}{d} \left(\frac{1}{d} \sum_{p=0}^{d-1} \omega^{p(k-l)} \right) \sum_{q=0}^{d-1} |k \oplus q\rangle\langle l \oplus q| \\
=& \frac{\delta_{k,l}}{d} \sum_{q=0}^{d-1} |k \oplus q\rangle\langle k \oplus q| \\
=& \frac{\delta_{k,l}}{d} \sum_{\tilde{q}=0}^{d-1} |\tilde{q}\rangle\langle \tilde{q}| = \frac{\delta_{k,l}}{d} \mathbb{I}_d.
\end{aligned}$$

The general claim now follows from decomposing $A = \sum_{k,l=0}^{d-1} A_{k,l} |k\rangle\langle l|$ and applying this relation:

$$\mathcal{T}(A) = \sum_{i,j=0}^{d-1} A_{i,j} \mathcal{T}(|k\rangle\langle l|) = \sum_{i,j=0}^{d-1} A_{i,j} \frac{\delta_{i,j}}{d} \mathbb{I}_d = \left(\sum_{i=0}^{d-1} A_{k,k} \right) \frac{1}{d} \mathbb{I}_d = \frac{\mathrm{tr}(A)}{d} \mathbb{I}.$$

□

C.4 Fast noisy simulation of Qudits-QAOA using the Fast Walsh Hadamard Transform

In this section, we discuss the details of our highly efficient simulation of the noisy Qudit-QAOA. First, let us recall that for the Qudit-QAOA we require two unitary operators. The classical

unitary, $e^{-i\gamma_t C}$, which is diagonal in the computational basis and $U_B(\boldsymbol{\beta}_t)$ which is only diagonal in the X-basis. These two unitary operators have to be applied on the initial state $|+\rangle^{\otimes n}$. Since the classical unitary is diagonal, we can in fact directly apply it to the initial state of length d^n as a vector-vector multiplication which has time complexity $\mathcal{O}(N)$ where N is the length of the vector (d^n in this case). To also implement the quantum unitary as a vector-vector multiplication, we use the Fast Walsh Hadamard Transform (FWHT) in order to transform into the X-basis where the operator becomes diagonal. The standard FWHT is defined for binary systems, below we present its generalization to arbitrary dimensionality. This is effectively achieved by recursively applying the generalized Hadamard transform on appropriate partitions of the state vector.

Algorithm 6 Generalized Fast Walsh-Hadamard Transform (gFWHT)

procedure GFWHT(ψ, d)
    Input: A state vector ψ, local Hilbert space dimension d
    Output: The transformed state vector

    $n \leftarrow \log_d \text{len}(\psi)$ ▷ Calculate the number of qudits

    $\tilde{\psi} \leftarrow \text{reshape}(\psi, (d, d^{n-1}))$ ▷ Reshape the state vector
    $w \leftarrow \text{zeros}((d, d^{n-1}))$ ▷ Initialize an empty transformed state vector
    for $i \in \{0, \ldots, d-1\}$ **do** ▷ Iterate over the range of local Hilbert space dimensions
        for $j \in \{0, \ldots, d-1\}$ **do** ▷ Iterate over the range of local Hilbert space dimensions
            $w_{ij} \leftarrow \sum_{k=0}^{d-1} \omega^{ijk} \tilde{\psi}_{ik}$ ▷ Compute the w_{ij} element
        end for
    end for

    $w \leftarrow \text{reshape}(w, \text{len}(\psi))$ ▷ Reshape the transformed state vector
    $\psi \leftarrow \text{gfwht}(w, d)$ ▷ Recursively apply the gFWHT

    return ψ ▷ Return the transformed state vector
end procedure

The generalized FWHT (gFWHT) has time complexity $\mathcal{O}(N \log N)$ and thus allows to efficiently transform into the X-basis and apply the quantum unitary using vector-vector multiplication. Once we have applied the quantum unitary we can transform back into the computational basis.

We can also use this trick to efficiently simulate noise, in particular, we can note that the noise term can be expressed as $X^q Z^p = H^\dagger Z^q H Z^p$ we see that can also implement the noise as a vector-vector multiplication by using the gFWHT (the gFWHT implements a Hadamard transform). In particular we can introduce a helper variable $\alpha_i \in \{0, 1\}$, which we use to express the noisy state $|\psi\rangle_{\text{noisy}}$ as

$$\begin{aligned} |\psi\rangle_{\text{noisy}} = \bigotimes_i (X^{q_i} Z^{p_i})^{\alpha_i} |\psi\rangle &= \bigotimes_i (H^\dagger Z^{q_i \alpha_i} H Z^{p_i \alpha_i}) |\psi\rangle \\ &= \text{gFWHT}^{-1}\Big[\bigotimes_i Z^{q_i \alpha_i} \text{gFWHT}(\bigotimes_i Z^{p_i \alpha_i} |\psi\rangle) \Big], \end{aligned} \tag{C.20}$$

if $\alpha_i = 0$ no error is applied to qudit i and for $\alpha = 1$ a random element of the Pauli group $X^{q_i} Z^{p_i}$ is applied, if we select the two values such that

$$P(\alpha_i = k) = p^k (1-p)^{1-k}, \quad k \in \{0, 1\}, \tag{C.21}$$

which will produce the depolarizing channel on average.

Note that in fact we only have to transform once into the X-basis for implementing X^q and only transform back once we have also implemented the U_B term. This procedure allows us to simulate system sizes that would otherwise be well beyond classical computational resources. A noisy observable is then obtained by sampling a set of $\{|\psi\rangle^i_{\text{noisy}}\}$ and computing the average of the expectation values

$$\langle O \rangle \approx \frac{1}{M} \sum_i \langle \psi |_{\text{noisy, i}} \, O \, |\psi\rangle_{\text{noisy, i}} \tag{C.22}$$

Fig. C.1 illustrates a $p = 1$ noisy qudit-QAOA circuit using the gFWHT that allows applying both noise and unitary evolution gates as diagonal matrices, i.e. as a fast vector-vector multiplication.

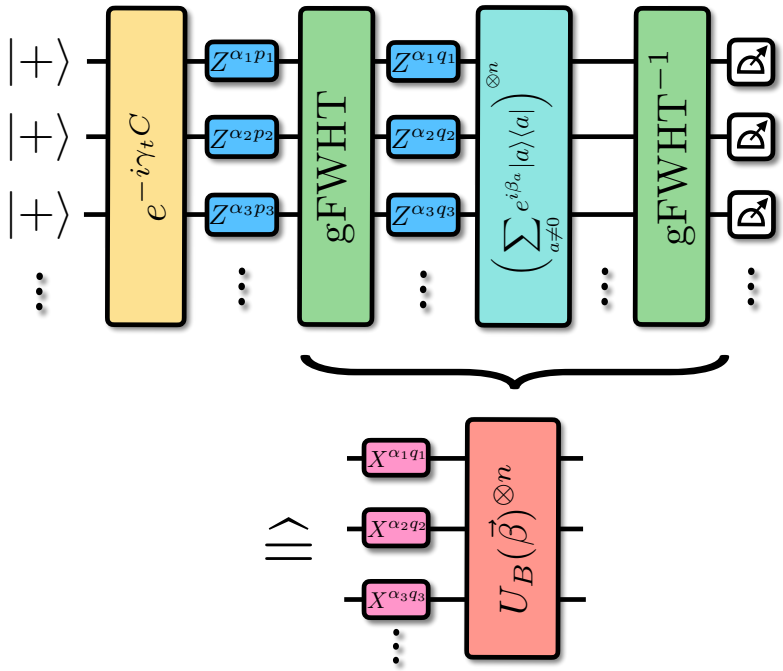


Figure C.1: Circuit for $p = 1$ that is used to approximate the single qudit depolarizing channel in our simulations. We use a gFWHT to implement both the noise and unitary gates as vector-vector multiplication. This has both lower time and memory complexity than naive matrix-vector multiplication.

C.5 Cost function expectation value for random sampling

Recall the antiferromagnetic Potts model, $C = \sum_{i,j \in E} \delta_{c_i, c_j}$, for a colorable graph we thus have $C \in [0, |E|]$. The upper limit is however more of a theoretical one since it is highly

unlikely that not a single edge does not have the same color. Let us now consider what the expectation value is for an equal superposition of all colorings, i.e. the average cost function under random coloring with uniform probability. Each vertex is a superposition of d colored states $|v_i\rangle = \frac{1}{\sqrt{d}}(|c_1\rangle + |c_2\rangle + ... + |c_{d-1}\rangle)$, thus have $|s\rangle = |v_1\rangle \otimes |v_2\rangle \otimes ... \otimes |v_n\rangle$. The expectation value is thus

$$\langle s| C |s\rangle = \sum_{i,j \in E} \sum_{c} \langle s| \left(|c_i\rangle \langle c_i|\right) \otimes \left(|c_j\rangle \langle c_j|\right) |s\rangle , \tag{C.23}$$

we thus have

$$= \sum_{i,j \in E} \sum_{c} \langle v_i|c_i\rangle \langle c_i|v_i\rangle \langle v_j|c_j\rangle \langle c_j|v_j\rangle = \sum_{i,j \in E} \sum_{c} \frac{1}{d}\frac{1}{d} = \frac{1}{d}|E|. \tag{C.24}$$

This implies that any value below $|E|/d$ is better than random guessing.

APPENDIX D

Mathematical details for classical shadows and further numerical results

D.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 use L to denote this cardinality). These techniques do, however, require copies of the underlying quantum state to be stored in parallel within a quantum memory and highly entangled gates to be performed 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 7. 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. (D.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. (D.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.

D.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 carry out the following.

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 "postmeasurement" 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 is measured 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. 5.1 (a) for a visualization.
5. (Implicitly) Construct the N-qubit *classical shadow*

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

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):

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

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, Eq. (D.1), plays nicely with taking partial traces. Let $A \subseteq \{1, \ldots, N\}$ be a collection of $|A| = k$ qubits. Then,

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

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

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

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. For example coupling terms in the Heisenberg Hamiltonian (5.1) have range $k = 2$, while the external field terms have range $k = 1$.

Theorem 7. *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. In other words,*

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

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

We emphasize that it is not necessary to form global shadow approximations. If O_l only acts nontrivially 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 7 is slightly stronger than a related result in Ref. [HKP20] (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 7 in their work. We present a formal proof in Appendix D.1.5 below.

D.1.2 Estimating subsystem purities

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

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

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

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

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. (D.3). Formula (D.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

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

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. (D.7) can be further improved to $\ln(1/\delta)$ by replacing simple empirical averaging in Eq. (D.6) by median-of-means estimation [HKP20]. 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 trade-off only becomes viable if one wishes to approximate polynomially many subsystem purities.

D.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. (1.15)], 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 expression of the gradient as a linear combination of $2L$ expectation values,

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

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 7 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. (D.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 produces only

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

This upper bound equals only ϵ if we rescale the accuracy of original approximation to ϵ/L. Inserting this rescaled accuracy into Theorem 7 produces an overall measurement cost of

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

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, Eq. (D.9), required 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, Eq. (D.7). To obtain the full gradient $\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$ the procedure has to be repeated pN times since the parameter-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.

D.1.4 Example of error accumulation in an Ising model

The extra scaling with L^2 in Eq. (D.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 1D 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

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

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 $\mathbb{E}\left[\hat{E}\right] = \mathrm{tr}\left(H\rho\right)$, 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 $|\psi\rangle = |00\cdots00\rangle$ is the Ising ground state and $|\phi\rangle = |01\cdots01\rangle$ 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

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

The task is to approximate this expectation value based on computational basis measurements. For each measurement, we either obtain outcome $0\cdots0$ (with probability $1-p$) or outcome $01\cdots01$ (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 (rescaled) 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 $\mathbb{E}\left[\hat{E}\right] = \mathrm{tr}\left(\rho(\lambda)H\right)$. 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 considered in the main text, see Eq. (5.1), is straightforward.

D.1.5 Proof of Theorem 7

Theorem 7 is a consequence of the following concentration inequality. Let $\|O\|_\infty$ denote the operator and spectral norm of an observable. We also use $\|\cdot\|_1$ to denote the trace norm.

Theorem 8. *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 $\epsilon \in (0,1)$,*

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

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

Theorem 9 (Bernstein inequality)**.** *Let $X^{(1)},\ldots,X^{(T)}$ be independent, centered (i.e., $\mathbb{E}\left[X_t\right] = 0$) random variables that obey $|X^{(t)}| \leq R$ almost surely. Then, for $\epsilon > 0$*

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

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

Proof of Theorem 8. 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. (D.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 also use $\|\rho_A\|_1 = \mathrm{tr}(\rho_A) = 1$ and the specific form of subsystem classical shadows Eq. (D.3), that factorizes nicely into tensor products. Estimating the variance is more difficult by comparison. However, Ref. [HKP20, 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

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

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:

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

□

D.2 Unitary t-designs

Here, we briefly review the notion of unitary t-designs. The Haar measure is the unique left and 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

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

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,

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

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, Zhu17a]. 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

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

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.

D.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 modeling quantum dynamics by random unitaries is an approximation, it 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

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

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

Given a bipartition $(A, \neg A)$ of the system, we begin by studying the distance of the reduced 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 $d = 2^N$.

D.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) $\|M\|_2 = \sqrt{\operatorname{tr}(M^\dagger M)}$

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

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 (D.12),

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

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

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

As we see, the only nontrivial dependence on U comes from the purity of the reduced 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 reduced 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. (D.11).

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

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

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

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

we get that the following simple expression for the expected purity

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

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

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

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.

D.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 [HKP20].

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

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

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

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

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

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

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 does not 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*:

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

[Pag93] conjectured 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

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

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 2, that is

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

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.

D.4 Entanglement growth and learning rate

Here we detail the derivation of Eq. (5.7). We first upper bound the trace distance via

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

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

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

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

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

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

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

where in the last equality we plug in the parameter change under GD (Eq. (5.2)), $\boldsymbol{\delta} = -\eta \nabla_{\boldsymbol{\theta}} E$.

D.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 SYK model [Kit15]. The SYK model provides a canonical example for a volume-law model where the ground state is nearly maximally entangled [HG19]. The nonlocal nature of the

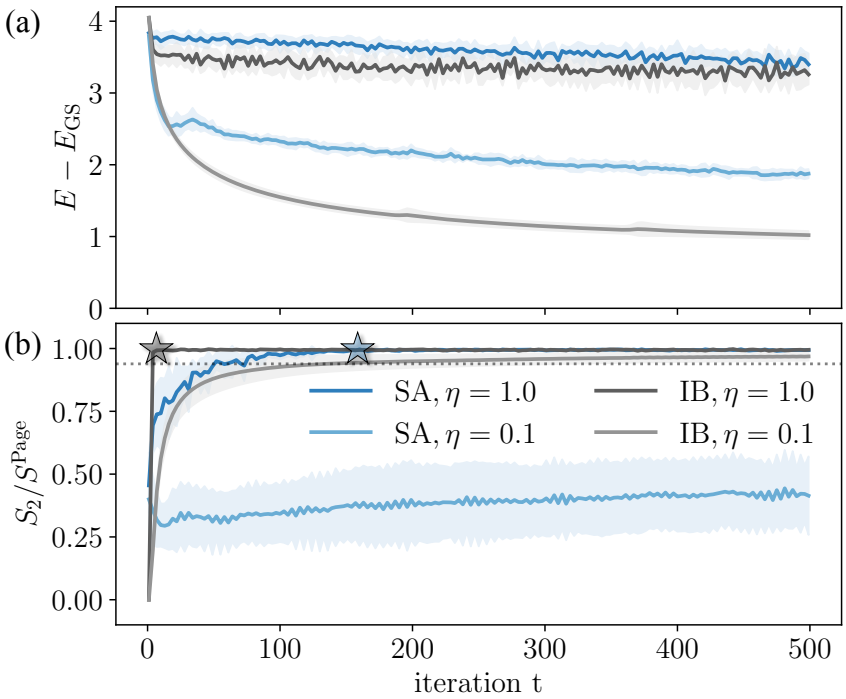


Figure D.1: (a-b) The application of our algorithm to the problem of finding the ground state of the SYK model. For the initialization we consider the small-angle (SA) ($\epsilon_\theta = 0.1$) and identity block (IB) initialization [GWOB19] (using one block). We can see that only through the reset of 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. (1.15) with circuit depth $p = 100$. Data is averaged over 100 random instances.

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 anticommutation 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

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

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:

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

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. D.1 (a)-(b) show that the 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 chosen 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.

Bibliography

[A+20a] Frank Arute et al. Observation of separated dynamics of charge and spin in the Fermi-Hubbard model. *arXiv e-prints*, page arXiv:2010.07965, October 2020.

[A+20b] Frank Arute et al. Quantum Approximate Optimization of Non-Planar Graph Problems on a Planar Superconducting Processor. *arXiv e-prints*, page arXiv:2004.04197, April 2020.

[AAA+20] Iskandar Akhalwaya, Cody Alexander, Nandini Anand, Alessandro Benfenati, Dominik Bucher, Giacomo Ciani, Wibe A de Jong, Marc Ganzhorn, Daniel Gottwald, Samuel Hawes, et al. Extending the applicability of the variational quantum eigensolver. *Quantum Science and Technology*, 5(2):025024, 2020.

[AAA+23] Igor Aleiner, Richard Allen, Trond I Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Juan Atalaya, Ryan Babbush, Dave Bacon, et al. Suppressing quantum errors by scaling a surface code logical qubit. *Nature*, 614(7949):676–681, 2023.

[AAB+19] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[AAB+20] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Sergio Boixo, Michael Broughton, Bob B. Buckley, David A. Buell, Brian Burkett, Nicholas Bushnell, Yu Chen, Zijun Chen, Benjamin Chiaro, Roberto Collins, William Courtney, Sean Demura, Andrew Dunsworth, Daniel Eppens, Edward Farhi, Austin Fowler, Brooks Foxen, Craig Gidney, Marissa Giustina, Rob Graff, Steve Habegger, Matthew P. Harrigan, Alan Ho, Sabrina Hong, Trent Huang, William J. Huggins, Lev Ioffe, Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Cody Jones, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Seon Kim, Paul V. Klimov, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Pavel Laptev, Mike Lindmark, Erik Lucero, Orion Martin, John M. Martinis, Jarrod R. McClean, Matt McEwen, Anthony Megrant, Xiao Mi, Masoud Mohseni, Wojciech Mruczkiewicz, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Hartmut Neven, Murphy Yuezhen Niu, Thomas E. O'Brien, Eric Ostby, Andre Petukhov, Harald Putterman, Chris Quintana, Pedram Roushan, Nicholas C. Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Doug Strain, Kevin J. Sung, Marco Szalay, Tyler Y. Takeshita, Amit Vainsencher, Theodore White, Nathan Wiebe, Z. Jamie Yao, Ping Yeh, and Adam Zalcman. Hartree-Fock on a superconducting qubit quantum computer. *Science*, 369(6507):1084–1089, August 2020.

[AAG20] Mahabubul Alam, Abdullah Ash-Saki, and Swaroop Ghosh. Accelerating Quantum Approximate Optimization Algorithm using Machine Learning. *arXiv e-prints*, page arXiv:2002.01089, February 2020.

[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.

[AGDLHG05] Alán Aspuru-Guzik, Anthony D Dutoi, Peter J Love, and Martin Head-Gordon. Simulated quantum computation of molecular energies. *Science*, 309(5741):1704–1707, 2005.

[AL18] Tameem Albash and Daniel A. Lidar. Adiabatic quantum computation. *Rev. Mod. Phys.*, 90:015002, Jan 2018.

[AM76] Neil W. Ashcroft and N. David Mermin. *Solid State Physics*. Brooks Cole, 1976.

[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. *Physical Review A*, 104(1):L010401, July 2021.

[ASS21] Atithi Acharya, Siddhartha Saha, and Anirvan M. Sengupta. Informationally complete POVM-based shadow tomography. *arXiv e-prints*, page arXiv:2105.05992, May 2021.

[ASZ+20] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *arXiv e-prints*, page arXiv:2011.00027, October 2020.

[AvK+08] Dorit Aharonov, Wim van Dam, Julia Kempe, Zeph Landau, Seth Lloyd, and Oded Regev. Adiabatic Quantum Computation Is Equivalent to Standard Quantum Computation. *SIAM Review*, 50(4):755–787, January 2008.

[BACS07] Dominic W. Berry, Graeme Ahokas, Richard Cleve, and Barry C. Sanders. Efficient Quantum Algorithms for Simulating Sparse Hamiltonians. *Communications in Mathematical Physics*, 270(2):359–371, March 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. *Physical Review Letters*, 126(7):070505, February 2021.

[BBF+18a] 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.

[BBF+18b] Fernando G. S. L. Brandao, 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, T. F. Rosenbaum, and G. Aeppli. Quantum Annealing of a Disordered Magnet. *Science*, 284:779, April 1999.

[Bel97] Richard Bellman. *Introduction to Matrix Analysis (2nd Ed.)*. Society for Industrial and Applied Mathematics, USA, 1997.

[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.

[BIS+18] Sergio Boixo, Sergei V Isakov, Vadim N Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J Bremner, John M Martinis, and Hartmut Neven. Characterizing quantum supremacy in near-term devices. *Nature Physics*, 14(6):595–600, 2018.

[BK98] Sergey Bravyi and Alexei Kitaev. Quantum error correction with imperfect gates. *arXiv preprint quant-ph/9811052*, 1998.

[BK21] Lennart Bittel and Martin Kliesch. Training Variational Quantum Algorithms Is NP-Hard. *Physical Review Letters*, 127(12):120502, September 2021.

[BKKT19] Sergey Bravyi, Alexander Kliesch, Robert Koenig, and Eugene Tang. Obstacles to State Preparation and Variational Optimization from Symmetry Protection. *arXiv e-prints*, page arXiv:1910.08980, October 2019.

[BKKT20] Sergey Bravyi, Alexander Kliesch, Robert Koenig, and Eugene Tang. Hybrid quantum-classical algorithms for approximate graph coloring. *arXiv e-prints*, page arXiv:2011.13420, November 2020.

[BKWS20] Bela Bauer, Valentin Kasper, David Wecker, and Gerd Schön. Quantum simulation of fermion-phonon models. *Nature Communications*, 11(1):1–9, 2020.

[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.

[BNR+18] Panagiotis Kl Barkoutsos, Giacomo Nannicini, Stefan Robert, Ivano Tavernelli, Keisuke Fujii, and Nikitas Gidopoulos. Quantum algorithms for electronic structure calculations: Particle/hole hamiltonian and optimized wave-function expansions. *Physical Review A*, 98(2):022322, 2018.

[BR12] R Blatt and C F Roos. Quantum computing: trapped-ion blues. *Nature*, 483(7389):304–305, 2012.

[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.

[BSK+17] Hannes Bernien, Sylvain Schwartz, Alexander Keesling, Harry Levine, Ahmed Omran, Hannes Pichler, Soonwon Choi, Alexander S Zibrov, Manuel Endres, Markus Greiner, et al. Probing many-body dynamics on a 51-atom quantum simulator. *Nature*, 551(7682):579–584, 2017.

[BVC21] Stefano Barison, Filippo Vicentini, and Giuseppe Carleo. An efficient quantum algorithm for the time evolution of parameterized circuits. *Quantum*, 5:512, July 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.

[CLB21] Alexandru Cîrstocea, Yuan Liao, and Bela Bauer. Variational quantum eigensolvers for time-dependent hamiltonians: Application to dissipative quantum many-body systems. *Physical Review Research*, 3(3):033117, 2021.

[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.

[CPSP19] Pieter W. Claeys, Mohit Pandey, Dries Sels, and Anatoli Polkovnikov. Floquet-engineering counterdiabatic protocols in quantum many-body systems. *Phys. Rev. Lett.*, 123:090602, Aug 2019.

[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+21] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. *Chemical reviews*, 121(2):879–948, 2021.

[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.

[CT17] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, February 2017.

[CTDL97] Claude Cohen-Tannoudji, Bernard Diu, and Franck Laloë. *Quantum Mechanics: Concepts and Applications*. John Wiley & Sons, 1997.

[CYZF21] Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia. Robust shadow estimation. *PRX Quantum*, 2:030348, Sep 2021.

[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.

[Dir58] P. A. M. Dirac. *The Principles of Quantum Mechanics*. Oxford University Press, 4 edition, 1958.

[DJ92] David Deutsch and Richard Jozsa. Rapid solutions of problems by quantum computation. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 439(1907):553–558, 1992.

[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.

[DS13] M H Devoret and R J Schoelkopf. Superconducting circuits for quantum information: an outlook. *Science*, 339(6124):1169–1174, 2013.

[EBK+23] Simon J. Evered, Dolev Bluvstein, Marcin Kalinowski, Sepehr Ebadi, Tom Manovitz, Hengyun Zhou, Sophie H. Li, Alexandra A. Geim, Tout T. Wang, Nishad Maskara, Harry Levine, Giulia Semeghini, Markus Greiner, Vladan Vuletic, and Mikhail D. Lukin. High-fidelity parallel entangling gates on a neutral atom quantum computer. *arXiv e-prints*, page arXiv:2304.05420, April 2023.

[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.

[EHF19] Tim J. Evans, Robin Harper, and Steven T. Flammia. Scalable Bayesian Hamiltonian learning. *arXiv e-prints*, page arXiv:1912.07636, December 2019.

[EKC+22] S. Ebadi, A. Keesling, M. Cain, T. T. Wang, H. Levine, D. Bluvstein, G. Semeghini, A. Omran, J. G. Liu, R. Samajdar, X. Z. Luo, B. Nash, X. Gao, B. Barak, E. Farhi, S. Sachdev, N. Gemelke, L. Zhou, S. Choi, H. Pichler, S. T. Wang, M. Greiner, V. Vuletić, and M. D. Lukin. Quantum optimization of maximum independent set using Rydberg atom arrays. *Science*, 376(6598):1209–1215, June 2022.

[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. *Physical Review Letters*, 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.

[EMW20] Daniel J. Egger, Jakub Marecek, and Stefan Woerner. Warm-starting quantum optimization. *arXiv e-prints*, page arXiv:2009.10095, September 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.

[FGG+01] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, Joshua Lapan, Andrew Lundgren, and Daniel Preda. A Quantum Adiabatic Evolution Algorithm Applied to Random Instances of an NP-Complete Problem. *Science*, 292(5516):472–476, April 2001.

[FGG14a] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A Quantum Approximate Optimization Algorithm. *arXiv e-prints*, page arXiv:1411.4028, November 2014.

[FGG14b] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A Quantum Approximate Optimization Algorithm. *arXiv e-prints*, page arXiv:1411.4028, November 2014.

[FGG20] 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.

[FGGS00] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, and Michael Sipser. Quantum Computation by Adiabatic Evolution. *arXiv e-prints*, pages quant–ph/0001106, January 2000.

[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.

[FMMC12] Austin G Fowler, Matteo Mariantoni, John M Martinis, and Andrew N Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86(3):032324, 2012.

[FMT+22] Laurin E. Fischer, Daniel Miller, Francesco Tacchino, Panagiotis Kl. Barkoutsos, Daniel J. Egger, and Ivano Tavernelli. Ancilla-free implementation of generalized measurements for qubits embedded in a qudit space. *Physical Review Research*, 4(3):033027, July 2022.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[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.

[GLF19] Vlad Gheorghiu, Daniel Litinski, and Austin G Fowler. Optimization of the surface code with a reduced number of control qubits. *npj Quantum Information*, 5(1):1–11, 2019.

[GM62] Murray Gell-Mann. Symmetries of baryons and mesons. *Physical Review*, 125(3):1067, 1962.

[GM19] G. G. Guerreschi and A. Y. Matsuura. QAOA for Max-Cut requires hundreds of qubits for quantum speed-up. *Scientific Reports*, 9(1):6903, 2019.

[Gol70] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[GORK+19] D. Guéry-Odelin, A. Ruschhaupt, A. Kiely, E. Torrontegui, S. Martínez-Garaot, and J. G. Muga. Shortcuts to adiabaticity: Concepts, methods, and applications. *Rev. Mod. Phys.*, 91:045001, Oct 2019.

[Got97] Daniel Gottesman. Stabilizer codes and quantum error correction. *arXiv preprint quant-ph/9705052*, 1997.

[Gro96] Lov K Grover. A fast quantum mechanical algorithm for database search. *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*, pages 212–219, 1996.

[GS18] David J. Griffiths and Darrell F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, 3 edition, 2018.

[GS19] Gian Giacomo Guerreschi and Vadim N Smelyanskiy. Variational quantum computation of excited states. *Physical Review Research*, 1(3):033062, 2019.

[GW95] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 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.

[Har21] 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.

[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+19a] 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. *Nature*, 567(7747):209–212, 2019.

[HCT+19b] 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. *Nature*, 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.

[HHZ19] Markus Heyl, Philipp Hauke, and Peter Zoller. Quantum localization bounds trotter errors in digital quantum simulation. *Science Advances*, 5(4), 2019.

[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.

[HMM+20] Kosuke Hashimoto, Sam Miller, Mario Motta, Nicholas C Rubin, Ryan Babbush, and Hartmut Neven. Expressibility of the alternating layered ansatz for quantum chemistry. *arXiv preprint arXiv:2008.02354*, 2020.

[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.

[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.

[HSSC08] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[HWG+22] Pavel Hrmo, Benjamin Wilhelm, Lukas Gerster, Martin W. van Mourik, Marcus Huber, Rainer Blatt, Philipp Schindler, Thomas Monz, and Martin Ringbauer. Native qudit entanglement in a trapped ion quantum processor. *arXiv e-prints*, page arXiv:2206.04104, June 2022.

[HWG+23] Pavel Hrmo, Benjamin Wilhelm, Lukas Gerster, Martin W. van Mourik, Marcus Huber, Rainer Blatt, Philipp Schindler, Thomas Monz, and Martin Ringbauer. Native qudit entanglement in a trapped ion quantum processor. *Nature Communications*, 14:2242, April 2023.

[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.

[Jor23] Stephen Jordan. The quantum algorithm zoo. `https://quantumalgorithmzoo.org/`, Accessed: 2023. Accessed: 2023.

[Kar72] Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103, Boston, MA, 1972. Springer.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980, December 2014.

[KEA+23] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout Van Den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, et al. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965):500–505, 2023.

[KG15] Richard Kueng and David Gross. Qubit stabilizer states are complex projective 3-designs. *arXiv e-prints*, page arXiv:1510.02767, October 2015.

[Kit02] Alexei Yu Kitaev. Classical and quantum computation. *Uspekhi Mat. Nauk*, 52(6):53–112, 2002.

[Kit03] Alexei Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303(1):2–30, 2003.

[Kit15] Alexei Kitaev. A simple model of quantum holography. Talks at KITP, April 7, 2015 and May 27, 2015., 2015.

[KMN07] Pieter Kok, William J Munro, and Kae Nemoto. Linear optical quantum computing. *Reviews of Modern Physics*, 79(1):135–174, 2007.

[KMT+17a] 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. *Nature*, 549(7671):242–246, September 2017.

[KMT+17b] 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. *Nature*, 549(7671):242–246, September 2017.

[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.

[KSC+19] Sami Khairy, Ruslan Shaydulin, Lukasz Cincio, Yuri Alexeev, and Prasanna Balaprakash. Learning to Optimize Variational Quantum Circuits to Solve Combinatorial Problems. *arXiv e-prints*, page arXiv:1911.11071, November 2019.

[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.

[LJG+21] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. Theory of overparametrization in quantum neural networks. *arXiv e-prints*, page arXiv:2109.11676, September 2021.

[LLL20] Daniel Liang, Li Li, and Stefan Leichenauer. Investigating quantum approximate optimization algorithms under bang-bang protocols. *Phys. Rev. Research*, 2:033402, Sep 2020.

[LMR+17] Norbert M Linke, Dmitri Maslov, Martin Roetteler, Shantanu Debnath, Caroline Figgatt, Kevin A Landsman, Kenneth Wright, and Christopher Monroe. Experimental comparison of two quantum computing architectures. *Proceedings of the National Academy of Sciences*, 114(13):3305–3310, 2017.

[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.

[MC20] Matija Medvidovic and Giuseppe Carleo. Classical variational simulation of the Quantum Approximate Optimization Algorithm. *arXiv e-prints*, page arXiv:2009.01760, September 2020.

[MD19] Joshua Morris and Borivoje Dakić. Selective Quantum State Tomography. *arXiv e-prints*, page arXiv:1909.05880, September 2019.

[MEAG+20] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C Benjamin, and Xiao Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1):015003, 2020.

[Mey21] Johannes Jakob Meyer. Fisher Information in Noisy Intermediate-Scale Quantum Applications. *Quantum*, 5:539, September 2021.

[Mi 21] Xiao Mi *et al.* Information Scrambling in Computationally Complex Quantum Circuits. *arXiv e-prints*, page arXiv:2101.08870, January 2021.

[MNKF18] Koji Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 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, 2016.

[NC00] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[NC02] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[NCV+21] Antoine Neven, Jose Carrasco, Vittorio Vitale, Christian Kokail, Andreas Elben, Marcello Dalmonte, Pasquale Calabrese, Peter Zoller, Benoît Vermersch, Richard Kueng, and Barbara Kraus. Symmetry-resolved entanglement detection using partial transpose moments. *npj Quantum Information*, 7(1):152, 2021.

[NLR20] Murphy Yuezhen Niu, Han-Hsuan Li, and Michael P Reisenberger. Qudit-based quantum error correction. *Physical Review Research*, 2(2):023063, 2020.

[NM65] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[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.

[NW06] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science and Business Media, 2006.

[OBK+16] Patrick J.J. O'Malley, Ryan Babbush, Ian D. Kivlichan, Jonathan Romero, Jarrod R. McClean, Rami Barends, Julian Kelly, Pedram Roushan, Andrew Tranter, Nan Ding, et al. Scalable quantum simulation of molecular energies. *Physical Review X*, 6(3):031007, 2016.

[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.

[Par62] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076, 09 1962.

[PBB+20] Guido Pagano, Aniruddha Bapat, Patrick Becker, L. Pedro Garcińa-Pintos, Paul W. Hess, Harvey B. Kaplan, Antonis Kyprianidis, Wen Lin Tan, Erik Birckelbaw, Miguel A. Morales Hernandez, et al. Quantum approximate optimization with a trapped-ion quantum simulator. *Proceedings of the National Academy of Sciences*, 117(41):25396–25401, 2020.

[PK19] Marco Paini and Amir Kalev. An approximate description of quantum states. *arXiv e-prints*, page arXiv:1910.10543, October 2019.

[PMS+14a] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, AlÃ!'n Aspuru-Guzik, and Jeremy L. Oâ Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1), Jul 2014.

[PMS+14b] 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*, 5:4213, July 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.

[Pow94] Michael JD Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. *Advances in optimization and numerical analysis*, 275:51, 1994.

[Pre18] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

[PSW06] Sandu Popescu, Anthony J. Short, and Andreas Winter. Entanglement and the foundations of statistical mechanics. *Nature Physics*, 2(11):754–758, November 2006.

[RBM+17] Jonathan Romero, Ryan Babbush, Jarrod R McClean, Cornelius Hempel, Peter J Love, and Alán Aspuru-Guzik. Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz. *Quantum Science and Technology*, 4(1):014008, 2017.

[RBMV21] Aniket Rath, Cyril Branciard, Anna Minguzzi, and Benoît Vermersch. Quantum Fisher Information from Randomized Measurements. *Physical Review Letters*, 127(26):260501, December 2021.

[RMP+22] Martin Ringbauer, Michael Meth, Lukas Postler, Roman Stricker, Rainer Blatt, Philipp Schindler, and Thomas Monz. A universal qudit quantum processor with trapped ions. *Nature Physics*, 18(9):1053–1057, July 2022.

[Ros56] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837, 09 1956.

[SBG+19] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, Jonas Meyer, Koji Mitarai, Florian Wilhelm, and Ville Bergholm. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

[SBSW20] Maria Schuld, Alex Bocharov, Krysta M. Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, March 2020.

[Sch11a] 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.

[Sch11b] Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96–192, January 2011.

[Sha70] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[Sho94] Peter W Shor. Algorithms for quantum computation: Discrete logarithms and factoring. pages 124–134, 1994.

[Sho95] Peter W Shor. Scheme for reducing decoherence in quantum computer memory. *Physical Review A*, 52(4):R2493–R2496, 1995.

[SIKC20] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum Natural Gradient. *Quantum*, 4:269, May 2020.

[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.

[SKPK19] Adam Smith, M. S. Kim, Frank Pollmann, and Johannes Knolle. Simulating quantum many-body dynamics on a current digital quantum computer. *npj Quantum Information*, 5:106, November 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(2):020365, June 2022.

[SP17] Dries Sels and Anatoli Polkovnikov. Minimizing irreversible losses in quantum systems by local counterdiabatic driving. *Proceedings of the National Academy of Sciences*, 114(20):E3909–E3916, 2017.

[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. *Quantum*, 5:491, July 2021.

[SS21b] Stefan H. Sack and Maksym Serbyn. Quantum annealing initialization of the quantum approximate optimization algorithm. *arXiv e-prints*, page arXiv:2101.05742, January 2021.

[SS21c] Stefan H. Sack and Maksym Serbyn. Quantum annealing initialization of the quantum approximate optimization algorithm. *Quantum*, 5:491, July 2021.

[SSL19] R. Shaydulin, I. Safro, and J. Larson. Multistart methods for quantum approximate optimization. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–8, 2019.

[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.

[SZZ+17] Yudong Shen, Xiang Zhang, Shuaining Zhang, Jun-Nan Zhang, Man-Hong Yung, and Kihwan Kim. Quantum implementation of the unitary coupled cluster for simulating molecular electronic structure. *Physical Review A*, 95(2):020501, 2017.

[Tay06] John R. Taylor. *Quantum Chemistry*. University Science Books, 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.

[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.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake

VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[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.

[W+19] K. Wright et al. Benchmarking an 11-qubit quantum computer. *Nature Communications*, 10:5464, November 2019.

[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.

[WL21] Jonathan Wurtz and Peter Love. MaxCut quantum approximate optimization algorithm performance guarantees for p >1. *Physical Review A*, 103(4):042612, April 2021.

[WL22] Jonathan Wurtz and Peter J. Love. Counterdiabaticity and the quantum approximate optimization algorithm. *Quantum*, 6:635, January 2022.

[Wu82] F. Y. Wu. The potts model. *Rev. Mod. Phys.*, 54:235–268, Jan 1982.

[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.

[WWJ+20] Madita Willsch, Dennis Willsch, Fengping Jin, Hans De Raedt, and Kristel Michielsen. Benchmarking the quantum approximate optimization algorithm. *Quantum Information Processing*, 19(7):197, June 2020.

[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.

[YRS+17] Zhi-Cheng Yang, Armin Rahmani, Alireza Shabani, Hartmut Neven, and Claudio Chamon. Optimizing variational quantum algorithms using pontryagin's minimum principle. *Phys. Rev. X*, 7:021027, May 2017.

[Zhu17a] Huangjun Zhu. Multiqubit clifford groups are unitary 3-designs. *Phys. Rev. A*, 96:062336, Dec 2017.

[Zhu17b] Jingbo B Zhu. Multilevel quantum systems: Mathematical foundation, quantum gates and algorithms. *Physics Reports*, 700:1–57, 2017.

[ZWC+18] 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. *arXiv e-prints*, page arXiv:1812.01041, December 2018.

[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.