Heinrich-Heine-Universität Düsseldorf

Complexity of Variational Quantum Algorithms

Inaugural dissertation

presented to the Faculty of Mathematics and Natural Sciences
of the Heinrich-Heine-Universität Düsseldorf
for the degree of

Doctor of Natural Sciences (Dr. rer. nat.)

by

Lennart Vincent Bittel

from Darmstadt

Düsseldorf, April 2023

From the Institute for Theoretical Physics III
at the Heinrich-Heine-Universität Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich-Heine-Universität Düsseldorf

Supervisor: Prof. Dr. Martin Kliesch

Second corrector: Prof. Dr. Dagmar Bruß

Date of the oral examination:

Declaration

Ich versichere an Eides statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist.

Düsseldorf, November 13, 2023

Lennart Bittel

Abstract

As the development of quantum computers progresses rapidly and larger physical chips become available, an important next step is to demonstrate their usefulness as computational devices. Unfortunately, current implementations operate below the error-correction threshold, which means that algorithms are limited to short gate sequences due to the gate noise. Established algorithms such as Shor's algorithm are therefore not yet implementable. As researchers, we are interested in a meaningful *quantum advantage*, i.e. performing a computational task on a quantum device for which the best known classical algorithms have infeasibly long runtimes. Prominent candidates are variational quantum algorithms (VQAs), which are a hybrid quantum-classical approach of solving an optimization problem. Here, a classical computer can choose tunable parameters of a quantum circuit which creates a variational state in order to minimize the expectation value of some cost observable. VQAs can be used for both classical optimization problems as well as finding the ground state energy of some quantum Hamiltonian.

In this thesis, we outline some challenges that VQAs must overcome to become a viable tool. Namely, there is the problem that the optimization converges to suboptimal local minima of the cost function. In our work, we show that the classical training of VQAs is NP-hard. This implies that, at least for some cases, no polynomial-time algorithm can converge to the global minima (assuming $\mathsf{P} \neq \mathsf{NP}$).

For VQAs, it is also important to find short circuit implementations to suppress the physical noise in the system and make their implementation feasible on near-term hardware. This means that the VQA ansatz must be expressive enough to approximate the ground-state energy while still maintaining low complexity. In our work, we show that finding the shortest circuit depth implementation is QCMA-hard, even if one only wants to get close within multiplicative factor scaling with the input size. Finally, there is the problem of the measurement effort required for VQAs. Here, the estimation of the gradient with respect to the tunable parameters can act as a bottleneck, mainly because the shot-noise statistics require multiple rounds of measurement. To alleviate this problem, we propose a gradient estimation routine based on a Bayesian framework to reduce the overall measurement effort. We motivate and numerically show that, for well-studied VQA proposals, the strategy can significantly reduce the number of measurement rounds while maintaining the same gradient quality.

Zusammenfassung

Da die Entwicklung von Quantencomputern rasch voranschreitet und immer größere Systeme verfügbar werden, besteht ein wichtiger nächster Schritt darin, ihren Nutzen zu demonstrieren. Leider arbeiten die derzeitigen Implementierungen unterhalb der Fehlerkorrekturschwelle, was bedeutet, dass mögliche Algorithmen durch kurze Gattersequenzen begrenzt sind. Etablierte Algorithmen wie der Shor-Algorithmus können daher noch nicht implementiert werden. Als Forscher sind wir an einem sinnvollen Quantenvorteil interessiert, d.h. an der Lösung eines Problems auf einem Quantengerät, für das die besten bekannten klassischen Algorithmen unausführbar lange Laufzeiten hätten. Ein Kandidat hierfür sind Variationsquantenalgorithmen (VQAs), die einen hybriden quanten-klassischen Ansatz zur Lösung eines Optimierungsproblems darstellen. Ein klassischer Computer kann Parameter eines Quantenschaltkreises wählen, die einen Variationszustand erzeugen um damit den Erwartungswert einer Observablen zu minimieren. VQAs können sowohl für klassische Optimierungsprobleme als auch die Schätzung der Grundzustandsenergie eines Quanten-Hamiltonoperators verwendet werden.

In dieser Arbeit analysieren wir einige Herausforderungen, die VQAs überwinden müssen, um ein nützliches Werkzeug zu werden. Ein Problem ist, dass die Optimierung zu suboptimalen lokalen Minima der Kostenfunktion konvergieren kann. Wir zeigen, dass das klassische Training von VQAs NP-schwer ist. Dies impliziert, dass kein Polynomialzeitalgorithmus immer gegen globale Minima konvergieren kann (unter der Annahme, dass $\mathsf{P} \neq \mathsf{NP}$).

Für VQAs ist es auch wichtig, kurze Quantenschaltkreise zu finden, damit ihre Implementierung auf zeitnah verfügbarer Hardware möglich wird. Dies bedeutet, dass der VQA-Ansatz in der Lage sein muss die Grundzustandsenergie zu approximieren, aber gleichzeitig eine geringe Komplexität aufweisen muss. Wir zeigen, dass es QCMA-schwer ist, eine Implementierung mit der geringsten Schaltkreistiefe zu finden, selbst wenn man sie nur bis auf einen multiplikativen Faktor finden möchte. Schließlich gibt es noch das Problem des Messaufwands, der für VQA Optimierungen erforderlich ist. Die Schätzung des Gradienten in Abhängigkeit von den Parametern kann sehr zeitaufwendig sein, auch weil aufgrund der Messstatistik mehrere Messrunden pro Messeinsteillung erforderlich sind. Um den Messaufwand zu reduzieren, entwickeln wir eine Gradientenschätzungsroutine, die auf einem Bayes'schen Rahmenwerk basiert. Wir motivieren und zeigen numerisch, dass diese Strategie die Anzahl der Messrunden bei gleichbleibender Gradientenqualität erheblich reduzieren kann.

Acknowledgements

I would like to thank everyone I met along the way. Especially my supervisor Martin Kliesch as well as my co-authors Sevag Gharibian, Jens Watty and Alex Gresch for the productive work we had. Also everyone else in our group, Raphael Brieger, Christopher Cedzich, Markus Heinrich, Nikolai Miklin, Mirko Arienzo, Pascal Baßler, Matthias Zipper, Salwa Shaglel, Juan Henning and Jonathan Schluck. I would also like to thank Dagmar Bruß who created a welcoming atmosphere, especially when our group was still small. This also includes all the people from her group, Hermann Kampermann, Lucas Tendic, Gláucia Murta, Giacomo Carrara, Thomas Wagner, Federico Grasselli, Sarnava Datta, Nikolai Wyderka, Julia Kunzelmann, Justus Neumann, Carlo Liorni, Giulio Gianfelici, Timo Holz, Felix Bischof and Daniel Miller with whom I have had stimulating discussions, shared memories at conferences and fun card game nights. I would also like to thank the other members of the institute Jens Bremer, Claudia Stader, Cornelia Glowacki and Cordula Hoffjan. Finally I would like to thank my parents Renate and Robert as well as my sister Nele and my girlfriend Johanna for their support throughout.

Contents

1 Introduction

Quantum information science is still a relatively new and rapidly evolving field. It has been described as a new paradigm in computing, where quantum computers can harness the power of quantum parallelism and entanglement to perform classically impossible operations. Unfortunately, we can only probe parts of the quantum state through destructive measurements, which can turn an entangled state into an effectively unusable, random classical outcome. Thus, good quantum algorithms must implement a quantum state in superposition, which constructively interferes to have a high probability of obtaining the desired classical result after measurement. This restriction is at the heart of quantum computers and the reason why we cannot simply think of them as more powerful classical computers, but instead need a nuanced understanding of their powers and limitations.

A good starting point is the idea of using quantum computers to simulate quantum systems. This was also the application proposed by Richard Feynman in 1981 [1]. The most successful approaches along these lines are systems of ultracold atoms [2, 3], which have been used successfully to simulate many different condensed-matter Hamiltonians and study their properties such as superfluidity or conductivity. However, these experiments are specialized for specific Hamiltonians and cannot be used to simulate any physical system.

To this end, we want a *universal quantum computer*, i.e. a device that can be used to perform arbitrary tasks. To this end, researchers have defined a quantum version of a Turing machine, which encapsulates the idea of a quantum computation. Since quantum computers can also simulate any classical computation efficiently, this new paradigm is at least as powerful as a classical computer. However, since any known simulation of quantum mechanics requires exponentially many resources, a quantum Turing machine could be much more powerful, leading to exponential speedups in runtime.

Further theoretical discoveries were needed to develop consistent model of quantum computing. In particular, the Solovay–Kitaev theorem [4] showed that, as long as the available gate operations are *universal*, one can translate between two gate-sets with at most poly-logarithmic time overhead.

The next key ingredient was to show that quantum error correction is indeed feasible, i.e. using multiple noisy qubits to encode a single logical qubit, but with significantly suppressed error rates. This effort culminated in the threshold theorem [5], which shows that one can suppress errors in quantum computation with

only poly-logarithmic overhead in resources, provided that the gate errors were sufficiently small to begin with. This means that real-world implementations, which will always contain some errors, can simulate an idealized quantum Turing machine with moderate overhead. In practice, the ratio of physical qubits to logical qubits may need to be on the order of $1000 : 1$, meaning that current implementations are only now starting to be able to perform first experiments with a single logical qubit [6]. The developed framework allows to define a computational complexity class called BQP, which describes all problems that can be solved on a quantum computer using at most polynomially many qubits and gate operations.

The key question is whether one can show that this quantum class is more powerful than the corresponding classical complexity class BPP, which describes problems that can be solved on a classical computer. Since an affirmative result would show an unconditional separation between the complexity classes P and PSPACE, an important open question in computer science [7], one has to settle for candidate problems that we know can be solved efficiently on a quantum computer, but we have reason to believe cannot be efficiently solved on a classical computer. Almost 30 years ago, Peter Shor [8] identified a family of problems in number theory, most notably prime factorization and discrete logarithms, that could be solved in polynomial time on a quantum computer, but for which the best classical algorithm, the General Number Field Sieve (GNFS) [9], requires superpolynomial time, meaning that the runtime grows faster than any polynomial. These problems are used in most public key encryption algorithms in use today (RSA, Diffie-Hellman). The fact that no one has been able to find an efficient classical algorithm to solve these problems, despite a great deal of interest, is a strong indication that quantum computers are indeed significantly more powerful than classical computers. However, it should also be noted that the list of candidate problems has not grown significantly in the past decades after these discoveries.

Before jumping to the conclusion that quantum computing is only a niche tool for solving a handful of problems in cryptography, it should be noted that there are complexity-theoretic reasons for this. As humans, we are mostly interested in problems that we can verify, that is, we need to be able to evaluate a proposed solution as either good or bad. Naturally these problems are encapsulated by the complexity class NP. It so happens that many studied problems in NP are either solvable in polynomial time (P) or NP-complete, i.e. as hard to solve as the most difficult problems in NP. Only a handful of problems are so-called NP-intermediate, not known to belong to either of the two groups. It is conjectured that quantum computers cannot solve NP-complete problems [10] and since we do not need quantum computers to solve problems for which efficient classical algorithms already exist, NP-intermediate problems are the only remaining natural candidates for a quantum speedup, some of which have been shown to run efficiently on a quantum

computer. For other NP-intermediate problems like graph isomorphism [11], even with considerable research, there is no clear indication that quantum computers can help solving these problems.

If we want to find more types of problems, where quantum computers will excel, we need problems that can only be verified on a quantum computer. Natural candidates are problems related to the time evolution of quantum systems [12]. For example, we believe that questions like phase classifications/transitions, topological transition, thermalization properties and many other problems are all efficiently solvable only on a quantum computer [13]. There are also efficiently solvable problems not directly related to quantum mechanics, most notably the computation of certain Jones polynomials [14] used in knot theory, which we believe cannot even be efficiently verified on a classical computer, but are known to be solvable on a quantum computer.

While complexity theory may give us insights into the nature of quantum computation in the long run, on a more near term basis, a direct comparison with known classical algorithms is very relevant. Known under the term *quantum advantage*, we want to find a specific problem instance that an existing quantum device can solve, but for which the best known classical algorithms would take an infeasible amount of time.

Currently available implementations operate above the error-correction threshold and a fully error corrected quantum device is still some time away. Since error correction is believed to be a requirement for quantum devices to implement a Quantum Fourier Transform (QFT), using Shor's algorithm is currently impossible. To obtain near term quantum advantage, we therefore must ask questions about the power of non-error-corrected quantum computation. The time frame before fully error- corrected quantum computers are available is called Noisy Intermediate-Scale Quantum (NISQ) [15]. Much is still unknown about the computational complexity of quantum devices without error correction. There is evidence that the task of sampling from certain quantum distributions is indeed difficult [16] even in the presence of noise and there have also been experimental implementations of this random circuit sampling [17]. This type of quantum advantage may be undesirable however, since it may be difficult to verify that the desired distribution was actually sampled [18].

Another interesting contender are Variational Quantum Algorithms (VQAs), which are a proposed framework for a hybrid, quantum-classical setup to solve either combinatorial or ground state Hamiltonian problems. We will introduce them in more detail in the following sections. In this context, a quantum advantage would be a VQA experiment that finds a better approximation of the ground-state energy of

a Hamiltonian, than what commonly used classical energy estimation algorithms can obtain. However, it is not yet clear whether VQAs can actually be used on near-term devices to gain a meaningful quantum advantage. In this thesis we explore the viability of VQAs as a near term application of NISQ quantum computers as well as the challenges involved.

Thesis Structure

This thesis is organised as follows:

- In chapter 2, we introduce the basics of complexity theory as well as the relevant complexity classes that we are interested in. Since this work requires a theoretical understanding of quantum computers, we also motivate the idea of a quantum Turing machine with an introduction to quantum mechanics and universal quantum computation.

- In chapter 3, we introduce VQAs, which describe a particular experimental setup that uses a hybrid/quantum classical approach to solve combinatorial and quantum chemical problems.

- Chapter 4 contains an overview of the difficulties that can arise when solving VQA optimization problems. We also present our result, which includes both an analysis of the required measurement effort and how to mitigate it, as well as complexity theoretic hardness arguments related to the optimization problem. This chapter also summarizes additional work done using machine learning approaches to predict MBL phase transitions.

- In chapter 5, we conclude this thesis as well as mention remaining open question and future research directions.

- The appendices A to D include the articles that were published during the dissertation.

2 Introduction to complexity theory

Complexity theory is the study of how hard it is to perform a certain mathematical task. Even before the advent of computers, we can think of multiplication by hand and long division as computational algorithms to perform arithmetic operations on numbers. For large numbers, it is more convenient to use a calculator, which has made multiplication and division a non-issue for practical purposes. This is an example of how increasing the available computational power has made a particular task easier. Today's computers operate in the Tera FLOPS range, which means they can perform a trillion multiplications/divisions per second. Alternatively, instead of increasing the computational power, one can also find a better strategy to tackle the algorithm in question. For multiplication, for example, there are algorithms that require significantly fewer operations [19–21] than the naive multiplication by hand. However, since they are quite complicated, these approaches become practical only for very large numbers, which is why they are seldom used in practice.

Practically relevant speedups in computing time are often celebrated. For instance, the Fast Fourier Transform (FFT) [22] is an algorithm that implements the discrete Fourier transform in almost linear runtime with the number of elements, whereas a naive implementation scales quadratically. In general, finding algorithms with short running times is a very important goal. Especially for complicated problems, it may be necessary to find efficient algorithms to make finding a solution practically feasible, even with all the existing increases in computational power. It is not always obvious which problems are easy to compute and which are hard. For example, finding the shortest route through a maze can be done efficiently, but finding the shortest route that passes through a given list of locations at least once, is potentially very hard. For the latter problem, called the Traveling Salesman Problem (TSP), the best known algorithms that guarantee to return the shortest route have running times that scale exponentially with the number of locations. While for many practical situations, there are decently fast approximation algorithms that find close to optimal routes [23], for the exact case, it is still unknown if significantly faster algorithms exist. Then there are also problems that we know require exponentially many resources to solve. Here, complexity theory helps to group certain computational problems and to find the relationship between them. The underlying framework also allows us to ask mathematically precise questions about the nature of computation.

In the following sections we present the complexity theoretic framework that has been developed to analyze computational problems as well as introduce certain complexity classes, including quantum complexity classes, that are used in this thesis.

2.1 Definition of decision problems and Turing machines

In this dissertation we focus on decision problems, which are problem that amount to answering yes/no questions. Thus, a decision problem assigns either a YES or a NO value to an input $I \in S$ from a set of allowed inputs S. Generally the problems are defined with an alphabet Σ. For example, we can choose a binary alphabet $\Sigma = \{0, 1\}$. The input can be an arbitrary string

$$I \in S \subset \bigcup_{N \in \mathbb{Z}_+} \Sigma^N =: \Sigma^* \tag{2.1}$$

here the associated N ($I \in \{0,1\}^N$) is called the input, or instance length of I. This allows the first, quite general definition of a decision problem.

Definition 1 (Decision Problem). *A decision problem X is a bipartition of the set of all inputs*

$$S \subset \bigcup_{N \in \mathbb{Z}_+} \Sigma^N =: \Sigma^* \tag{2.2}$$

into a set of YES instances ($S_y \subset S$) and NO instances ($S_n \subset S$). If $I \in S_y$, we say that the input I is in the language *of X.*

Since we can encode any input in binary strings, we can also think of decision problems as a function on a subset of a binary strings to YES/NO or $1/0$. An algorithm that solves the decision problem is therefore an implementation of this function.

While allowing only YES/NO answer type problems seems restrictive, for more general problems it is often possible to define decision versions with closely related complexity.

For example we can have *function problems*, where the goal is to compute some function

$$F : \{0,1\}^N \to \{0,1\}^M \tag{2.3}$$
$$x \mapsto F(x)\,. \tag{2.4}$$

By defining a decision problem "Is the value of the a-th bit of $F(x)$ a 1?" (i.e. $F(x)_a = 1$?) with input $I = (x, a)$, we can obtain $F(x)$ after M rounds of questioning.

Another example are *optimization problems*, where the goal is to minimize some cost function $F : \{0,1\}^N \to \mathbb{R}$. We can define the decision problem "Does an $x \in \{0,1\}^N$ exist, such that $F(x) \leq \alpha$?" with input $I = (x, \alpha)$. A binary search strategy can find the minimum value of F with exponential precision in the number of calls to the decision problem. For this reason, although the problems we are interested in are mostly optimization problems, we can instead consider their respective decision versions.

As a model of computation we use a (deterministic) Turing machine, which is a concept similar to the execution of computer code on a powerful computer. A Turing machine is described by

1. an internal state $\Xi \in \Sigma^K$ of fixed alphabet Σ and size K,

2. an infinite tape which can be accessed with a *head* that can exchange bits between the tape and the internal state as well as move along it,

3. a transition function (δ), which depending on Ξ can perform operations on the head and change the internal state Ξ. If no transition is defined, the Turing machine halts.

Starting from an initial state Ξ_0, the Turing machine manipulates the tape according to its transition function. The input I can be encoded as the initial state on the infinite tape (single tape Turing machine) or on a second tape to which the Turing machine also has access to (multi tape Turing machine). If the Turing machine halts, there is also an additional condition which makes it either accept or reject the input I.

The idealization of an infinitely long tape is advantageous, because it allows a single Turing machine to handle arbitrary input size N. An example of a two tape Turing machine is shown in fig. 2.1. We will not be concerned about the precise architecture of the machine, but only require Turing completeness, which means that the machine

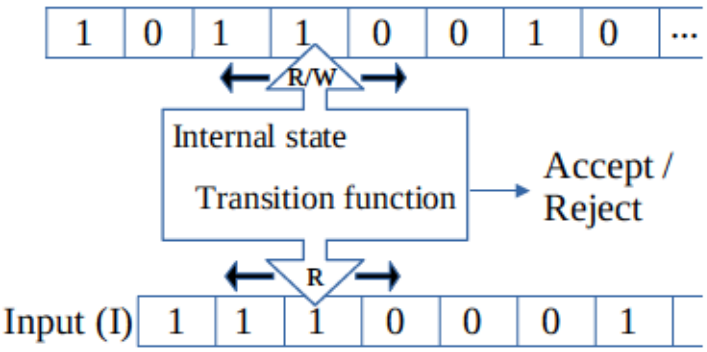

Fig. 2.1. – A two tape Turing machine. The Turing machine gets the input I on the input tape. It can read bits of I into its finite internal state, or move along the tape. Similarly, it has access to another working tape of infinite size, on which it can read, as well as write and move along. Finally, the machine can also terminate its operation to accept/reject the input I. Which operation is performed is determined by the Turing machine's internal transition function, which can be any function acting on the internal state and the tip.

can simulate any other Turing machine. For a more detailed introduction into Turing machines see [24].

When we consider computational resources such as runtime (the number of transition steps performed by the Turing machine before it terminates) or required memory (the size of the infinite tape that was accessed during the computation), they can strongly depend on the precise architecture of Turing machine used. An example of this is to consider how information is stored. In a single tape model, similar to cassettes, one assumes a long tape where the tip physically moves along the tape, meaning that accessing information from a bit in memory that is n positions takes a linear ($O(n)$) amount of time. In contrast, many algorithms assume random-access memory (RAM), which describes a memory system where any bit of information can be accessed in constant ($O(1)$) time. Since accessing memory is present in almost all algorithms, machines with a tape memory may experience significant time overhead compared to those using RAM. Similar polynomial speedups are possible between single and multiple tape Turing machines. In order to have an agnostic definition of complexity, we compare resource use only up to polynomial overheads.

With this understanding, we are able to define our first complexity classes. In fig. 2.2, all the complexity classes considered in this thesis are shown, as well as their relationship to each other.

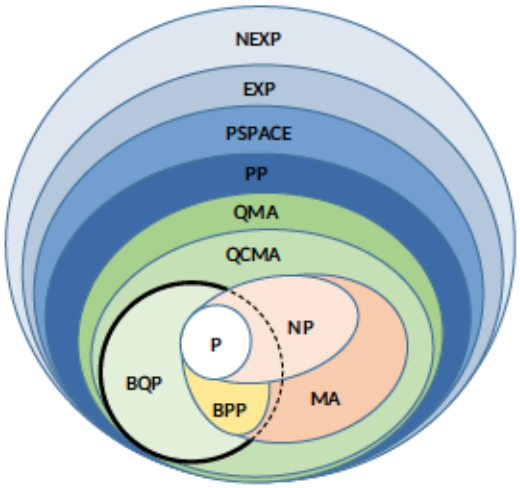

Fig. 2.2. – Complexity classes shown as a Venn diagram to show inclusions. Unconditional proper subsets are only proven between P and EXP as well as NP and NEXP. The black circle (BQP) shows all problems we call efficiently solvable, meaning they can be solved in polynomial time on a quantum computer.

2.2 Basic complexity classes

Using the framework of deterministic Turing machines, we can define the first complexity classes. First we consider time complexity.

Definition 2 (Complexity class P (or PTIME)). *A decision problem is in* P*, if there exists a deterministic Turing machine (TM), such that for all inputs $I \in S$ of size N, TM terminates in polynomial time $O(\text{poly}(N))$ and T accepts when I is a YES instance, but TM rejects when I is a NO instance.*

Here polynomial time $O(\text{poly}(N))$ means that there exists a problem size independent exponent α, such that the Tuning machine terminates in at most $O(N^\alpha)$ time for all $I \in S$. We use the big-O notation to describe the scaling behavior. $g(n) = O(f(n))$ means that for large n ($\forall n \geq n_0$) there exists a constant $\beta \in \mathbb{R}$ such that $g(n) \leq \beta \times f(n)$. From a complexity theoretic point of view, we call any algorithm with a polynomial running time *efficient*, which is known as Cobham's thesis [25]. Note that if n_0, α or β are very large, the algorithm may still be practically infeasible, so this is more a theoretical term, with only limited direct practical implications. We can also define the class of problems which can be decided with at most exponential runtime.

Definition 3 (Complexity class EXP (or EXPTIME)). *A decision problem is in* EXP, *if there exists a deterministic Turing machine (TM), such that for all inputs $I \in S$ of size N, TM terminates in exponential time $O(2^{\text{poly}(N)})$ and TM accepts when I is a YES instance, but TM rejects when I is a NO instance.*

As a consequence of the time hierarchy theorem [26] we know that P is a proper subset of EXP, i.e. there are computational problems that can be solved only in exponential time and for which no polynomial-time algorithm exists. Similarly, one can define the class 2EXP for problems where the Turing machine terminates in doubly exponential time $O(2^{2^{\text{poly}(N)}})$, or in general nEXP for $n \in \mathbb{N}$- times exponential time. The class R, or *decidable* problems, describes problems for which there exists a Turing machine that terminates with the correct result, without any time limit. This gives the inclusion

$$\mathsf{P} \subsetneq \mathsf{EXP} \subsetneq \mathsf{2EXP} \subsetneq \mathsf{3EXP} \subsetneq \cdots \subsetneq R. \tag{2.5}$$

Notably not every decision problem is in R. The halting problem, which asks whether a Turing machine with a input I terminates in a finite amount of time or never terminates is not in R. This is shown by a proof of contradiction: We consider a Turing machine T which takes as input a description of another Turing machine $I = \tilde{T}$, where T only halts if $\tilde{T}$ does not halt and otherwise loops forever. Does T halt if it is given itself $\tilde{T} = T$ as input? Since T cannot both halt and not halt, this implies that it is not possible to construct T. The only way how this can be true is if there cannot exist a Turing machine which solves the halting problem for all instances. This proof shows that there are well defined problems that cannot be solved by a Turing machine.

When instead of time, we consider the amount of tape accessed during the algorithm, we can define another class known as PSPACE.

Definition 4 (Complexity class PSPACE). *A decision problem is in* PSPACE, *if there exists a deterministic Turing machine (TM), such that for all inputs $I \in S$ of size N, TM terminates requiring only polynomial $O(\text{poly}(N))$ bits of tape and TM accepts when I is a YES instance, but TM rejects when I is a NO instance.*

Similarly, the class EXPSPACE contains problems that require at most exponential memory and in general the class nEXPSPACE requires at most n-times exponential memory. We also have the following inclusion

$$\mathsf{P} \subseteq \mathsf{PSPACE} \subseteq \mathsf{EXP} \tag{2.6}$$

because a Turing machine can only access polynomially many bits in polynomial time and because a Turing machine which accesses at most $M = O(\text{poly}(N)$ bits of memory can only be in at most $O(2^M) = O(2^{\text{poly}(N)})$ different states before it has to return to a previous state. As such, it needs to terminate within $O(2^{\text{poly}N})$ time steps as it would otherwise loop forever. We do not know if these relations are proper subsets. While it is strongly assumed to be the case, our known proof techniques have been unable to show an unconditional separation of these classes. This also holds for all sub-classes of PSPACE which we will consider moving forward.

Another complexity class is NP, which stands for nondeterministic polynomial time. The name originates from considerations about nondeterministic Turing machines, which we will not consider here. But there is also an alternative definition, which describes problems where the YES instances can be verified with an additional proof.

Definition 5 (Complexity class NP). *A decision problem is in* NP*, if there exists a deterministic Turing machine (TM), such that for all inputs $I \in S$ of size N, TM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a Yes instance ($I \in S_y$), there exists a proof $p \in \{0,1\}^{M=O(\text{poly}(N))}$ such that TM accepts on input $I' = (I, p)$.*
- *a NO instance ($I \in S_n$), for all $p \in \{0,1\}^M$, TM rejects on input $I' = (I, p)$.*

The class describes problems of the type: "Does there exist a solution to this problem?", or "$\exists x : F(x) = \text{TRUE}$?", where the function F can be computed in polynomial time. In the YES case, one can verify that a proposed solution $p = x^*$ is indeed correct, i.e. verify that $F(x^*) = \text{TRUE}$ holds. In the NO case all proofs are rejected because, by definition, no such solution exists. An important example of NP problems is the decision version of many optimization problems ("$\exists x : F(x) \leq \alpha$?") . Naturally, one can also define the class NEXP, where the proof string can be exponentially large and the Turing machine can take exponentially long to terminate. With the time hierarchy theorem, we also know that NEXP is strictly greater than NP. We also have the inclusion

$$\text{P} \subseteq \text{NP} \subseteq \text{PSPACE}\,, \tag{2.7}$$

since P is the class NP with an empty proof p and an algorithm can simply enumerate all possible proofs, which takes exponential time, but only requires polynomial space.

A related class is coNP, which describes decision problems where one can verify the NO instances efficeintly.

Definition 6 (Complexity class coNP). *A decision problem is in* coNP*, if there exists a deterministic Turing machine (TM), such that for all inputs $I \in S$ of size N, TM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a NO instance ($I \in S_n$), there exists a proof $p \in \{0,1\}^{M=O(\text{poly}(N))}$ such that TM accepts on input $I' = (I, p)$.*
- *a YES instance ($I \in S_y$), for all $p \in \{0,1\}^M$, TM rejects on input $I' = (I, p)$.*

A problem in NP can be transformed into a coNP problem by negating the formulation of the question ("$\forall x : F(x) = \text{FALSE}$?"). However, there are good reasons to think of them as two different classes. For instance, a problem can be in both classes NP $\cap$ coNP meaning that both the YES and NO instances can be efficiently verified. The relation here is

$$\mathsf{P} \subseteq \mathsf{NP} \cap \mathsf{coNP} \subseteq \mathsf{NP}\,. \tag{2.8}$$

A relevant example of a problem in NP $\cap$ coNP is the decision version of prime factorization which asks whether an integer N has a prime factor less than some threshold q. This is the problem which can be solved efficiently on a quantum computer using Shor's algorithm. A proof is a list of all prime factors $p = (p_1, \ldots, p_l)$ with $N = p_1 \cdot p_2 \cdots p_l$. In polynomial time, one can verify that all p_i are indeed prime and that this describes the correct decomposition. The proof allows us to determine the smallest prime factor exactly and thus to verify both the YES and NO instances. This means that prime factorization is in NP $\cap$ coNP.

2.3 Reductions and completeness

In this section we introduce reductions and the notions of completeness and hardness. The idea of a reduction is to use a solver for one problem to solve another problem. We say that problem A is reducible to B ($A \leq_R B$), if one can solve the problem A by having access to a solver of B. Here R describes the type of reduction, i.e. the general rules of how one is allowed to reduce A to B. There are several types of reductions. A natural one is a Turing reduction where problem A is solved with an algorithm that is allowed to make multiple calls to a solver of B. A polynomial-time Turing reduction, also called a Cook reduction [27], requires that both the runtime of the algorithm and the number of calls to the solver are polynomially bounded. While

this captures our understanding of complexity, since it means that problem A can be solved efficiently if problem B can be solved efficiently, it has the consequence that problems in NP can be reduced to problems in coNP by negating the problem and returning the opposite result. To still distinguish between these classes, it is useful to use a more restrictive form of reductions known as a *many-one* reductions and their polynomial time versions known as a Karp reductions [28].

Definition 7 (Karp Reduction). *A decision problem A reduces to B under Karp reductions ($A \leq_K B$), if an instance in A can be solved by an algorithm that performs polynomially many operations followed by a call to a solver for problem B and returning the result from the solver without modification.*

The main differences from a Turing reduction are that only one call to the subroutine is allowed and that this result is also what the algorithm returns. The latter point prevents the reduction of an NP problem to its complement in coNP. The idea is that not only the complexity of the problems is similar, but also the structure itself, i.e. an instance of A can be mapped as an instance of B. Many-one reductions are therefore also refered to as *mapping* reductions. Based on Karp reductions, we can define a concept of completeness.

Definition 8 (Completeness). *A decision problem $X \in \mathsf{C}$ in some complexity class C is C-complete, if all other problems in C can be reduced to X by Karp reductions.*

$$\forall Y \in \mathsf{C} : Y \leq_K X \tag{2.9}$$

Completeness encapsulates that X is among the hardest problem in C, i.e. an efficient solver for problem X can be used to solve all problems in C efficiently. Karp reductions are preferred in the literature, but in principle Cook reductions can also be used [29]. Not all complexity classes have complete problems, but all we have introduced so far do indeed have them. This also motivates why we believe prime factoring is not NP-complete, since its inclusion in coNP would imply that NP = coNP. If we remove the requirement that the problem is in the complexity class, we get something called hardness.

Definition 9 (Hardness). *A decision problem X is C-hard, if all problem in C can be reduced to X by Karp reductions.*

$$\forall Y \in \mathsf{C} : Y \leq_K X \tag{2.10}$$

As such an EXP-complete problem is also NP-hard, but unless NP = EXP, not NP-complete. Colloquially, the notion of hardness is often also extended to non-decision problems, like optimization problems. This is technically inaccurate since a Karp reduction requires X to be a decision problem. In general, this means either that the associated decision version is NP-hard or that there is a single call Turing reduction from some NP-complete problem to the optimization problem. However, this second definition implies that a coNP-complete problem is also NP-hard. To avoid these complications, it is best to stick to decision problems and thus the above definition.

2.4 Probabilistic complexity classes

So far we have only looked at deterministic Turing machines and algorithms. However, there are many practical algorithms, notably Monte Carlo algorithms, that use randomness and sampling in their computation. This will also be important for quantum computers, since quantum measurements are intrinsically random. If we want to encapsulate a complexity class that also describes efficient probabilistic algorithms, we need to change two aspects. First, we need to define a probabilistic Turing machine. To do this, we need to add an operation that can flip a truly random coin and write the result to memory. Second, we need to allow for a small probability of failure. This is necessary because if the algorithm would work regardless of the outcome of the coin toss, you could simply replace it with a coin that always writes 0 into memory, avoiding the need for randomness all together.

We can define the class of problems that can be solved in polynomial time on a probabilistic Turing machine, which is called bounded probabilistic polynomial-time or BPP. For classical computers, this describes the largest class of decision problems that are considered to be efficiently solvable.

Definition 10 (Complexity class BPP). *A decision problem is in* BPP*, if there exists a probabilistic Turing machine (PTM), such that for all inputs $I \in S$ of size N, TM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a YES instance ($I \in S_y$), PTM accepts with probability at least $P[\text{accepts}] \geq \frac{2}{3}$.*
- *a NO instance ($I \in S_n$), PTM accepts with probability at most $P[\text{accepts}] \leq \frac{1}{3}$.*

Here the probabilities are arbitrary. They only need to be a pair of constants, independent of the input size (N) that are strictly greater and smaller than $1/2$. This is because running the algorithm multiple times and choosing the most common result (majority vote) allows the probability of success to be exponentially close to 1.

It is not known if BPP and P are actually two distinct classes. The question is related to whether there exist good pseudorandom number generators [30]. Mainly given the success of derandomizing many probabilistic algorithms, most famously prime testing [31], many researchers believe that the classes are indeed equivalent. Currently, polynomial identity testing [32] is a remaining BPP problem that is not known to be in P. We can define a larger class by allowing the probabilities of success and failure to be arbitrarily close to $\frac{1}{2}$, called probabilistic polynomial time or PP.

Definition 11 (Complexity class PP). *A decision problem is in* PP, *if there exists a probabilistic Turing machine (PTM), such that for all inputs $I \in S$ of size N, TM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a YES instance ($I \in S_y$), PTM accepts with probability at least $P[\text{accepts}] \geq \frac{1}{2}$.*
- *a NO instance ($I \in S_n$), PTM accepts with probability at most $P[\text{accepts}] \leq \frac{1}{2}$.*

The reason this class is assumed to be significantly larger is that the success probability can be arbitrarily close to $1/2$ and also depend on input size. We can see this by showing that NP $\subset$ PP. If we assume a problem $X \in$ NP, we can solve it in the following way:

1. Pick a random bit string $p \in \{0,1\}^M$.
2. Check if p is accepted by the NP verifier.
3. If p is a valid proof, accept the instance, $P[\text{accepts}] = 1$.
4. If p is not a valid proof, accept with probability $P[\text{accepts}] = \frac{1}{2} - \epsilon_M$.

Effectively, if p is a valid proof, we can be sure that it is a YES instance. If it is not, we still do not know the answer, but it is ever so slightly more likely to be a NO instance. If we set $\epsilon_M = \frac{1}{2^{M+1}}$, even if there is only one valid proof we get the probabilities

$$\begin{aligned}
\text{YES: } P[\text{accepts}] &= P[p \text{ is accepted}] \times 1 + P[p \text{ is rejected}] \left(\frac{1}{2} - \epsilon_M\right) \\
&\geq \frac{1}{2^M} + \left(1 - \frac{1}{2^M}\right)\left(\frac{1}{2} - \frac{1}{2^{M+1}}\right) = \frac{1}{2} + \frac{1}{2^{2M+1}} > \frac{1}{2} \\
\text{NO: } P[\text{accepts}] &= \frac{1}{2} - \frac{1}{2^{M+1}} < \frac{1}{2}.
\end{aligned}$$

This algorithm is very impractical, since it takes exponentially many rounds to achieve any kind of certainty about the result.

It also holds that PP is included in PSPACE, since one can simulate all possible random outcomes with polynomial memory and accept if the average acceptance rate is greater than $1/2$. This gives final relationship

$$\mathsf{NP} \subset \mathsf{PP} \subset \mathsf{PSPACE}. \tag{2.11}$$

The last probabilistic class we introduce here is a natural generalization of NP where the verifier has access to a probabilistic Turing machine. This class is called Merlin-Arthur (MA).

Definition 12 (Complexity class MA). *A decision problem is in* MA*, if there exists a probabilistic Turing machine (PTM), such that for all inputs $I \in S$ of size N, PTM terminates in polynomial time ($O(\mathsf{poly}(N))$) and if I describes*

- *a YES instance ($I \in S_y$), then there exists a proof $p \in \{0,1\}^{M=O(\mathsf{poly}(N))}$ such that PTM accepts on input $I = (I, p)$ with probability at least $P[\text{accepts}] \geq \frac{2}{3}$.*
- *a NO instance ($I \in S_n$), for all $p \in \{0,1\}^M$, PTM accepts on input $I' = (I, p)$ with probability at most $P[\text{accepts}] \leq \frac{1}{3}$.*

It follows from this structure that P = BPP also implies NP = MA. The inclusions here are

$$\mathsf{NP}, \mathsf{BPP} \subset \mathsf{MA} \subset \mathsf{PP}. \tag{2.12}$$

The last inclusion follows with a similar algorithm as we saw for NP ⊂ PP.

2.5 Quantum complexity classes

Below we introduce the complexity classes that arise in quantum computing. To do this we need to define a quantum Turing machine (QTM). For this we can take a traditional Turing machine and replace the classical tape with a quantum state, the classical gate-set with a quantum gate-set and introduce the concept of a quantum measurement to decide acceptance/rejection. A quantum state is represented in a bra-ket notation, which describes column and row vectors respectively. A qubit can

be described by a two dimensional complex vector. For a state $|a\rangle \in \mathcal{H}_2 = \mathbb{C}^2$ we have

$$|a\rangle = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \qquad |0\rangle \coloneqq \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad |1\rangle \coloneqq \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{2.13}$$

$$\langle a| = \begin{pmatrix} a_0^* & a_1^* \end{pmatrix} \qquad \langle 0| \coloneqq \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad \langle 1| \coloneqq \begin{pmatrix} 0 & 1 \end{pmatrix} \tag{2.14}$$

where $(\cdot)^*$ refers to the complex conjugation and we also introduce the basis states $\{|0\rangle, |1\rangle\}$. To create multi-qubit systems we use the tensor product

$$|a, b\rangle = |a\rangle \otimes |b\rangle = \begin{pmatrix} a_0 b_0 \\ a_0 b_1 \\ a_1 b_0 \\ a_1 b_1 \end{pmatrix} , \tag{2.15}$$

which makes the Hilbert space dimension of a quantum system scale exponentially with the number of qubits. An n qubit state $|\boldsymbol{a}\rangle \in \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_2 = \mathbb{C}^{2^n} = \mathbb{C}^d$ can be represented as

$$|\boldsymbol{a}\rangle = \sum_{i \in \{0,1\}^n} a_{i_1 \cdots i_n} |i_1 \dots i_n\rangle , \tag{2.16}$$

where $a_{i_1 \cdots i_n} \in \mathbb{C}$ are some complex coefficients. Quantum states are also normalized, meaning that

$$\sum_{i \in \{0,1\}^n} |a_{i_1 \cdots i_n}|^2 = 1 \tag{2.17}$$

holds. Quantum gates are described by unitary operations $U \in U(d) \subset C^{d,d}$ which are linear maps with the additional condition $UU^\dagger = \mathbb{1}$, where $(\cdot)^\dagger$ is the Hermitian conjugate. The gate is then the map $|a\rangle \mapsto U|a\rangle$.
For physical systems, the time evolution is described by the Schrödinger equation

$$\mathrm{i}\partial_t |\Psi(t)\rangle = H |\Psi(t)\rangle , \tag{2.18}$$

where H refers to the Hamiltonian or energy operator of the system, which is a Hermitian operator $H \in \mathbb{C}^{d,d}$ with $H^\dagger = H$. Throughout, we set the Planck constant $\hbar = 1$. The unitary time evolution is then given by

$$|\Psi(t)\rangle = \mathrm{e}^{-\mathrm{i}Ht} |\Psi(0)\rangle = U(t) |\Psi(0)\rangle . \tag{2.19}$$

For this reason we also call H the generator of $U(t)$. Permutations, a discrete subclass of the unitary groups are called *classical* gates. This is because they map basis states

to basis states. As an example, we have the X gate, which describes a NOT operation or a bit flip

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad X|1\rangle = |0\rangle, \quad X|0\rangle = |1\rangle \tag{2.20}$$

Similarly two qubits gates can be defined, for example the controlled NOT operation (CNOT), which flips the second bit, but only if the first bit is in the state 1

$$\mathrm{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathrm{CNOT}|i,j\rangle = |i, j \oplus i\rangle \ . \tag{2.21}$$

Here $\oplus$ refers to addition modulo 2. For a universal quantum gate set, we also need some non-classical gates. For example, the Hadamard gate (H) and the T gate are both quantum gates without a classical counterpart.

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad H|0\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}} =: |+\rangle \ , \qquad H|1\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}} =: |-\rangle$$

$$T = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{e}^{\mathrm{i}\frac{\pi}{4}} \end{pmatrix}, \qquad T|0\rangle = |0\rangle \ , \qquad T|1\rangle = \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}|1\rangle$$

If a gate acts on a specific site in a larger Hilbert space we use the tensor product. For example the matrix representation of a T gate at position k in a n-qubit system is

$$T_k := \mathbb{1}_{2^{k-1}} \otimes T \otimes \mathbb{1}_{2^{n-k}} \ . \tag{2.22}$$

It can be shown that the set of gates T, H and CNOT acting on arbitrary qubits are universal [33]. Since these gates describe a countable set of operations, but the total number of unitaries is uncountable, we have to use a weaker version of universality. Universal in the context of quantum computing means that for any $U \in U(2^n)$ and any precision $\epsilon > 0$, one can find a sequence of gates from the gate-set that implements the gate V, where $\|U - V\| \leq \epsilon$, i.e. the gate-error can be made arbitrarily small. Here $\|\cdot\|$ refers to some operator norm.

Finally, we also define the measurement operation $\mathcal{M}_k$, which measures the k-th qubit and returns the measured value. The absolute square of the amplitude $|a_{i_1 \cdots i_n}|^2 = |\langle i_1 \cdots i_m | a \rangle|^2$ describes the probability of the post measurement state

being $|i_1 \dots i_n\rangle$ after a full Z basis measurement is performed. As such the likelihood of measuring 1 on site k is

$$P[M_k = 1] = \langle a | M_k | a \rangle = \sum_{i \in \{0,1\}^n | i_k = 1} |a_{i_1 \cdots i_n}|^2 , \tag{2.23}$$

where $M_k = |1\rangle\langle 1|_k$ describes the measurement operator $M = |1\rangle\langle 1|$ acting on site k.

With this background we are able to define a quantum Turing machine. The key components are

- The tape is replaced by a quantum state. A classical state is associated with the corresponding basis state. The internal state of the Turing machine also includes a quantum register.
- Logical operations are extended to include a universal quantum gate-set.
- For the final return, the Turing machine can measure a qubit in its internal state and return the classical result.

Similar to classical Turing machines, there are different ways to define the exact architecture. Popular alternative definitions include having both a quantum and a classical memory or allowing multiple measurements during execution. However, it can be shown that this does not affect the overall performance of such a machine (within polynomial time overhead) because the quantum tape is able to simulate both classical computation and intermediate measurements. With this in mind, we can define the quantum analog of BPP called bounded quantum polynomial-time or BQP.

Definition 13 (Complexity class BQP). *A decision problem is in* BQP, *if there exists a quantum Turing machine (QTM), such that for all inputs $I \in S$ of size N, QTM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a YES instance ($I \in S_y$), QT accepts with probability at least $P[\text{accepts}] \geq \frac{2}{3}$*
- *a NO instance ($I \in S_n$), QT accepts with probability at most $P[\text{accepts}] \leq \frac{1}{3}$.*

This class describes the largest class of problems which can be solved in polynomial time, at least according to our current understanding of physics. It is not trivial to show that a (limited-memory) quantum Turing machine can be practically implemented. The reason is that any errors occurring during the execution of quantum

gates in real implementations could accumulate and lead to a wrong result. The threshold theorem [5] shows that error correction is indeed possible, and thus also long running quantum algorithms.

We can also define quantum generalizations of the class MA. First, the class QCMA (quantum-classical Merlin-Arthur) describes problems that can be verified on a quantum computer with a classical proof.

Definition 14 (Complexity class QCMA). *A decision problem is in* QCMA*, if there exists a quantum Turing machine (QTM), such that for all inputs $I \in S$ of size N, QTM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a YES instance ($I \in S_y$), there exists a proof $p \in \{0,1\}^{M=O(\text{poly}(N))}$ such that QMT accepts on input $I' = (I, p)$ with probability at least $P[\text{accepts}] \geq \frac{2}{3}$.*
- *a NO instance ($I \in S_n$), for all $p \in \{0,1\}^M$, QTM accepts on input $I' = (I, p)$ with probability at most $P[\text{accepts}] \leq \frac{1}{3}$.*

This is the complexity class that best encapsulates VQAs, the main focus of this thesis. Another, more popular generalization is the class QMA (Quantum Merlin-Arthur). This describes a setup where the proof is given as a quantum state $|p\rangle$.

Definition 15 (Complexity class QMA). *A decision problem is in* QMA*, if there exists a quantum Turing machine (QTM), such that for all inputs $I \in S$ of size N, QTM terminates in polynomial time ($O(\text{poly}(N))$) and if I describes*

- *a YES instance ($I \in S_y$), there exists a proof state $|p\rangle \in \mathbb{C}^{2^{M=O(\text{poly}(N))}}$ such that QT accepts on input $I' = (I, |p\rangle)$ with probability at least $P[\text{accepts}] \geq \frac{2}{3}$.*
- *a NO instance ($I \in S_n$), for all states $|p\rangle \in \mathbb{C}^{2^M}$, QT accepts on input $I' = (I, |p\rangle)$ with probability at most $P[\text{accepts}] \leq \frac{1}{3}$.*

The class is particularly relevant because the ground state energy problem (up to polynomial precision) is QMA-complete [34]. We know that both classes contain BQP and MA and it has also been shown that they are included in the class PP [35]

$$\text{MA}, \text{BQP} \subseteq \text{QCMA} \subset \text{QMA} \subseteq \text{PP}. \tag{2.24}$$

Much is still unknown about the relationship between the two classes, QCMA and QMA. For them to be meaningfully different, there has to be problem instances,

for which a proof state $|p\rangle$ cannot be efficiently prepared on a quantum computer, i.e. a superpolynomial number of gates is required to implement the proof state. Otherwise, instead of $|p\rangle$, the proof could be a classical description of how to prepare $|p\rangle$ instead. This has direct implications for the ground state problem. Since the ground state can act as a proof state, QCMA $\neq$ QMA implies that there are ground states that cannot be prepared in polynomial time. Similar to the difficulties when proving circuit lower bounds for classical problems, so far only linear circuit lower bounds [36] are known.

3 Variational quantum algorithms

Variational Quantum Algorithms (VQAs) [37–39], also referred to as Variational Quantum Eigensolver (VQE), have gained prominence in recent years as candidate algorithms for achieving useful quantum advantages on NISQ devices. They are used to estimate the ground state energy of a quantum many-body Hamiltonian via variational optimization. Unlike alternatives like the Quantum Phase Estimation (QPE) algorithm [40], which requires very accurate quantum circuits, VQAs can be run on near-term hardware. VQAs are a hybrid quantum-classical algorithm, where the quantum hardware of VQAs prepares a variational quantum state by applying a parameterized quantum circuit. To interface with a classical computer, the expectation value of the parameterized state is estimated. Then a classical computer, typically running a gradient descent based algorithm [41, 42] or Nelder-Mead [43], is used to minimize the energy via repeated parameter updates and new estimation of the energy functional. Detailed reviews of state of the art methods are given in [44, 45]. VQAs can, in principle, be run on any circuit design and at any depth, allowing for a hardware-tailored ansatz. It is hoped that the additional classical optimization can be used to improve the performance of the still limited capabilities of existing quantum hardware. In the following section we will define VQAs more rigorously as well as introduce a popular version known as the Quantum Alternating Operator Ansatz (QAOA) [46, 47], which is inspired by the adiabatic theorem.

3.1 Definition of VQAs

For each VQA instance we have three components which we will discuss.

- The initial state $|\Psi_0\rangle \in \mathbb{C}^{2^n}$
- The parameterized circuit $U(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in D \subset \mathbb{R}^L$
- The cost observable $O \in \mathbb{C}^{2^n,2^n}$, with $O = O^\dagger$

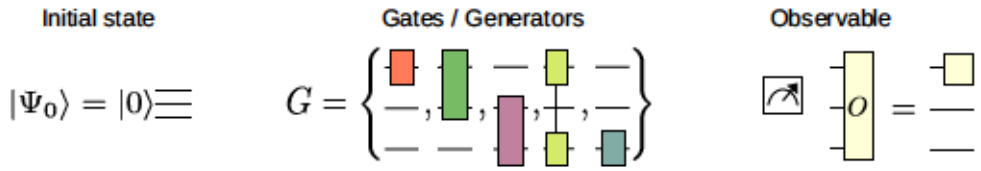

Fig. 3.1. – A graphical depiction of a VQA instance. Here the initial state is chosen as the all zero state. The gate generators are depicted by individual colors.

A sketch of a VQA instance with the relevant components is shown in fig. 3.1. The overall objective of the algorithm is to minimize the expectation value of the parameterized state $|\Psi(\boldsymbol{\theta})\rangle \coloneqq U(\boldsymbol{\theta})\,|\Psi_0\rangle$. The cost function is then given by

$$\langle O(\boldsymbol{\theta})\rangle \coloneqq \langle \Psi_0|U^\dagger(\boldsymbol{\theta})OU(\boldsymbol{\theta})|\Psi_0\rangle = \langle \Psi(\boldsymbol{\theta})|O|\Psi(\boldsymbol{\theta})\rangle \tag{3.1}$$

$$\langle O_{\min}\rangle \coloneqq \min_{\boldsymbol{\theta}\in D} \langle O(\boldsymbol{\theta})\rangle \ , \tag{3.2}$$

which is minimized with a classical algorithm.

3.1.1 Intital state

The initial state $|\Psi_0\rangle$ must be simple. For our complexity analysis this means that it is easy to implement in polynomial time on a quantum computer. We assume that the state is given by a circuit description of $V_{\text{prep}} = \prod_{i=1}^{D} V_{\text{prep}}^i$ where V_{prep}^i are some local unitaries. So we can write

$$|\Psi_0\rangle = V_{\text{prep}}\,|0\rangle \ . \tag{3.3}$$

For real life implementations, we could replace $|0\rangle$ with the actual initial state of the physically existing hardware. In addition, it is necessary to ensure that V_{prep} can actually be implemented and that the overall circuit complexity is small enough to keep noise to a minimum. Therefore it may be useful to set the initial state directly as the all zero state ($|\Psi_0\rangle = |0\rangle$) or some product state ($|\Psi_0\rangle = |s\rangle$, $s \in \{0,1\}^n$).

3.1.2 Circuit design

The circuit $U(\boldsymbol{\theta})$ is the key component in an VQA. Ideally, the circuit should

- be easy to implement with a small circuit depth. On hardware this means that $U(\boldsymbol{\theta})$ should adhere to the hardware's connectivity constraints as well as use its natural gate-set.

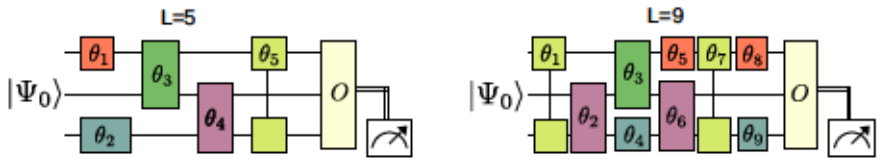

Fig. 3.2. – Two possible VQA settings in a variable VQA ansatz. The values $\{\theta_1, \ldots, \theta_L\}$ describe the respective gate times, while the colour represents the type of generator used, which are optimized as additional hyper parameters. In this setting, the same generators can be reused multiple times. Thus the circuit depth L also becomes an optimization parameter.

- be able to reach expectation values close to the ground state energy of O for a good variational state. When we use $\lambda_{\min}(O)$ to refer to the ground state energy, this means $\langle O_{\min} \rangle \approx \lambda_{\min}(O)$.

- not be overparameterized. While we want the circuit to be able to reach low energy states of O, we do not want an excessively large parameter space (L) since this makes the optimization procedures more expensive and harder to implement on NISQ hardware.

In general, there are many ways to design the circuit. Which type is preferred may depend on the hardware used as well as the structure of the problem at hand. The parameters $\theta_i \in \mathbb{R}$ are used to represent some parameterization of the unitary. Commonly one can assume that gates are generated by some time evolution with a tunable Hamiltonian

$$U(\boldsymbol{\theta}) = e^{\mathrm{i}\left(\sum_{i=1}^{L} \theta_i H_i + H_0\right)} . \tag{3.4}$$

Here θ_i can describe interaction strengths or the size of some external fields of a physically existing Hamiltonian. Common proposals use a single parameter gate for the circuit model, meaning a sequence of gates where each parameter θ_i describes some effective evolution time of one generator H_i

$$U(\boldsymbol{\theta}) = e^{\mathrm{i}H_L\theta_L} \ldots e^{\mathrm{i}H_2\theta_2} e^{\mathrm{i}H_1\theta_1} = \prod_{i=1}^{L} e^{\mathrm{i}H_i\theta_i} . \tag{3.5}$$

This is a natural structure which is compatible with the quantum circuit model, if each H_i describes a local gate generator.
The order of the individual H_i can either be fixed by the ansatz class, or be chosen as part of the optimization procedure similar to how hyper parameters are chosen in machine learning. This more variable approach is used for instance in adapt-VQA [48]. An example instance of this is shown in fig. 3.2.

3.1.3 Observables

Finally, the observable is usually constrained by the particular problem of interest. It is often convenient to represent an observable in the Pauli basis. Here, we have the four Pauli matrices

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{1}\,, \qquad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sigma_x\,, \tag{3.6}$$

$$\sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \sigma_y\,, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \sigma_z\,, \tag{3.7}$$

where the matrix representation is given in the Z or computational basis. Often the idenitity (σ_0) is not considered a Pauli matrix, but for our purposes it is beneficial to include it. Pauli matrices have the interesting property of being both Hermitian and unitary. They are also a basis for 2×2 matrices for any operator O

$$O = a_0\sigma_0 + a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3\,, \tag{3.8}$$

where the constraint $a_i \in \mathbb{R}$ is equivalent to O being Hermitian. For multi-qubit systems, we can consider tensor products of Pauli matrices, so-called Pauli-strings. They are given as

$$P_{\boldsymbol{k}} = \sigma_{k_1} \otimes \sigma_{k_2} \otimes \cdots \otimes \sigma_{k_n}\,, \tag{3.9}$$

with $\boldsymbol{k} \in \{0,1,2,3\}^n$. In total there are 4^n different Pauli-strings. We also use σ_α^i to refers to a Pauli matrix σ_α acting at position i.
Pauli-strings form a basis for the operators in the 2^N dimensional Hilbert space, i.e. every observable can be represented as a linear combination of Pauli-strings

$$O = \sum_{\boldsymbol{k}\in\{0,1,2,3\}^n} a_{\boldsymbol{k}} P_{\boldsymbol{k}}\,, \tag{3.10}$$

where $\boldsymbol{a} \in \mathbb{C}^{4^n}$ for arbitrary operators, and $\boldsymbol{a} \in \mathbb{R}^{4^n}$ for Hermitian operators. The locality or weight $\mathrm{w}(P_{\boldsymbol{k}}) = l$ refers to the number of non-identity Pauli matrices, i.e. the Pauli-string acts non-trivially on l sites. Importantly, there are only $4^l\binom{n}{l} = O(n^l)$ different l-local Pauli-strings. The definition of an l-local observable is that it can be written as a linear combination of l-local Pauli-strings

$$O = \sum_{\substack{\boldsymbol{k}\in\{0,1,2,3\}^n \\ \mathrm{w}(P_{\boldsymbol{k}})\leq l}} a_{\boldsymbol{k}} P_{\boldsymbol{k}}\,. \tag{3.11}$$

If the observable acts trivially on at least $n - l$ sites we say that the observable is *strictly* l-local. Constant l-local observables are often used when considering

quantum computational problems since they can be describes with only polynomially many parameters in the number of qubits.

There are two categories of observables that are proposed for VQAs.

Classical observables: The observable is chosen to represent an NP problem. The general approach is to use a diagonal observable in the Z basis, also called an Ising model

$$O = \sum_{i=1}^{n} a_i \sigma_z^i + \sum_{i,j=1}^{n} b_{ij} \sigma_z^i \sigma_z^j \,, \tag{3.12}$$

where $a \in \mathbb{R}^n, b \in \mathbb{R}^{n,n}$ defines the instance. It is also possible to add higher locality σ_z terms if desired. This approach is feasible because finding the ground-state of such an Ising model is NP-hard and its corresponding decision problem is NP-complete. Thus an NP problem can be reduced to an Ising model, i.e. it can expressed by an observable of the form in eq. (3.12). For example, the MaxCut problem can be represented with the cost observable

$$O = \frac{1}{2} \sum_{i,j=1, i<j}^{n} A_{ij} \left(\mathbb{1} - \sigma_z^i \sigma_z^j \right) \,, \tag{3.13}$$

where A_{ij} is the adjacency matrix of the graph.

Quantum observables: This category generally involves physically inspired Hamiltonians that come from diverse fields such as solid-state, particle, atomic or nuclear physics [49] as well as molecular chemistry [50] or materials science [51]. Ground state energies are important for determining chemical reaction energies, the stability of molecules or atoms and much more. Here, classical techniques such as Hartree-Fock and Density Functional Theory (DFT) [52] are powerful tools that have been used for many numerical studies. A quantum advantage in this context would therefore be a VQA implementation that can significantly outperform the best classically derived energy estimates for a given ground state energy problem. A local observable can be represented as a sum of strictly local operators

$$O = \sum_{i=1}^{n_O} O_i \,, \tag{3.14}$$

where the O_i act non-trivially on at most l qubits. For practical purposes, the locality constraint can be an obstacle. This is because most physical Hamiltonians involve

fermions, namely electrons. The orbital basis set discretization is often used to obtain a discrete, local fermionic observable

$$O = \sum_{i,j,k,l} O_{ijkl} c_i^\dagger c_j^\dagger c_k c_l \,, \tag{3.15}$$

where $c_i^\dagger, c_i$ are the fermionic creation and annihilation operators and $O_{ijkl} \in \mathbb{C}$ some coefficients resulting from the discretization of the Hamiltonian. Fermionic systems obey the commuatiton relations

$$\left\{c_i, c_j^\dagger\right\} = \delta_{ij} \,, \quad \{c_i, c_j\} = 0 \,, \tag{3.16}$$

where $\{A, B\} = AB + BC$ is the anti-commutator. This makes it difficult to find a good Pauli basis representation of c_i, which observes these relations. The most common approach is the Jordan-Wigner transformation [53] which uses the definition

$$c_i = (-\sigma_z) \otimes \cdots \otimes (-\sigma_z) \otimes \sigma_- \otimes \mathbb{1} \otimes \cdots \otimes \mathbb{1} \,, \tag{3.17}$$

where σ_- acts on the i-th site and

$$\sigma_- = \frac{\sigma_x - \mathrm{i}\sigma_y}{2} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} . \tag{3.18}$$

It can be verified that this definition recovers the desired commutation relations. This means that a local fermionic observable O (eq. (3.15)) can be efficiently represented in the Pauli basis, but are not necessarily be describable as a local observable. Finding a good qubit representation for fermionic Hamiltonians is still a very active research area [54, 55].

3.2 Measurements and measurement effort

A quantum device needs to have a measurement protocol to probe the prepared quantum state. A good model for many architectures is that of the Z basis measurement. This assumes that the device can perform a complete measurement in the computational basis to obtain a classical string $\boldsymbol{s} \in \{0, 1\}^n$ corresponding to the measurement outcome. If the architecture is not capable performing a full basis measurement, as may be the case for NV-centers [56] and some quantum dot implementations, the protocols described below must be modified.

For classical observables, compuational basis measurements suffice to estimate the observable. For quantum observables, which have an eigenbasis different from the compuational basis, additional steps must to be taken. In the following we describe

measurement strategies for estimating the expectation value $\langle O \rangle = \langle \Psi | O | \Psi \rangle$ from basis measurements. Finding good strategies is very important because reducing the number of required measurement setting and the total number of measurement rounds, can greatly benefit the overall performance of a VQA, especially when the state preparation is resource intensive. In general, the strategy is to decompose the observable into terms $\tilde{O}_j$

$$O = \sum_{j=1}^{n_m} \tilde{O}_j \tag{3.19}$$

that can be efficiently measured. For this, each $\tilde{O}_j$ must be diagonalizable

$$\tilde{O}_j = U_j^\dagger D_j U_j \,, \tag{3.20}$$

where U_j describes a unitary operator and D_j is a diagonal operator. We require that each entry of D_j can be efficiently computed classically and that the unitary U_j is implementable on the physical device.
The problem of finding the best decomposition is generally very difficult. However, simply finding some working measurement strategy is straightforward. One option is to choose $\tilde{O}_j = O_i$ since a local observable can be efficiently diagonalized. Similarly, if $[O_i, O_k] = 0$ holds for some $i, k \in [n_O]$, the observable $\tilde{O}_j = O_i + O_k$ can also be used, since the observables can be measured simultaneously.
Due to hardware limitations, it might also be required to perform multiple measurement for a single local term leading to $n_m > n_O$. This might be the case if the device is only able to perform local Pauli operations. There are also alternative strategies such as classical shadows [57], which can be useful when estimating multiple observables at once. Thus, the measurement strategy by the following algorithm.

- For $j \in [n_m]$:
    1. Prepare the state to be measured: $|\Psi\rangle$
    2. Apply the unitary U_j: $U_j |\Psi\rangle$
    3. Perform a Z basis measurement. Get the measured string $s \in \{0,1\}^n$.
    4. Calculate the entry $(D_j)_s$.
    5. Repeat steps (1-4) m_j times, setting o_j as the average value of all $(D_j)_s$.
- Return the estimate $o = \sum_{j=1}^{n_m} o_j$.

This protocol uses $m = \sum_j m_j$ total measurement rounds and gives the correct expectation value on average.

$$\langle o \rangle = \sum_{j=1}^{n_m} \sum_{s \in \{0,1\}^n} (D_j)_s p_{j,s} \tag{3.21}$$

$$= \sum_{j=1}^{n_m} \sum_{s \in \{0,1\}^n} (D_j)_s \langle \Psi | U_j^\dagger |s\rangle\langle s| U_j | \Psi \rangle \tag{3.22}$$

$$= \sum_{j=1}^{n_m} \langle \Psi | \tilde{O}_j | \Psi \rangle \tag{3.23}$$

$$= \langle \Psi | O | \Psi \rangle \ , \tag{3.24}$$

where $p_{j,s}$ describes the probability of measuring s when the measurement setting is chosen for the $\tilde{O}_j$ observable. This assumes that both the gate operations and the basis measurements are performed without error, i.e. there is no systematic error. To estimate the statistical error, we first calculate the variance of a single measurement (single shot) of $\tilde{O}_j$.

$$\sigma^2_{\tilde{O}_j} = \langle o_j^2 \rangle - \langle o_j \rangle^2 \tag{3.25}$$

$$= \sum_{j,s} D^2_{j,s} p_{j,s} - \left(\sum_{j,s} D_{j,s} p_{j,s} \right)^2 \tag{3.26}$$

$$= \langle \Psi | \tilde{O}_j^2 | \Psi \rangle - \langle \Psi | \tilde{O}_j | \Psi \rangle^2 \tag{3.27}$$

$$\tag{3.28}$$

The variance is in general state dependent. A useful first estimate is to use the maximally mixed state

$$|\Psi\rangle\langle\Psi| \to \frac{\mathbb{1}}{2^n} , \tag{3.29}$$

for which the estimate is

$$\sigma^2_{\tilde{O}_j} = \mathrm{Tr}[\tilde{O}_j^2]/2^n - \mathrm{Tr}[\tilde{O}_j]^2/2^{2n} \tag{3.30}$$

This is also the result when $|\Psi\rangle$ is a random state.according to the Haar measure. The Haar measure describes a distribution that is invariant under any unitary application. It is a convenient first estimate when $|\Psi\rangle$ does not have a sparse representation in the eigenbasis of $\tilde{O}_j$. In contrast, if the state is in an eigenstate of $\tilde{O}_j$, the variance vanishes for this measurement setting. It is important to note that even if the state is the ground state of O, the individual measurements of each $\tilde{O}_j$ can still have large variances, meaning that the final estimate will as well.
If we choose $\tilde{O}_j$ such that $\mathrm{Tr}[\tilde{O}_j] = 0$, the estimate for the variance w.r.t. the

maximally mixed state is $\sigma_{o_j}^2 = \mathrm{Tr}[\tilde{O}_j^2]/2^n$ or the mean squared entry of D_j. For a Pauli operator, this gives $\sigma_P^2 = 1$ or for a sum of Pauli operators

$$\tilde{O}_j = \sum_i^{n_P} \alpha_i P_i \tag{3.31}$$

the variance is

$$\sigma_{\tilde{O}_j}^2 = \sum_i |\alpha_i|^2 \,. \tag{3.32}$$

However any particular state $|\Psi\rangle$ will have deviations from this estimate. To obtain estimation guarantees, Hoeffding's inequalities can be used [58]

Repeated measurements can be described by a multinomial distribution, meaning that the variance of the estimation after m_j rounds of measurement is given by

$$\sigma_{\tilde{O}_j,m_j}^2 = \frac{\sigma_{\tilde{O}_j}^2}{m_j} \,. \tag{3.33}$$

For the total variance for the estimation routine this gives

$$\sigma_m^2 = \sum_j \frac{\sigma_{\tilde{O}_j}^2}{m_j} \,. \tag{3.34}$$

If the total number of measurement rounds is fixed ($m = \sum_j m_j$), we can find the optimal measurement budget allocation using Lagrange multipliers. The Lagrange function is given by

$$\sigma_m^2 = \sum_j \frac{\sigma_{\tilde{O}_j}^2}{m_j} + \lambda \left(\sum_{j=1}^{n_m} m_j - m \right) \,. \tag{3.35}$$

Taking the derivative w. r.t. m_j yields the condition

$$0 = -\frac{\sigma_{\tilde{O}_j}^2}{m_j^2} + \lambda \,, \tag{3.36}$$

which together with the measurement budget constraint gives the optimal measurement budget allocation

$$m_j = m \frac{\sigma_{\tilde{O}_j}}{\sum_j \sigma_{\tilde{O}_j}} \,. \tag{3.37}$$

The final measurement variance is therefore given by

$$\sigma_m^2 = \frac{(\sum_j \sigma_{\tilde{O}_j})^2}{m} \,. \tag{3.38}$$

While the values of σ_{O_j} are a priori unknown, one can use the collected empirical estimates of the variances to implement a near-optimal measurement strategy, at least for a large measurement budget $m \gg n_m$. The total standard deviation scales proportionally to the sum of the individual single shot standard deviations and not to the variances. Therefore it is generally helpful to reduce the number of measurement settings n_m.

Example: We can consider an observable given by a sparse, representation in the Pauli basis $O = \sum_{i=1}^{n_O} P_i$, with Pauli operators P_i. If one can measure each Pauli operator simultaneously (as is the case for classical observables), then for a random state one can expect a variance of

$$\sigma_m^2 = \frac{n_O}{m}, \tag{3.39}$$

meaning the number of measurement rounds required scales proportional to the number of Pauli terms in the observable. If each component is measured in an individual measurement setting, as may be required if the Pauli operators do not commute, we expect instead

$$\sigma_m^2 = \frac{n_O^2}{m}, \tag{3.40}$$

i.e. the required measurement rounds scale quadratically with n_O. This shows that, especially for large n_O, efficient grouping can significantly reduce the measurement effort. If the observable has $n_O = 100$ local terms, and we want the estimate to be within chemical accuracy, which can be around $\sigma_m = 0.01$, this requires between $m = 1,000,000$ and $m = 100,000,000$ measurements depending on the method. Of cause for a particular state $|\psi\rangle$, the required rounds may be significantly reduced.

3.3 Quantum Alternating Operator Ansatz (QAOA)

One particular VQA construction that has attracted a great deal of interest is known as Quantum Alternating Operator Ansatz (QAOA). First proposed under the name Quantum Approximate Optimization Algorithm as a specific quantum algorithm to solve the MaxCut problem [46], it has been generalized to a broader class of algorithms [47] which are inspired by the adiabatic theorem. We will motivate the protocol by deriving the adiabatic theorem, which was first shown by Max Born in 1928 [59].

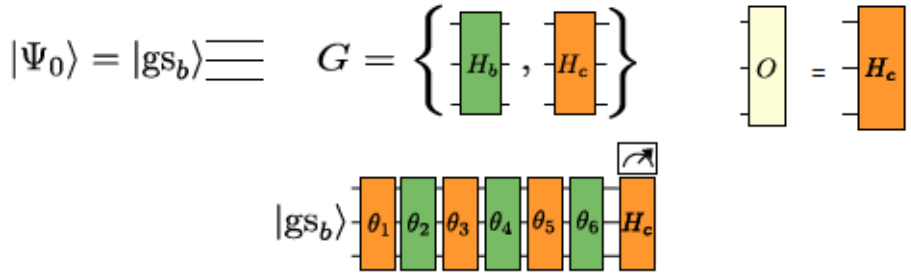

Fig. 3.3. – A graphical representation of the QAOA ansatz. The initial state is the ground state of H_b and the cost observable is H_c. The gate generators are H_b and H_c.

Theorem 1 (Adiabatic theorem). *For two Hamiltonians H_b and H_c, when one starts in the ground state of H_b, $|\Psi_0\rangle = |\mathrm{gs}_b\rangle$ and time evolves under a Hamiltonian defined by*

$$H_\tau = (1-\tau)H_b + \tau H_c\,, \tag{3.41}$$

with a smooth transition from $\tau(t=0) = 0$ to $\tau(t=T) = 1$. If there are no level crossings of the ground-state energy of H_τ, then the final state converges to the ground state of H_c, $|\Psi_f\rangle = |\mathrm{gs}_c\rangle$ for very long transition times $T \to \infty$ and a nearly vanishing change $\dot{\tau} \to 0$.

The intuitive reason here is that the large transition time ensures that any leakage into higher energy states as H_τ changes are averaging out.

Proof sketch: We can derive the theorem by using the Schrödinger equation for time dependent Hamiltonians

$$\mathrm{i}|\dot{\Psi(t)}\rangle = H(t)\,|\Psi(t)\rangle \tag{3.42}$$

$$E_i(t)\,|i(t)\rangle = H(t)\,|i(t)\rangle\,, \tag{3.43}$$

where the first expression describes the time evolution of the system, while the second expression defines the time dependent energy eigenbasis of $H(t)$. We can derive a differential equation for the change of the eigenstate by taking the time derivative of eq. (3.43)

$$0 = (\dot{E_i}(t) - \dot{H(t)})\,|i(t)\rangle + (E_i(t) - H(t))|\dot{i(t)}\rangle \tag{3.44}$$

$$= (\dot{E_i}(t) - \dot{H(t)})\,|i(t)\rangle + \sum_{j=0}^{d} |j(t)\rangle\langle j(t)|\,(E_i(t) - E_j(t))|\dot{i(t)}\rangle \tag{3.45}$$

We can use $\langle i(t)|\dot{i(t)}\rangle = 0$, which can be derived by taking the derivative of the normalization constraint $\langle i(t)|i(t)\rangle = 1$ and fixing the relative phase of the eigenvectors at different times. With the pseudo inverse, this gives the solution

$$|\dot{i(t)}\rangle = \sum_{j=0|j\neq i}^{d} |j(t)\rangle \frac{\langle j(t)|\dot{H(t)}|i(t)\rangle}{E_i(t) - E_j(t)} . \tag{3.46}$$

The amplitude of a system being in a particular eigenstate at time t is given as

$$c_i(t) \coloneqq \langle i(t)|\Psi(t)\rangle . \tag{3.47}$$

Using eq. (3.42) and eq. (3.46), we can derive the differential equation

$$\dot{c_i(t)} = \langle \dot{i(t)}|\Psi(t)\rangle + \langle i(t)|\dot{\Psi(t)}\rangle \tag{3.48}$$

$$= \sum_{j\neq i} \frac{\langle i(t)|\dot{H(t)}|j(t)\rangle}{E_i(t) - E_j(t)} \langle j(t)|\Psi\rangle - \mathrm{i} \langle i(t)| H(t) |\Psi\rangle \tag{3.49}$$

$$\mathrm{i}\dot{c_i(t)} = \dot{\tau(t)} \sum_{j\neq i} \mathrm{i} \frac{\langle i(t)|\Delta H|j(t)\rangle}{E_i(t) - E_j(t)} c_j(t) + E_i(t)c_i(t) . \tag{3.50}$$

Thus, the evolution in the eigenbasis of $H(t)$ is governed by a Schrödinger equation with a Hamiltonian

$$H_c(t) = H_0(t) + \dot{\tau(t)} V(t) \tag{3.51}$$

$$= \sum_{i=0}^{2^n-1} E_i(t) |i(t)\rangle\langle i(t)| + \mathrm{i}\dot{\tau(t)} \sum_{j\neq i} \frac{\langle i(t)|\Delta H|j(t)\rangle}{E_i(t) - E_j(t)} |i(t)\rangle\langle j(t)| , \tag{3.52}$$

where the rate of change $\dot{\tau}$ is assumed to be very small. Here, H_c is effectively a diagonal Hamiltonian with small off-diagonal terms. As such a large energy barrier $\Delta_t = E_1(t) - E_0(t)$ prevents any transition away from the ground state. This can be shown more rigorously using the Riemann-Lebesgue lemma. With the upper bound $|\langle i(t)|\Delta H|j(t)\rangle| \leq \|H_b - H_c\|_\infty$, we get the condition

$$\dot{\tau}(t) \ll \frac{\Delta_t^2}{\|H_b - H_c\|_\infty} \tag{3.53}$$

or when the transition rate is constant,

$$T \gg \frac{\|H_b - H_c\|_\infty}{\Delta^2} . \tag{3.54}$$

If the spectral gap is known, it is advisable to use a varying transition speed $\dot{\tau}(t) \propto \Delta_t^2$. However, even if the full spectra are known, finding an optimal transition speed $\tau(t)$ is not a trivial matter, since the analysis requires methods beyond the Riemann-Lebesgue lemma. In the literature many different error approximation strategies are used [60–62]. In practical implementations, the spectral gap Δ_t will be mostly

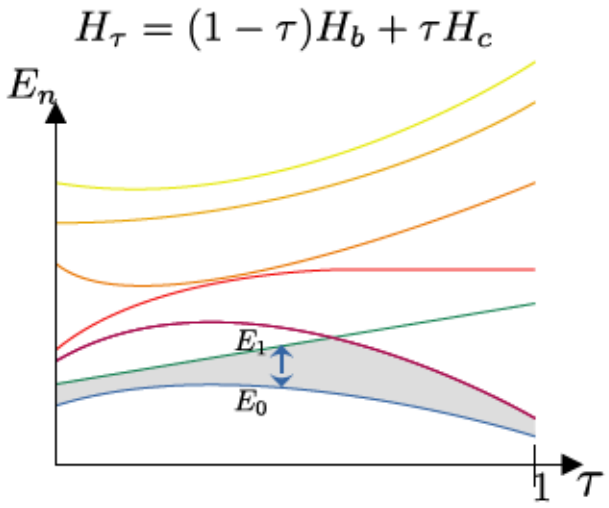

Fig. 3.4. – Sketch of a possible energy spectrum during an adiabatic transition. A large spectral gap $\Delta_t = E_1(t) - E_0(t)$ allows the use of a quicker transition speed. A level crossing (purple/green lines) means that the adiabatic transition converges to a different eigenstate.

unknown. Thus, heuristic schemes are needed to find a good transition functions $\tau(t)$.

If the transition time T is shortened, this leads to state migration to higher energy states. The error propagates first into the low energy excited states since $\Delta E_i = E_i - E_0$ will still be large enough to suppress transition to high energy states. As such, reducing T within reason will reduce the fidelity to the ground state, but may still give a sufficiently small expectation value.

For the purposes of quantum computation, one can discretize the evolution

$$|\mathrm{gs}_c\rangle = \lim_{T\to\infty} \mathcal{T} \int_0^1 e^{-iT((1-\tau)H_b + \tau H_c)} \mathrm{d}\tau \, |\mathrm{gs}_b\rangle \tag{3.55}$$

$$= \lim_{m\to\infty} \lim_{T\to\infty} \prod_{j=1}^{m} e^{-iT((1-\frac{j}{m})H_b + \frac{j}{m}H_c)} \, |\mathrm{gs}_b\rangle \tag{3.56}$$

$$= \lim_{k\to\infty} \lim_{m\to\infty} \lim_{T\to\infty} \prod_{j=1}^{m} \left(e^{-\mathrm{i}\frac{T}{k}(1-\frac{j}{m})H_b} e^{-\mathrm{i}\frac{T}{k}\frac{j}{m}H_c} \right)^k |\mathrm{gs}_b\rangle \tag{3.57}$$

$$\approx e^{-\mathrm{i}\theta_{2L}H_b} e^{-\mathrm{i}\theta_{2L-1}H_c} \cdots e^{-\mathrm{i}\theta_2 H_b} e^{-\mathrm{i}\theta_1 H_c} \, |\mathrm{gs}_b\rangle \, . \tag{3.58}$$

Here the first line is simply the adiabatic theorem restated, using the time order operator $\mathcal{T}$, which is an operator used to solve the time dependent Schrödinger evolution.

The second line is a discretization of the time ordered evolution, which amounts to solving the constant Hamiltonian Schrödinger equation for very small time steps $\frac{T}{m}\|H_\tau\| \ll 1$.

The next line is a first order trotterization. This allows us to describe the time evolution of H_τ using only the evolution from H_b and H_c. For this we also require that $\frac{T}{k}\|H_i\| \ll 1$ with $H_i \in \{H_b, H_c\}$.
If m and k are chosen sufficiently large, this gives us a strategy to prepare the ground state $|\mathrm{gs}_c\rangle$ by only applying the time evolution of H_b and H_c with computable phases $\{\theta_1, \theta_2, \ldots, \theta_{2L}\}$.

The outlined strategy is far from optimal. This is because there are many modifications that can reduce the number of required gates. We have already seen that a varying transition speed $\tau(t)$ can reduce the size of T. It is also advisable to use higher order and time dependent trotterization procedures [63] which allow for significantly larger step sizes and therefore smaller k and m while maintaining the same error guarantees. It should also be noted that unlike in original adiabatic theorem, any leakages will also go into high energy states as the eigenstates of H_τ are different to those of H_b and H_c. This means that great care needs to be taken to ensure that the correct $\boldsymbol{\theta}$ are applied.

Following the discretization strategy outlined above, even with higher order trotterization strategies, the evolution times θ_i are generally very small. This means the circuit consists of many gates, each very close to the identity gate and any implementation error will lead to leakages into high energy regimes. On near term devices, where gates are expensive to apply and individual gate errors are significant, such algorithms become quickly infeasible.

The approach of QAOA is to instead let the evolution times $\boldsymbol{\theta} \in \mathbb{R}^{2L}$ be classically tunable parameters

$$|\boldsymbol{\theta}\rangle := e^{-\mathrm{i}\theta_{2L}H_b} e^{-\mathrm{i}\theta_{2L-1}H_c} \cdots e^{-\mathrm{i}\theta_2 H_b} e^{-\mathrm{i}\theta_1 H_c} |\mathrm{gs}_b\rangle \ , \tag{3.59}$$

with the objective function

$$\langle O(\boldsymbol{\theta})\rangle = \langle \boldsymbol{\theta}|H_c|\boldsymbol{\theta}\rangle \ . \tag{3.60}$$

For the initial vector $\boldsymbol{\theta}_0 \in \mathbb{R}^L$ one can choose an estimate obtained from the discretized adiabatic theorem [64, 65], but then use VQA strategies to further amplify the overlap with low energy states of H_c. This allows to utilize the entire available gate set, as this heuristic approach is able to operate without requiring stringent error guarantees.

Overall QAOAs have the potential of greatly reducing the total gate count compared other adiabatic approaches, making them more feasible to run on near term devices.

However, they do experience the same difficulties that are also present in general VQA setups.

4 Results

During the optimization of VQA and QAOA instances, major challenges can arise. These include (i) the convergence into persistent local minima which may not be avoidable, (ii) a vanishing gradient with system size, which leads to significant measurement overheads, (iii) for physical ground states, the ansatz class may need a very large circuit complexity, making implentation difficult and infeasible on NISQ devices. In the following, we will introduce these challenges in more detail as well as summarize our contributions toward analyzing, understanding and potentially mitigating them. Finally we also mention not directly related work about a machine learning approach to classify quantum phase transitions. Similarly to VQAs, this proposal can be also seen as using classical computation to boost the predictive power of quantum measurement data. The full detail for each project is given in the attached papers in the appendices.

4.1 Optimization and local minima

Approaches to solve VQAs are mostly based on local optimization strategies. This means that one minimizes the cost function $\langle O(\boldsymbol{\theta})\rangle = \langle \boldsymbol{\theta}|O|\boldsymbol{\theta}\rangle$ based on their local behavior. A very basic and popular approach is gradient descent, where one chooses to change the current parameters $\boldsymbol{\theta}^{(i)}$ along the direction of the steepest descent

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta_i \frac{\partial}{\partial \boldsymbol{\theta}} \langle O(\boldsymbol{\theta})\rangle \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}} , \tag{4.1}$$

where η_i describes a step size which can be chosen by some predetermined strategy, or is itself optimized with a line search optimization. We discuss how to best estimate the gradient in section 4.2. There are many versions and generalizations of this approach. Some, like Newton's method, also use second order information about the Hessian of the cost function, while a popular intermediate method, BFGS, uses only some second order information to improve upon gradient descent. There has also been some research [66] showing that natural gradient decent [67] can work well for VQAs. The issue with all local methods is that while they generally converge to a local minimum, there is no guarantee that this minimum is close to the optimal

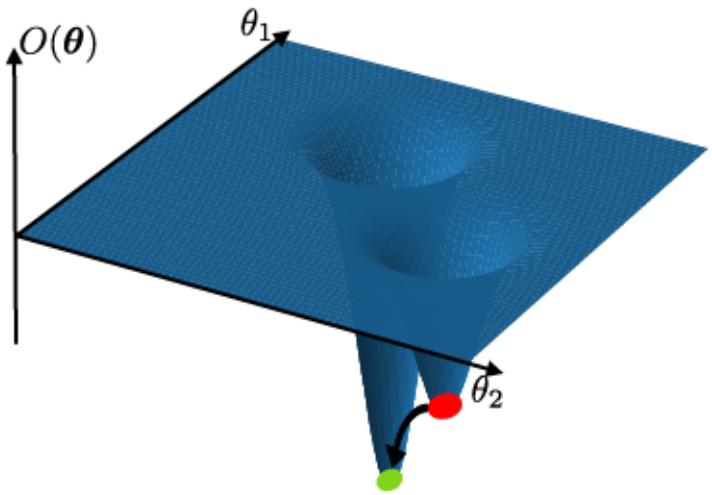

Fig. 4.1. – A sketch of the problem of local minima in two dimensions. The current position (red dot) has converged to a local minimum. However, there exists a significantly smaller minimum (green dot), which the local optimization procedure is unable to reach. This optimization error ϵ_{opt} is threrefore the difference between the two minima.

solution. Figure 4.1 shows an example of a sub-optimal convergence. For a single layer, we obtain the cost function

$$\langle O(\theta)\rangle = \langle\Psi_0|e^{i\theta H}Oe^{-i\theta H}|\Psi_0\rangle \tag{4.2}$$

$$= \sum_{i,j=1}^{n_\lambda} e^{-i(\lambda_i-\lambda_j)\theta}\langle\Psi_0|P_jOP_i|\Psi_0\rangle \tag{4.3}$$

$$= \sum_{k=1}^{n_\mu} e^{-i\mu_k\theta}c_k\,, \tag{4.4}$$

where P_i is the projector to the eigenspace of the eigenvalue λ_i of the generator $H = \sum_i \lambda_i P_i$. The frequencies μ_k describe all eigenvalue differences of the generator. As such, the VQA cost function is given by a Fourier series. In general, especially if the evolution is not periodic and many different frequencies (μ_k) contribute, even a single layer can have many local minima. In the simplest case $n_\lambda = 2$, (wlog. $\lambda = (0,1)$), it follows that the time evolution is given by

$$\langle O(\theta)\rangle = a_1\cos(\theta+\phi_0) + a_0\,, \tag{4.5}$$

with some real values (a_0, a_1, ϕ_0). This trigonometric function has two local extrema at $\theta = -\phi_0$ and $\theta = \pi - \phi_0$, where one describes a minimum and the other a maximum. In the multidimensional case, where each generator has two energy levels each, we obtain

$$\langle O(\boldsymbol{\theta})\rangle = \langle\boldsymbol{\theta}|O|\boldsymbol{\theta}\rangle \tag{4.6}$$

$$= \sum_{s,s'\in\{0,1\}^L} \mathrm{e}^{-\mathrm{i}\sum_{i=1}^L (s_i-s_i')\theta_i}\langle\Psi_0|P_{s'}^\dagger OP_s|\Psi_0\rangle \tag{4.7}$$

$$= \sum_{s\in\{-1,0,1\}^L} a_s\cos\left(\sum_{i=1}^L s_i\theta_i + \phi_s\right)\,, \tag{4.8}$$

where $P_s = P_{s_L}^{(L)}\dots P_{s_1}^{(1)}$ is a product of projectors and $\boldsymbol{a}\in\mathbb{R}^{3^L}, \boldsymbol{\phi}\in[0,2\pi)^{3^L}$ are real valued vectors. In general, for L layers, the cost function has at least 2^L local extrema, or points of vanishing gradient

$$\forall j\in[L]\,:\,0 = \frac{\partial\langle O(\boldsymbol{\theta})\rangle}{\partial\theta_j} = -\sum_{s\in\{-1,0,1\}^L} a_s s_j \sin\left(\sum_{i=1}^L s_i\theta_i+\phi_s\right)\,. \tag{4.9}$$

The second derivative or Hessian is given by

$$H_{j,k} = \frac{\partial^2\langle O(\boldsymbol{\theta})\rangle}{\partial\theta_j\partial\theta_k} = -\sum_{s\in\{-1,0,1\}^L} a_s s_j s_k\cos\left(\sum_{i=1}^L s_i\theta_i+\phi_s\right)\,. \tag{4.10}$$

A local minimum is given when the Hessian H is positive definite ($H>0$). Generally, the vast amount of the 2^L local extrema will describe saddle points, with only a

vanishing fraction being true local minimum. While it is difficult to estimate the number of local minima precisely, one can expect their number to grow quickly, potentially even exponentially with the number of layers [68]. The optimization can therefore be stuck in a far from optimal local minimum. We call the difference between the global minimum and the value returned by some optimization algorithm the *optimization error*. A common strategy to reduce this error is to reinitialize the optimization with different parameters, with the goal of terminating in a different local minima closer to the optimal value. This approach increases the overall runtime significantly since the entire protocol is now run multiple times, but the strategy works reliably if there is a significant likelyhood of a new initialization yielding a low energy state. In contrast, if there are many far from optimal local minima, reinitialization may not suffice to minimize the objective function. It may also be difficult to estimate the optimization error. As such, finding good termination and reinitialization criteria can be challenging.

4.1.1 Paper B - Training variational quantum algorithms is NP-hard

In the paper B, we show a NP-hardness of VQAs. Studying the VQA optimization problem with complexity theory can be useful. If we can show that the relevant problem is NP-hard, then there cannot exist a polynomial time algorithm which solves the required task (assuming P $\neq$ NP). In the particular case outlined above, this implies that exponentially many reinitializations may be required in the worst case.
We further define multiple decision versions of the VQA optimization problem and then show their respective NP-hardness with Karp reductions. The decision versions are so-called promise problems. The promise problems ask to decide between a YES instance, where $\langle O_{\text{min}} \rangle \leq \alpha$ for some $\alpha \in \mathbb{R}$ and a NO instance where $\langle O_{\text{min}} \rangle \geq \beta$. The promise is that the expectation value is either smaller than α or greater than β and nothing in between. For intermediate values of $\langle O_{\text{min}} \rangle$, the algorithm is effectively free to accept/reject at will. The promise gap $g = \beta - \alpha$ corresponds to desired precision of the optimization. If $\frac{1}{g} = \Omega(\text{poly}(N))$ is required for the problem, then a solver of the decision problem can only be used to determine $\langle O_{\text{min}} \rangle$ up to polynomial precision. If $\frac{1}{g} = \text{const.}$, then $\langle O_{\text{min}} \rangle$ can only be determined up to constant precision.

First, we define a setting where the classical optimization routine has access to the quantum expectation value, up to polynomial precision, through an oracle call. In this framework, not only is the optimization problem hard, but finding any non trivial approximation to the optimal value is NP-hard as well. As such, at least for a generic enough ansatz class, any practical algorithms will need to be heuristic in

nature and are unable to provide rigorous guarantees. Furthermore, the hardness is not even a result of the exponential Hilbert space dimension. Even if we consider qudit systems, where the Hilbert space dimension scales linearly in input size, we obtain an approximate hardness result. This result is related to known hardness results about multi-variable trigonometric polynomials [69]. The final category we were interested in involves free fermionic systems. They describe fermionic systems where there is no internal particle-particle interaction, effectively they are at most quadratic in the creation and annihilation operators $(c_i^\dagger, c_i)$. As such, the set of operation that can be performed is significantly reduced. If the system is initialized in a so-called *Gaussian state* the evolution remains in a family of states which can be fully described with at most quadratically many real parameters w.r.t. the number of Fermions, which is an exponential reduction in complexity. However, similarly to the qudit case, this restricted scenario already suffices for the hardness of approximation result.
In the paper we conclude that it is indeed the local minima present in the optimization which are the cause of the NP-hardness.

4.2 Measurement effort

Another major bottleneck for the optimization of VQAs will be the required measurement effort. When we consider gradient based methods, there are many components which can contribute to a significant measurement overhead:

- Multiple reinitialization to avoid local minima n_{reinit}
- Computation of the cost function and gradient for each optimization step $n_{\text{opt steps}}$
- Mulitple measurement settings for the estimation of the observable n_m
- Gradient estimation requires estimates of each partial derivative L
- Measurement at multiple measurement positions to determine the partial derivative $n_{\text{derivative}}$
- Sufficient statistics in each measurements to obtain the required accuracy $n_{\text{sn}} \propto \epsilon^{-2}$ due to shot noise errors

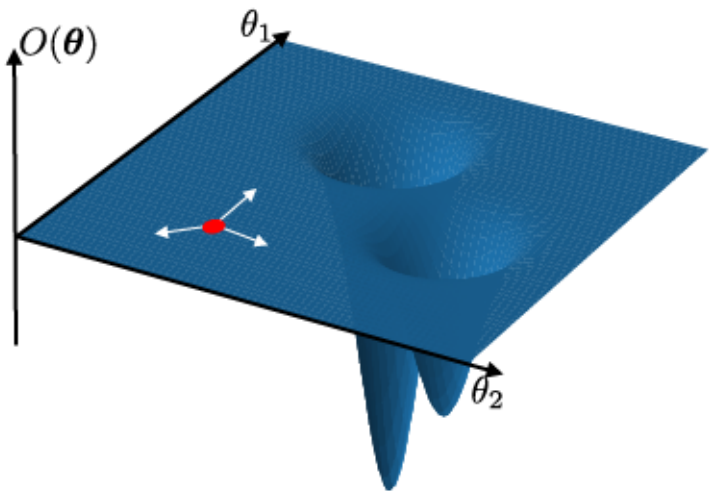

Fig. 4.2. – A sketch of a barren plateau in two dimensions. The current position (red dot) has an effectively vanishing gradient. As such, local optimizations routines are unable to converge to a local minimum, or only with tremendous measurement budgets.

The overall number of measurement rounds therefore is given by

$$n_{\text{tot}} = n_{\text{reinit}} \times n_m \times n_{O\text{ est}} \times L \times n_{\text{derivative}} \times O(\epsilon^{-2})\,. \tag{4.11}$$

Excluding the required measurement statistics, even with optimistic assumptions about the convergence rate, one can already expect a full optimization to have of the order of a thousand to a million different measurement positions, significantly depending on the specific problem observable and VQA ansatz. State preparation and measurement times vary massively between different implementations. Individual gate times for super conducting qubits are generally on the order of $10^{-8}s$ [70] with measurement times on the order of $10^{-7}s$ [71]. Trapped ion system are generally slower by nearly 3 orders of magnitude [72]. For a full one shot implementation of a circuit, one can therefore expect times of the order of microseconds to seconds. Adding a large multiplicative factor to obtain sufficient sampling statistics for each setting may therefore make the optimization procedure run incredibly slow, if not make it outright infeasible to perform.

4.2.1 Barren Plateaus

The problem of the measurement overhead is exacerbated by an effect that has come to be known as *barren plateus* [73]. The phenomena describes the general trend of VQA systems to have a gradient amplitude that is suppressed in the Hilbert space dimension, meaning exponentially small in the number of qubits. Naturally, this is a problem in local optimization schemes since determining the direction of an exponentially small gradient requires exponentially many measurements. To derive this suppressed behavior, we can estimate the amplitude of a gradient for some generic state. For the analysis, we represent a generic state as one drawn randomly from the Haar measure. We will motivate this assumption later. Explictly this means that $|\boldsymbol{\theta}\rangle$ is well described by

$$|\boldsymbol{\theta}\rangle = U\,|\Psi_0\rangle \tag{4.12}$$

where U is a Haar random unitary. For the derivative w.r.t. the last layer, we obtain

$$\begin{aligned}\langle O'(0)\rangle := \frac{\partial}{\partial\theta_L}\langle O(\theta)\rangle\,| &= \lim_{h\to 0}\frac{\langle O(h)\rangle - \langle O(0)\rangle}{h} && (4.13)\\ &= \lim_{h\to 0}\langle\boldsymbol{\theta}|\mathrm{e}^{\mathrm{i}Ht}O\mathrm{e}^{-\mathrm{i}Ht} - O|\boldsymbol{\theta}\rangle\,/h && (4.14)\\ &= \mathrm{i}\,\langle\boldsymbol{\theta}|[H,O]|\boldsymbol{\theta}\rangle\ , && (4.15)\end{aligned}$$

where $[A,B] := AB - BA$ is the commutator. For Haar random unitaries, we have the following identities

$$\int_{\text{Haar}} U\,|\Psi\rangle\langle\Psi|\,U^\dagger \mathrm{d}U = \frac{\mathbb{1}}{d}\,, \tag{4.16}$$

$$\begin{aligned}\int_{\text{Haar}} U^\dagger A U \rho U^\dagger B U \mathrm{d}U = \frac{\mathbb{1}\,\mathrm{Tr}(\rho)}{d}&\left(\frac{d\,\mathrm{Tr}(AB)}{d^2-1} - \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d^2-1}\right) \\ &+ \rho\left(\frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d^2-1} - \frac{\mathrm{Tr}(AB)}{d(d^2-1)}\right),\end{aligned} \tag{4.17}$$

where d is the Hilbert space dimension and the integral is taken over all unitaries according to the Haar measure. With this we obtain that the derivative vanishes in expectation,

$$\langle O'(0)\rangle_{\text{Haar}} = \mathrm{i}\int_{\text{Haar}}\langle\Psi_0|U^\dagger[H,O]U|\Psi_0\rangle\,dU = \mathrm{i}\frac{\mathrm{Tr}[\mathbb{1}[H,O]]}{d} = 0\,. \tag{4.18}$$

We can also derive the expected square of the gradient

$$\langle O'(0)^2\rangle_{\text{Haar}} = \int_{\text{Haar}} \langle\Psi_0|U^\dagger[H,O]U|\Psi_0\rangle \langle\Psi_0|U^\dagger[H,O]U|\Psi_0\rangle \, dU \tag{4.19}$$

$$= \frac{|\langle\Psi_0|\Psi_0\rangle}{d}\left(\frac{d\,\text{Tr}[[H,O]^2]}{d^2-1} - \frac{\text{Tr}[[H,O]]^2}{d^2-1}\right) \tag{4.20}$$

$$+ |\langle\Psi_0|\Psi_0\rangle|^2\left(\frac{d\,\text{Tr}[[H,O]]^2}{d^2-1} - \frac{\text{Tr}[[H,O]^2]}{d(d^2-1)}\right) \tag{4.21}$$

$$= \frac{\text{Tr}[[H,O]^2]/d}{d+1} \tag{4.22}$$

$$\leq \frac{4\|O\|_\infty^2\|H\|_\infty^2}{d+1}, \tag{4.23}$$

which used that $\text{Tr}([H,O]) = 0$. The expected square of the gradient is therefore suppressed by the Hilbert space dimension. Similar results also follow for derivatives with respect to other layers. The assumption that $|\boldsymbol{\theta}\rangle$ is Haar random is generally not satisfied. The technical requirement is only that the VQA circuit behaves as if it is drawn from an approximate 2-design [74], which is significantly weaker and often already occurs within quadratic cirucit complexity [73, 75]. This effect will always occur if two aspects are satisfied: (1) The VQA circuit describes a universal gate-set and (2) a random initialization is used for the initial parameters $\boldsymbol{\theta}_0$. To avoid barren plateaus, one therefore either needs to find a good initialization which is not in the barren plateau regime or use a set of generators which are not universal.

4.2.2 Paper C- Fast gradient estimation for variational quantum algorithms

In the paper C we are developing algorithms to best estimate the partial derivative of a cost function with a fixed measurement budget. In the literature, there are generally two approaches used to find the partial derivative. The first is using a finite difference methods like central differences

$$\langle O'(0)\rangle = \frac{\langle O(h)\rangle - \langle O(-h)\rangle)}{2h} + O(h^2), \tag{4.24}$$

which while being simple, has the problem that it does lead to systematic errors. Alternatively, if the spectrum is a known discrete spectrum, the Parameter Shift Rule (PSR) [76, 77] can be used to find the gradient exactly. For a cost function as defined in eq. (4.4), with $\lambda = \{0,1\}$ one obtains

$$\langle O'(0)\rangle = \frac{\langle O(x^*)\rangle - \langle O(-x^*)\rangle)}{2\sin(x^*)} = \frac{\langle O(\pi/2)\rangle - \langle O(-\pi/2)\rangle}{2} \tag{4.25}$$

where the statement holds for all x^*, but $x^* = \pi/2$ is preferred since it gives the best statistics. This result follows since, as we saw, the function is described by

$\langle O(\theta)\rangle = a_1 \cos(\theta + \phi_0) + a_2$. In general, if the cost function oscillates with n_μ many different frequencies, $2n_\mu$ measurement positions are required to find the exact gradient. The optimal strategy can be obtained by inverting a discrete sine transform.

When it comes to unitaries with an unknown spectrum, or many different eigenvalues, the PSR method becomes impractical, especially on a limited measurement budget. For central differences, if $h \to 0$, the statistics become very poor, since $\langle O(h)\rangle \sim \langle O(-h)\rangle$ means simply separating the two values already requires high precision. Our proposal combines both approaches to find an optimal measurement strategy depending on the available measurement rounds. The idea here is to use a Bayesian framework that can also model systematic errors in the gradient estimation. For this we use prior estimates to predict the expected size of the Fourier coefficient of the cost function, which allows us to find the optimal measurement strategy for the gradient. To obtain prior estimates of the coefficient from eq. (4.7)

$$\langle c_{s,s'}^2 \rangle = \langle \Psi | P_{s'} O P_s | \Psi \rangle \ , \tag{4.26}$$

we develop techniques based around design convergences properties and a scaling analysis from smaller systems. Numerically, we also find very good agreement with the theoretical predictions. In the regime before the 2-design convergence has occurred, we develop interpolation methods which allow us to make sufficiently accurate predictions for the purpose of the estimation procedure.
Our analysis shows that the PSR is the optimal strategy for very large measurement budgets, while a central differences approach with a large step size h is preferable when the measurement budget is relatively small. In the intermediate regime, we use convex optimization methods to obtain the optimal measurement allocation strategy. Here, with an increasing measurement budget, the number of positions gradually increases with the overall strategy slowly approaching that of PSR. We show in numerical simulations for a gradient descent optimization, that the overall measurement effort can be reduced significantly by using our method while leading the same quality of the result.

4.3 Efficient ansatz classes - circuit depth optimization

As we saw in the previous section, allowing an ansatz class that spans too much of the state space will lead to barren plateaus and therefore makes the VQA untrainable, which can occur already with moderate circuit depth. An issue when the observable O describes some quantum many body Hamiltonian is that, when our complexity

theoretic assumptions are correct (QCMA $\neq$ QMA), a low energy state of a physical Hamiltonian can have an exponentially high circuit complexity, meaning their preparation is practically infeasible, especially on NISQ devices. Since finding circuit lower and upper bounds is very hard in general [78], it is still an open question which circuit depth will be required to prepare many practically relevant ground states. Finding a sufficiently expressive ansatz with the shortest circuit complexity istherefore another main challenge of VQAs.

4.3.1 Paper D - Optimizing the depth of variational quantum algorithms is strongly QCMA-hard to approximate

In the paper D, we analyze the complexity of VQAs with a particular focus on the number of required layers L. A natural question is to ask what the smallest number of layers L_{opt} is to reach the ground state energy. Since the depth is now a scaling parameter, we choose the circuit design where each layer can select from an allowed gate set. In the paper we define the VQA decision problem and show a many-one (Karp) reduction to the Quantum Monontone Satisfying Assignment (QMSA) problem which is known to be QCMA-hard. Additionally we use the hardness of approximation result for the Hamming weight of a QMSA solution to show a similar hardness result w.r.t. the circuit depth of VQAs. As such, if BQP $\subsetneq$ QCMA, finding the smallest number of layers cannot be done efficiently on a quantum computer. More explicitly we show that for any polynomial time algorithm and for every $\epsilon > 0$ there have to exist at least some instances, where the ratio between the depth given by the algorithm L_{alg} and the optimal depth L_{opt} scales

$$\frac{L_{\text{alg}}}{L_{\text{opt}}} \geq N^{1-\epsilon}, \tag{4.27}$$

where N is the encoding size of the VQA instance.

We saw that QAOAs convergence for arbitrarily large circuit depth L, but that it is unclear by how much the depth can be reduced. In the paper, we also derive the same hardness of approximation result for QAOAs. This gives an indication that finding strategies which significantly outperform the adiabatic theorem may be possible, but that finding them strategy may be potentially very difficult.

4.4 Many-body localization

In this section we discuss work on Many-Body Localization (MBL) [79], which is a quantum mechanical, high energy phase transition between a localized phase (MBL),

and a delocalized phase. A common physical model that is used both for numerical and analytical studies is the Heisenberg Hamiltonian with a local random potential. The Hamiltonian is given by

$$H = \sum_{i=1}^{n} h_i \sigma_i^z + \sigma_i^x \sigma_{i+1}^x + \sigma_i^y \sigma_{i+1}^z + \sigma_i^z \sigma_{i+1}^z \,, \tag{4.28}$$

where the latter terms describe interaction terms, while the first term describes a local magentic field. The amplitudes are sampled uniformly from an interval $h_i \in [-h, h]$, where h is the magnetic field strength. For small h the system thermalizes, meaning that a typical evolution creates correlations on large distances and reaches high entanglement entropy states. This behavior is often analyzed in the framework of the Eigenstate Thermalization Hypothesis (ETH), which is a series of conditions which allow a reversible unitary evolution to be described in the terminology of statistical physics, meaning irreversibility of time, agnosticity of the initial state, convergence to local Gibbs ensembles, etc.
In contrast, for a large field strength h, the system experiences a transition into a localized phase called many body localization (MBL). Here, no long range correlations can be observed and the system has localized eigenstates. This phase is still poorly understood, mostly because many analytic approaches which work for Anderson localization [80, 81], a similar phase but without particle-particle interaction, do not work for the many-body case and numerical simulations are restricted to few qubits, making an analysis of the scaling behavior into the thermodynamic limit very difficult. It is assumed that the system has a phase transition for a magnetic field strength of around $h \approx 6$. Especially for MBL in higher dimensions, near term experiments can be a crucial tool to understand the MBL behavior beyond the limits existing simulation algorithms.

4.4.1 Paper E - Scalable approach to many-body localization via quantum data

The paper E is a project using machine learning to understand and predict the behavior of MBL Hamiltonians. Our approach uses individual disorder realizations (i.e. the vector $\boldsymbol{h} \in [-h, h]^n$) to predict properties of the system. As is common in numerical approaches [82], we use multiple indicators to characterize the system. These indicators include correlation functions, entanglement entropies, as well as expectation values for time evolved states. Spectral properties are also used since they are affected by the transition. A machine learning algorithm uses numerically obtained values for different magnetic fields

$$\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_N\} \tag{4.29}$$

to learn values of the indicators and therefore also the MBL/ETH phase transition. On finite sizes, the localization behavior will naturally depend on the particular sample (h) of the magnetic field. As such our approach is also able to correlate these indicators with a particular magnetic field distribution. With the help of recurrent neural networks, we are able to learn the indicators for small system sizes and generalize the prediction for larger system sizes. While this approach may be susceptible to finite size effects present in the training data, it does offer a potential path forward, especially when future quantum experiments can offer a more reliable set of training data for larger system sizes. This project falls in the category of classical learning algorithms using quantum data for training. In the future, when quantum computers become more readily available, it will be an important task to actually interpret the quantum measurement results. Currently, it is still very difficult estimate the precise impact of the quantum data, since future experimental and theoretical developments in quantum computing may have tremendous effects. Instead one can focus on the much better understood classical data analysis tools which will be required for the post-processing to show potential avenues for quantum advantage with quantum data.

5 Conclusion and open questions

There is still a lot unknown about the computational power of quantum computers. While there is good evidence to suggest that they can significantly outperform classical computers in practically relevant tasks, it is still unclear if they will be broadly used or only utilized as a specialized tool. For quantum computation without error correction (NISQ) the question of practical usefulness is still unresolved. Our work focuses on one particular NISQ proposal, VQAs and we have derived specific hardness results for them. Our results seem to imply that VQAs cannot be used not a black box tool to solve arbitrary optimization problems. However, if the outlined measurement bottleneck can be mitigated, they might proof themselves as a useful to enhance the performance of other ground state reparation protocols.

Additionally, the work also opens up new questions and research directions. We are currently looking at the question of how well a particular ansatz class for VQAs can find low energy states. Indeed, we are working on showing that it is NEXP-hard to decide if a particular ansatz contains low energy states or not. We are also working on extending these results to more physics focused problems about time evolving systems and thermalization. Similarly, we believe the Bayesian framework introduced for the gradient estimation can be applied more widely, for instance in the context of Hamiltonian learning [83, 84] and questions about the thermalization of quantum systems.

Bibliography

[1] R. P. Feynman, Simulating physics with computers, in *Feynman and computation* (CRC Press, 2018), pp. 133–153 (cit. on p. 1).

[2] C. Gross and I. Bloch, Quantum simulations with ultracold atoms in optical lattices, Science **357**, 995 (2017) (cit. on p. 1).

[3] F. Schäfer, T. Fukuhara, S. Sugawa, Y. Takasu, and Y. Takahashi, Tools for quantum simulation with ultracold atoms in optical lattices, Nature Reviews Physics **2**, 411 (2020) (cit. on p. 1).

[4] A. Y. Kitaev, Quantum computations: algorithms and error correction, Russian Mathematical Surveys **52**, 1191 (1997) (cit. on p. 1).

[5] D. Aharonov and M. Ben-Or, „Fault-tolerant quantum computation with constant error“, in Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (1997), pp. 176–188 (cit. on pp. 1, 22).

[6] Suppressing quantum errors by scaling a surface code logical qubit, Nature **614**, 676 (2023) (cit. on p. 2).

[7] L. Stockmeyer, Classifying the computational complexity of problems, The journal of symbolic logic **52**, 1 (1987) (cit. on p. 2).

[8] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, SIAM review **41**, 303 (1999) (cit. on p. 2).

[9] P. Carl, A tale of two sieves, Notices of the American Mathematical Society **43**, 1473 (1996) (cit. on p. 2).

[10] S. Aaronson, „BQP and the polynomial hierarchy“, in Proceedings of the forty-second ACM symposium on Theory of computing (2010), pp. 141–150 (cit. on p. 2).

[11] M. Grohe and P. Schweitzer, The graph isomorphism problem, Communications of the ACM **63**, 128 (2020) (cit. on p. 3).

[12] J. Haah, M. B. Hastings, R. Kothari, and G. H. Low, Quantum algorithm for simulating real time evolution of lattice Hamiltonians, SIAM Journal on Computing, FOCS18 (2021) (cit. on p. 3).

[13] A. J. Daley, I. Bloch, C. Kokail, S. Flannigan, N. Pearson, M. Troyer, and P. Zoller, Practical quantum advantage in quantum simulation, Nature **607**, 667 (2022) (cit. on p. 3).

[14] D. Aharonov and I. Arad, The BQP-hardness of approximating the Jones polynomial, New Journal of Physics **13**, 035019 (2011) (cit. on p. 3).

[15] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018) (cit. on p. 3).

[16] S. Aaronson and A. Arkhipov, „The computational complexity of linear optics“, in Proceedings of the forty-third annual ACM symposium on Theory of computing (2011), pp. 333–342 (cit. on p. 3).

[17] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, et al., Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019) (cit. on p. 3).

[18] D. Hangleiter, M. Kliesch, J. Eisert, and C. Gogolin, Sample complexity of device-independently certified "quantum supremacy", Physical review letters **122**, 210502 (2019) (cit. on p. 3).

[19] A. Schonhage, Schnelle multiplikation grosser zahlen, Computing **7**, 281 (1971) (cit. on p. 7).

[20] M. Fürer, „Faster integer multiplication“, in Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (2007), pp. 57–66 (cit. on p. 7).

[21] D. Harvey, J. Van Der Hoeven, and G. Lecerf, Even faster integer multiplication, Journal of Complexity **36**, 1 (2016) (cit. on p. 7).

[22] W. T. Cochran, J. W. Cooley, D. L. Favin, H. D. Helms, R. A. Kaenel, W. W. Lang, G. C. Maling, D. E. Nelson, C. M. Rader, and P. D. Welch, What is the fast Fourier transform?, Proceedings of the IEEE **55**, 1664 (1967) (cit. on p. 7).

[23] S. Arora, „Polynomial time approximation schemes for Euclidean TSP and other geometric problems“, in Proceedings of 37th Conference on Foundations of Computer Science (IEEE, 1996), pp. 2–11 (cit. on p. 7).

[24] C. H. Papadimitriou, Computational complexity, in *Encyclopedia of computer science* (2003), pp. 260–265 (cit. on p. 10).

[25] A. Cobham, The intrinsic computational difficulty of functions, (1965) (cit. on p. 11).

[26] J. Hartmanis and R. E. Stearns, On the computational complexity of algorithms, Transactions of the American Mathematical Society **117**, 285 (1965) (cit. on p. 12).

[27] S. A. Cook, „The complexity of theorem-proving procedures“, in Proceedings of the third annual ACM symposium on Theory of computing (1971), pp. 151–158 (cit. on p. 14).

[28] R. M. Karp, *Reducibility among combinatorial problems* (Springer, 2010) (cit. on p. 15).

[29] L. Longpré and P. Young, Cook reducibility is faster than Karp reducibility in NP, Journal of Computer and System Sciences **41**, 389 (1990) (cit. on p. 15).

[30] O. Goldreich, In a World of P= BPP. Studies in Complexity and Cryptography **6650**, 191 (2011) (cit. on p. 17).

[31] M. Agrawal, N. Kayal, and N. Saxena, PRIMES is in P, Annals of mathematics, 781 (2004) (cit. on p. 17).

[32] N. Saxena, Progress on Polynomial Identity Testing. Bull. EATCS **99**, 49 (2009) (cit. on p. 17).

[33] Y. Shi, Both Toffoli and controlled-NOT need little help to do universal quantum computation, arXiv preprint quant-ph/0205115 (2002) (cit. on p. 20).

[34] A. Y. Kitaev, A. Shen, M. N. Vyalyi, and M. N. Vyalyi, *Classical and quantum computation*, 47 (American Mathematical Soc., 2002) (cit. on p. 22).

[35] A. Kitaev and J. Watrous, „Parallelization, amplification, and exponential time simulation of quantum interactive proof systems“, in Proceedings of the thirty-second annual ACM symposium on Theory of computing (2000), pp. 608–617 (cit. on p. 22).

[36] A. Anshu and C. Nirkhe, Circuit lower bounds for low-energy states of quantum code Hamiltonians, arXiv preprint arXiv:2011.02044 (2020) (cit. on p. 23).

[37] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, A variational eigenvalue solver on a photonic quantum processor, Nature communications 5, 4213 (2014) (cit. on p. 25).

[38] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New Journal of Physics **18**, 023023 (2016) (cit. on p. 25).

[39] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, et al., Noisy intermediate-scale quantum algorithms, Reviews of Modern Physics **94**, 015004 (2022) (cit. on p. 25).

[40] U. Dorner, R. Demkowicz-Dobrzanski, B. J. Smith, J. S. Lundeen, W. Wasilewski, K. Banaszek, and I. A. Walmsley, Optimal quantum phase estimation, Physical review letters **102**, 040403 (2009) (cit. on p. 25).

[41] D. Wierichs, C. Gogolin, and M. Kastoryano, Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer, Physical Review Research **2**, 043246 (2020) (cit. on p. 25).

[42] S. Wei, H. Li, and G. Long, A full quantum eigensolver for quantum chemistry simulations, Research (2020) (cit. on p. 25).

[43] J. A. Nelder and R. Mead, A simplex method for function minimization, The computer journal **7**, 308 (1965) (cit. on p. 25).

[44] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, et al., Variational quantum algorithms, Nature Reviews Physics **3**, 625 (2021) (cit. on p. 25).

[45] J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, E. Grant, L. Wossnig, I. Rungger, G. H. Booth, et al., The variational quantum eigensolver: a review of methods and best practices, Physics Reports **986**, 1 (2022) (cit. on p. 25).

[46] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv preprint arXiv:1411.4028 (2014) (cit. on pp. 25, 34).

[47] S. Hadfield, Z. Wang, B. O'gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019) (cit. on pp. 25, 34).

[48] H. L. Tang, V. Shkolnikov, G. S. Barron, H. R. Grimsley, N. J. Mayhall, E. Barnes, and S. E. Economou, qubit-adapt-vqe: An adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor, PRX Quantum **2**, 020310 (2021) (cit. on p. 27).

[49] R. Miceli and M. McGuigan, „Effective matrix model for nuclear physics on a quantum computer", in 2019 New York Scientific Data Summit (NYSDS) (IEEE, 2019), pp. 1–4 (cit. on p. 29).

[50] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. Sawaya, et al., Quantum chemistry in the age of quantum computing, Chemical reviews **119**, 10856 (2019) (cit. on p. 29).

[51] V. Lordi and J. M. Nichol, Advances and opportunities in materials science for scalable quantum computing, MRS Bulletin **46**, 589 (2021) (cit. on p. 29).

[52] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Physical review **136**, B864 (1964) (cit. on p. 29).

[53] P. Jordan and E. P. Wigner, *Über das paulische äquivalenzverbot* (Springer, 1993) (cit. on p. 30).

[54] A. Tranter, P. J. Love, F. Mintert, and P. V. Coveney, A comparison of the Bravyi–Kitaev and Jordan–Wigner transformations for the quantum simulation of quantum chemistry, Journal of chemical theory and computation **14**, 5617 (2018) (cit. on p. 30).

[55] K. Setia and J. D. Whitfield, Bravyi-Kitaev Superfast simulation of electronic structure on a quantum computer, The Journal of chemical physics **148**, 164104 (2018) (cit. on p. 30).

[56] A. Nizovtsev, S. Y. Kilin, F Jelezko, T Gaebal, I. Popa, A Gruber, and J. Wrachtrup, A quantum computer based on NV centers in diamond: optically detected nutations of single electron and nuclear spins, Optics and spectroscopy **99**, 233 (2005) (cit. on p. 30).

[57] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nature Physics **16**, 1050 (2020) (cit. on p. 31).

[58] A. Gresch and M. Kliesch, Guaranteed efficient energy estimation of quantum many-body Hamiltonians using ShadowGrouping, arXiv preprint arXiv:2301.03385 (2023) (cit. on p. 33).

[59] M. Born and V. Fock, Beweis des adiabatensatzes, Zeitschrift für Physik **51**, 165 (1928) (cit. on p. 34).

[60] G. A. Hagedorn and A. Joye, Elementary exponential error estimates for the adiabatic approximation, Journal of mathematical analysis and applications **267**, 235 (2002) (cit. on p. 36).

[61] A. Ambainis and O. Regev, An elementary proof of the quantum adiabatic theorem, arXiv preprint quant-ph/0411152 (2004) (cit. on p. 36).

[62] M. H. Amin, Consistency of the adiabatic theorem, Physical review letters **102**, 220401 (2009) (cit. on p. 36).

[63] T. N. Ikeda, A. Abrar, I. L. Chuang, and S. Sugiura, Minimum Fourth-Order Trotterization Formula for a Time-Dependent Hamiltonian, arXiv preprint arXiv:2212.06788 (2022) (cit. on p. 38).

[64] B. F. Schiffer, J. Tura, and J. I. Cirac, Adiabatic spectroscopy and a variational quantum adiabatic algorithm, PRX Quantum **3**, 020347 (2022) (cit. on p. 38).

[65] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices, Physical Review X **10**, 021067 (2020) (cit. on p. 38).

[66] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum natural gradient, Quantum **4**, 269 (2020) (cit. on p. 41).

[67] S.-I. Amari, Natural gradient works efficiently in learning, Neural computation **10**, 251 (1998) (cit. on p. 41).

[68] B. Kalantari, Quadratic functions with exponential number of local maxima, Operations research letters **5**, 47 (1986) (cit. on p. 44).

[69] L. Pfister and Y. Bresler, Bounding multivariate trigonometric polynomials, IEEE Transactions on Signal Processing **67**, 700 (2018) (cit. on p. 45).

[70] D. Basilewitsch, C. Dlaska, and W. Lechner, Comparing planar quantum computing platforms at the quantum speed limit, arXiv preprint arXiv:2304.01756 (2023) (cit. on p. 46).

[71] E. Jeffrey, D. Sank, J. Mutus, T. White, J Kelly, R Barends, Y Chen, Z Chen, B Chiaro, A Dunsworth, et al., Fast accurate state measurement with superconducting qubits, Physical review letters **112**, 190504 (2014) (cit. on p. 46).

[72] V. Schäfer, C. Ballance, K Thirumalai, L. Stephenson, T. Ballance, A. Steane, and D. Lucas, Fast quantum logic gates with trapped-ion qubits, Nature **555**, 75 (2018) (cit. on p. 46).

[73] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nature communications **9**, 4812 (2018) (cit. on pp. 47, 48).

[74] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Physical Review A **80**, 012304 (2009) (cit. on p. 48).

[75] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, Communications in Mathematical Physics **291**, 257 (2009) (cit. on p. 48).

[76] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, General parameter-shift rules for quantum gradients, Quantum **6**, 677 (2022) (cit. on p. 48).

[77] G. E. Crooks, Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition, arXiv preprint arXiv:1905.13311 (2019) (cit. on p. 48).

[78] I. Wegener, *The complexity of Boolean functions* (John Wiley & Sons, Inc., 1987) (cit. on p. 50).

[79] D. A. Abanin and Z. Papić, Recent progress in many-body localization, Annalen der Physik **529**, 1700169 (2017) (cit. on p. 50).

[80] P. W. Anderson, Absence of diffusion in certain random lattices, Physical review **109**, 1492 (1958) (cit. on p. 51).

[81] S. Fishman, D. Grempel, and R. Prange, Chaos, quantum recurrences, and Anderson localization, Physical Review Letters **49**, 509 (1982) (cit. on p. 51).

[82] F. Alet and N. Laflorencie, Many-body localization: An introduction and selected topics, Comptes Rendus Physique **19**, 498 (2018) (cit. on p. 51).

[83] C. E. Granade, C. Ferrie, N. Wiebe, and D. G. Cory, Robust online Hamiltonian learning, New Journal of Physics **14**, 103013 (2012) (cit. on p. 53).

[84] J. Wang, S. Paesani, R. Santagati, S. Knauer, A. A. Gentile, N. Wiebe, M. Petruzzella, J. L. O'brien, J. G. Rarity, A. Laing, et al., Experimental quantum Hamiltonian learning, Nature Physics **13**, 551 (2017) (cit. on p. 53).

[85] L. Tendick, H. Kampermann, and D. Bruß, Activation of Nonlocality in Bound Entanglement, Phys. Rev. Lett. **124**, 050401 (2020) (cit. on p. 65).

[86] L. Bittel and M. Kliesch, Training variational quantum algorithms is np-hard, Physical review letters **127**, 120502 (2021) (cit. on p. 75).

[87] L. Bittel, J. Watty, and M. Kliesch, Fast gradient estimation for variational quantum algorithms, arXiv preprint arXiv:2210.06484 (2022) (cit. on p. 77).

[88] L. Bittel, S. Gharibian, and M. Kliesch, Optimizing the depth of variational quantum algorithms is strongly QCMA-hard to approximate, arXiv preprint arXiv:2211.12519 (2022) (cit. on p. 105).

[89] A. Gresch, L. Bittel, and M. Kliesch, Scalable approach to many-body localization via quantum data, arXiv preprint arXiv:2202.08853 (2022) (cit. on p. 139).

List of Acronyms

VQA	Variational Quantum Algorithm	3
VQE	Variational Quantum Eigensolver	25
QAOA	Quantum Alternating Operator Ansatz	25
QMSA	Quantum Monontone Satisfying Assignment	50
NISQ	Noisy Intermediate-Scale Quantum	3
PSR	Parameter Shift Rule	48
ETH	Eigenstate Thermalization Hypothesis	51
MBL	Many-Body Localization	50
SDP	semidefinite program	65

A

Paper: Activation of nonlocality in bound entanglement

Title:	Activation of Nonlocality in Bound Entanglement
Authors:	Lucas Tendick, Hermann Kampermann, and Dagmar Bruß
Journal:	Physical Review Letters
Impact factor:	9.185 (2021)
Date of submission:	24 April 2019
Publication status:	Published
Contribution by LT:	First author (input approx. 85%)

This publication corresponds to the paper [85]. The summary of the results is presented in chapter ??.

This work was a continuation of the studies in my Master's thesis. I had the initial idea to work on hidden nonlocality. Together with HK and DB, I collected some ideas for states that might be interesting to investigate, with bound entangled states among them. Then, I checked the literature for papers with appropriate methods to use. HK suggested relevant sources to me to get familiar with semidefinite programs (SDPs). In the following, I conducted the central part of the research, i.e., the numerical search for local bound entangled states that reveal hidden nonlocality. In between I had several discussions with DB and especially HK. After obtaining the numerical results, I showed that it is possible to retrieve an analytical description of the state from the numerics. The results were discussed together with HK and DB. Finally, I wrote the whole manuscript, which my co-authors proofread. I improved the manuscript based on my co-authors' comments on several drafts of the paper.

Activation of Nonlocality in Bound Entanglement

Lucas Tendick,* Hermann Kampermann, and Dagmar Bruß
Institute for Theoretical Physics III, Heinrich-Heine-Universität Düsseldorf, D-40225 Düsseldorf, Germany

(Received 24 April 2019; revised manuscript received 26 September 2019; accepted 10 January 2020; published 3 February 2020)

We discuss the relation between entanglement and nonlocality in the hidden nonlocality scenario. Hidden nonlocality signifies nonlocality that can be activated by applying local filters to a particular state that admits a local hidden-variable model in the Bell scenario. We present a fully biseparable three-qubit bound entangled state with a local model for the most general (nonsequential) measurements. This proves for the first time that bound entangled states can admit a local model for general measurements. We furthermore show that the local model breaks down when suitable local filters are applied. Our results demonstrate the first example of activation of nonlocality in bound entanglement. Hence, we show that genuine hidden nonlocality does not imply entanglement distillability.

DOI: 10.1103/PhysRevLett.124.050401

Performing local measurements on certain entangled quantum states can lead to the phenomenon of quantum nonlocality. That is, the correlations obtained from the measurements are not compatible with the principle of local realism, witnessed by the violation of a so-called Bell inequality [1]. Although entanglement and nonlocality were extensively studied since the foundation of quantum theory [2,3], the relation between both is still not fully understood.

After the seminal work by Bell [1] as an answer to the EPR-Gedankenexperiment [4], it was widely believed that entanglement and nonlocality are just two different notions of the inseparability of quantum states. Indeed, for pure entangled states nonlocality is a generic feature [5,6]. However, Werner [7] showed that there exist mixed entangled states (so-called Werner states) which admit a local model for projective measurements. Later, Barrett [8] extended this result by showing that certain Werner states admit a local model even when positive-operator valued measures (POVMs), i.e., most general nonsequential measurements are considered. This displays the inequivalence of entanglement and nonlocality in the Bell scenario.

It was first noticed by Popescu [9] and more recently by Hirsch *et al.* [10] that some local entangled states can violate a Bell inequality when the observers apply judicious local filters as probabilistic preselection before the Bell test. This phenomenon is referred to as hidden nonlocality, or as genuine hidden nonlocality when one considers an entangled quantum state ρ with a local model even for POVMs. However, it was shown that genuine hidden nonlocality is not a general feature [11]. For example, a particular two-qubit Werner state remains local even after arbitrary local filtering.

Note that hidden nonlocality is not the only extension of the Bell scenario. For instance, nonlocality can also be superactivated [12,13] by allowing the parties to perform joint measurements on multiple copies of a local entangled state. An even more general concept is that of entanglement distillation [3]. In this scenario the parties have access to both local filters and multiple copies of a given state, with the goal to probabilistically obtain pure entangled states. Distillable states can therefore always be seen as nonlocal resource in the so-called asymptotic scenario [14]. However, there exist entangled states which are not distillable to pure entangled states. These states build the famous set of bound entangled states [15], which were the subject of various scientific works in the past [16–20]. Studying the nonlocal properties of bound entangled states will approach the answer of the fundamental open question of whether all entangled states are nonlocal resources. Since bound entanglement is the weakest form of entanglement, it was conjectured by Peres [21] that bound entangled states cannot lead to any nonlocal correlations at all. However, nowadays we know that the Peres conjecture is false [22,23]: bound entangled states can violate a Bell inequality. Despite these results and more advanced scenarios [24], nothing is known about the activation of local bound entanglement.

In this Letter, we answer the open question of whether bound entangled states with genuine hidden nonlocality exist in the affirmative. Specifically, we show that a certain three-qubit bound entangled state with a local model for POVMs can violate a Bell inequality when local filters are applied. This proves that genuine hidden quantum nonlocality does not imply entanglement distillability. Our results and possible extensions are visualized in Fig. 1.

Preliminaries.—Consider three distant parties Alice, Bob, and Charlie sharing an entangled quantum state ρ. The parties can perform local measurements via the positive semidefinite operators $M_{a|x}$, $M_{b|y}$, and $M_{c|z}$ with the settings x, y, z and the outcomes a, b, c. These operators form POVMs, as they satisfy the completeness relation

FIG. 1. Abstract overview of our results. We show that the set of nonlocal bound entangled states (BE states) can be enlarged in the hidden nonlocality scenario (HNL). This is the first step towards a possible equivalence of all BE states and all nonlocal BE states. Further enlargements of the set of nonlocal BE states could be provided by superactivation (SA) and the asymptotic scenario (Asymp.), similar to the case for distillable states. It is also an open question, whether the set can be enlarged to all BE states in such scenarios.

$\sum_a M_{a|x} = \mathbb{1}$ (and analogously for Bob and Charlie), where $\mathbb{1}$ denotes the identity operator. The resulting statistics is given by

$$p(abc|xyz) = \mathrm{Tr}[(M_{a|x} \otimes M_{b|y} \otimes M_{c|z})\rho]. \quad (1)$$

The state ρ is said to be local (for $\{M_{a|x}\}$, $\{M_{b|y}\}$, and $\{M_{c|z}\}$) if the distribution (1) admits a local decomposition of the following form:

$$p(abc|xyz) = \int \pi(\lambda) p(a|x\lambda) p(b|y\lambda) p(c|z\lambda) d\lambda. \quad (2)$$

That is, the statistics can be explained by a local hidden-variable model (LHV), where $\lambda \in \mathbb{R}$ is the shared local hidden variable, distributed according to the density $\pi(\lambda)$ such that $\int \pi(\lambda) d\lambda = 1$. The probability distributions $p(a|x\lambda)$, $p(b|y\lambda)$, and $p(c|z\lambda)$ are typically called local response functions in this context. A state ρ with such a decomposition *for all* possible measurements cannot violate any Bell inequality; otherwise it does violate (at least) one Bell inequality.

A concept which is easier to handle and necessary for Bell nonlocality is the concept of quantum steering [25]. The steering scenario is an asymmetric scenario where one or more parties remotely steer the state of the remaining parties by performing measurements on their part of the state. Here, we focus on the so-called one-sided steering scenario where Alice tries to steer Bob and Charlie. We say a state ρ demonstrates steering if its probability distribution does not admit a decomposition of the form

$$p(abc|xyz) = \int \pi(\lambda) p(a|x\lambda) \mathrm{Tr}(M_{b|y} \sigma_\lambda^B) \mathrm{Tr}(M_{c|z} \sigma_\lambda^C) d\lambda. \quad (3)$$

That is, the statistics can be explained by a so-called local hidden-state model (LHS), where the local response functions of Bob and Charlie are obtained from measurements on predetermined quantum states σ_λ^B and σ_λ^C, respectively. The set of (unnormalized) conditional states $\{\sigma_{a|x}^{BC}\}$ that Alice can prepare for Bob and Charlie, the so-called assemblage, is given by

$$\sigma_{a|x}^{BC} = \mathrm{Tr}_A[(M_{a|x} \otimes \mathbb{1} \otimes \mathbb{1})\rho], \quad (4)$$

where Tr_A denotes the partial trace and $\mathrm{Tr}(\sigma_{a|x}^{BC}) = p(a|x)$ is the probability that Alice obtains outcome a. Here, the measurement sets of Bob and Charlie $\{M_{b|y}\}$ and $\{M_{c|z}\}$ are assumed as tomographically complete. Further, note that any LHS can be considered as an LHV, while the converse does not hold [26]. An assemblage is said to demonstrate steering if it does not admit the decomposition

$$\sigma_{a|x}^{BC} = \int \pi(\lambda) p(a|x\lambda) \rho_\lambda^{BC}, \quad (5)$$

here ρ_λ^{BC} is a separable quantum state shared by Bob and Charlie.

We present now the hidden nonlocality scenario in the spirit of [10]. In this scenario the parties perform a probabilistic preselection according to a desired outcome before the Bell test. Hence, they apply a sequence of measurements on the shared state ρ_L which can lead to nonlocal correlations even if ρ_L admits an LHV for POVMs. In particular, this idea can be implemented by the use of local filters given by arbitrary Kraus operators F_x, fulfilling $F_x^\dagger F_x \leq \mathbb{1}$, $x \in \{A, B, C\}$ and acting on the respective local Hilbert space of the observers. The state which the parties share after filtering is given by

$$\rho = \frac{F_A \otimes F_B \otimes F_C \rho_L F_A^\dagger \otimes F_B^\dagger \otimes F_C^\dagger}{\mathrm{Tr}(F_A \otimes F_B \otimes F_C \rho_L F_A^\dagger \otimes F_B^\dagger \otimes F_C^\dagger)}, \quad (6)$$

where the success probability of the filtering is given by the normalization factor. We say that a state ρ_L possesses genuine hidden nonlocality if it admits an LHV for POVMs but the state ρ for some judiciously chosen filters F_A, F_B, F_C violates a Bell inequality. Note that local invertible filters do not change the entanglement character of a given state [3], i.e., bound entangled states remain bound entangled. Nevertheless the filters can increase the amount of entanglement (probabilistically) between the parties [27], which gives an intuitive reason why local filters can be useful. Further, by bound entangled states we mean entangled states with positive partial transpose (PPT).

Methods.—In order to derive our results, we will solve two main tasks: we show that the filtered state does violate a Bell inequality and that the state before filtering admits a local model for POVMs. The first task can be solved efficiently by an iterative sequence of semidefinite programs (SDPs) [28], using the so-called seesaw [29] method.

Consider a Bell inequality of the form

$$I = \sum_{a,b,c,x,y,z} c_{abc|xyz} p(abc|xyz) \leq L, \tag{7}$$

with given (real) coefficients $c_{abc|xyz}$ and a local bound L. The Bell operator according to this inequality is then given by

$$\mathcal{B} = \sum_{a,b,c,x,y,z} c_{abc|xyz} M_{a|x} \otimes M_{b|y} \otimes M_{c|z}. \tag{8}$$

The goal is to maximize the quantum value $Q = \mathrm{Tr}(\mathcal{B}\rho)$ for PPT entangled states ρ. Optimizing such an expression over all local measurements and the state is a problem, which cannot be solved by an SDP in general. However, the seesaw method provides a solution: we fix the measurements for two of the parties for a given state ρ, such that the problem becomes linear in the remaining party, let us say Alice. We maximize the expression Q subject to the constraints $M_{a|x} \geq 0$, $\sum_a M_{a|x} = \mathbb{1}$, which leads us to the optimal measurements of Alice. This strategy is iteratively applied over the individual parties and the state, to optimize the quantum value Q, without being guaranteed that it is a global maximum.

The second task is more difficult to solve. Even though there exist analytical constructions for LHVs, they mostly restrict to certain classes of states with high symmetry or they are restricted to projective measurements. Recently in [30,31] a method was presented to algorithmically construct local models, again making use of SDPs. Here, we only point out the main use of this construction (for details see [30,31]). Consider a discrete set of measurements $\{M_{a|x}\}$ associated with a so-called shrinking factor $0 \leq \eta \leq 1$ and the target state ρ_L. Further, consider the following SDP:

$$\begin{aligned}
&\text{given } \rho_L, \{M_{a|x}\}, \eta \\
&\text{find } q^* = \max q \\
&\text{s.t. } \mathrm{Tr}_A[(M_{a|x} \otimes \mathbb{1} \otimes \mathbb{1})\chi] = \sum_\lambda D_\lambda(a|x)\sigma_\lambda^{BC}, \quad \forall\ a,x \\
&\qquad \sigma_\lambda^{BC} \geq 0, (\sigma_\lambda^{BC})^{T_B} \geq 0 \quad \forall\ \lambda \\
&\qquad \eta\chi + (1-\eta)\frac{\mathbb{1}}{d_A} \otimes \mathrm{Tr}_A(\chi) = q\rho_L + (1-q)\frac{\mathbb{1}}{d_A d_B d_C},
\end{aligned} \tag{9}$$

where the Hermitian matrices χ and σ_λ^{BC} are the optimization variables. The SDP can be understood as follows. The first constraint ensures that (not necessarily positive-semidefinite quasistate) χ does admit an LHS for the finite set of measurements $\{M_{a|x}\}$, where $D_\lambda(a|x)$ are the deterministic strategies corresponding to Alice's set of inputs and outputs. More specifically, $D_\lambda(a|x) = \delta_{a,\lambda_x}$, where $\lambda = \lambda_1\lambda_2\cdots\lambda_{m_A}$ is a string of length m_A, where m_A is the number of Alice's settings. The (subnormalized) states σ_λ^{BC} have to be separable between Bob and Charlie which is in general a nontrivial task, but for two qubits can simply be enforced by the partial transpose constraint $(\sigma_\lambda^{BC})^{T_B} \geq 0$ [32]. The last constraint contains the shrinking factor $0 \leq \eta \leq 1$ and ensures that also a noisy version of the target state ρ_L admits an LHS, but this time for the continuous set of measurements $\mathcal{M}$ (e.g., four-outcome POVMs) which was approximated by the discrete set $\{M_{a|x}\} \subset \mathcal{M}$.

The SDP is based on the fact that the statistics from noisy measurements on a noiseless state are equal to the statistics of a noisy state with noiseless measurements, i.e.,

$$\mathrm{Tr}_A[(M_a^\eta \otimes \mathbb{1} \otimes \mathbb{1})\chi] = \mathrm{Tr}_A[(M_a \otimes \mathbb{1} \otimes \mathbb{1})\rho_L], \tag{10}$$

where the target state is defined by

$$\rho_L = \eta\chi + (1-\eta)\frac{\mathbb{1}}{d_A} \otimes \mathrm{Tr}_A(\chi), \tag{11}$$

and the noisy measurements are given by

$$M_a^\eta = \eta M_a + (1-\eta)\mathrm{Tr}(M_a)\frac{\mathbb{1}}{d_A}, \tag{12}$$

for any $M_a \in \mathcal{M}$.

Note that because χ admits an LHS for the discrete set $\{M_{a|x}\}$, by convexity it admits also a local model for the noisy measurements M_a^η. From the equality in (10) it follows that ρ_L does also admit an LHS for a set of continuous noiseless measurements.

Here, the shrinking factor η is the largest number such that all noisy measurements M_a^η can be written as a convex mixture of elements from the discrete set $\{M_{a|x}\}$, i.e.,

$$M_a^\eta = \sum_x p_x M_{a|x}, \tag{13}$$

with $\sum_x p_x = 1$ and $p_x \geq 0\ \forall\ x$.

The shrinking factor can only be obtained analytically in the case of qubit projective measurements, but for general measurements it can be obtained by an SDP [31].

Results.—We now display our main result by first presenting a nonlocal three-qubit bound entangled state and in a second step show that this state originates from local filtering of a different state with an LHS model for POVMs. Note that the following results were recovered from the numerical data and are therefore exact in an analytical sense, unless indicated differently. Consider the (real-valued) density matrix in the basis $\{|000\rangle, |001\rangle, |010\rangle, \ldots, |111\rangle\}_{ABC}$ given by

$$\rho_{\mathrm{NL}} = (r_{ij})_{1\leq i,j\leq 8}, \tag{14}$$

with the following defining entries:

$$
\begin{aligned}
&r_{11} = 0.0290, \qquad r_{12} = r_{13} = r_{15} = -0.0098,\\
&r_{14} = r_{16} = r_{17} = r_{23} = r_{25} = r_{35} = -0.0083,\\
&r_{18} = r_{27} = r_{36} = r_{45} = 0.0646,\\
&r_{22} = r_{33} = r_{55} = 0.0412,\\
&r_{24} = r_{26} = r_{34} = r_{37} = r_{56} = r_{57} = -0.0335,\\
&r_{28} = r_{38} = r_{46} = r_{47} = r_{58} = r_{67} = -0.0598,\\
&r_{44} = r_{66} = r_{77} = 0.1352,\\
&r_{48} = r_{68} = r_{78} = 0.0102, \qquad r_{88} = 0.4418.
\end{aligned}
$$

Note that ρ_{NL} is invariant under partial transpose with respect to any party, as well as invariant under permutation of parties, by construction. Therefore, the state is PPT and also biseparable with respect to any bipartite cut [23,33]. Note further that ρ_{NL} has the same symmetry properties as the family of states in [23] without being a member of this family. Nevertheless, using the seesaw method it can be shown to violate Śliwa's inequality number 5 [34] (which implies ρ_{NL} is entangled), which reads

$$
\begin{aligned}
I = \langle \mathrm{sym}[&A_1 + A_1B_2 - A_2B_2 - A_1B_1C_1\\
&- A_2B_1C_1 + A_2B_2C_2]\rangle \leq 3,
\end{aligned}
\tag{15}
$$

where $\mathrm{sym}[X]$ denotes the symmetrization of X over the three parties, e.g., $\mathrm{sym}[A_1B_2] = A_1B_2 + A_1C_2 + A_2B_1 + A_2C_1 + B_1C_2 + B_2C_1$. Here, $A_j = B_j = C_j, j \in \{1,2\}$, and $A_j = M_{1|j} - M_{2|j}$. We choose $A_1 = -0.7909\sigma_z - 0.6119\sigma_x$, $A_2 = -0.2344\sigma_z + 0.9721\sigma_x$, which leads to a quantum violation $Q \approx 3.0152 > 3$ of inequality (15). Note that the maximal quantum value achievable by PPT states only allows violations up to $Q \approx 3.0187$ [35].

Next, we show that ρ_{NL} can originate from a local state by filtering. Consider the state ρ_L defined via the relation

$$
\rho_{\mathrm{NL}} = \frac{F_A \otimes F_B \otimes F_C \rho_L F_A^\dagger \otimes F_B^\dagger \otimes F_C^\dagger}{\mathrm{Tr}(F_A \otimes F_B \otimes F_C \rho_L F_A^\dagger \otimes F_B^\dagger \otimes F_C^\dagger)}, \tag{16}
$$

with the local filters

$$
F_A = \begin{bmatrix} 0.4310 & -0.2971 \\ -0.2488 & 0.7291 \end{bmatrix},
$$

$$
F_B = \begin{bmatrix} 0.0342 & -0.0808 \\ -0.3664 & 0.8688 \end{bmatrix},
$$

$$
F_C = \begin{bmatrix} 0.3268 & -0.1873 \\ -0.1773 & 0.6440 \end{bmatrix}.
$$

For more details, see the Supplemental Material [36]. Note that it is immediately clear that there exists a valid quantum state ρ_L fulfilling the above equation. This can be seen by using the fact that the above local filters are invertible and the only constraint $F^\dagger F \leq \mathbb{1}$ can always be achieved, since the filters F and cF map onto the same state for any $c \in \mathbb{C}\backslash\{0\}$.

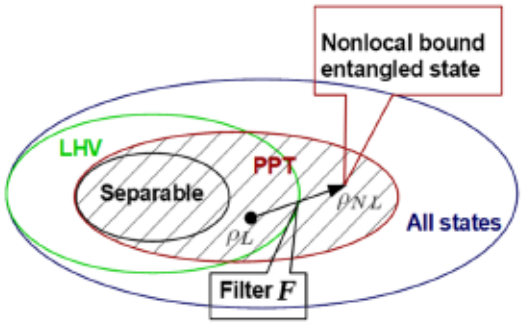

FIG. 2. Schematic overview over the relevant sets of states. The states in the shaded area are undistillable. Our results confirm the existence of bound entangled states with an LHV for POVMs. However, (invertible) local filters F are able to reveal the hidden nonlocality of these states. They map a state ρ_L from the set of states admitting an LHV onto a nonlocal state ρ_{NL}.

In order to finally show that ρ_L possesses genuine hidden nonlocality, we need to show that it admits a local model for all POVMs. Therefore, we use the same parametrization as in [31] for Alice's finite set of measurements $\{M_{a|x}\}$. It consists of all relabellings of $\{P_+, P_-, 0, 0\}$ where P_+ is a projector onto a vertex of an icosahedron in the Bloch sphere and P_- onto the opposite direction, as well as all relabellings of the trivial set $\{1, 0, 0, 0\}$. This leads to a set of 76 elements with a shrinking factor of $\eta \approx 0.673$. Note that it is sufficient to consider only extremal POVMs, which for qubits have at most four outcomes [38]. The optimization for the LHS, according to (9) results in $q^* = 1$. The precision of this result is subject to the standard precision of MATLAB [39] as well as the SDP solvers SeDuMi [40] and Mosek [41] for Yalmip [42]. Hence, ρ_L admits a local model for POVMs without the need of additional noise. For a graphical illustration of our main results, see Fig. 2.

Conclusions and outlook.—In the present Letter, we have shown that a fully biseparable bound entangled state of three qubits can admit a local model for POVMs, but can give rise to nonlocal correlations when local filters were applied before the Bell test. Hence, we have shown that bound entangled states can possess genuine hidden nonlocality. This is the first example of activation of nonlocality in bound entanglement. Furthermore, this is also the first example of an LHV of a bound entangled state for all POVMs, while previous models were restricted to projective measurements [31,43]. One important conclusion of our results is that genuine hidden nonlocality (since it also exists for bound entangled states) does not imply entanglement distillability. Together with the result of [11] it shows that genuine hidden nonlocality and entanglement

distillation are inequivalent. Note that since the local model we have constructed is an LHS model, our results are also relevant for the steering scenario.

It would be interesting to know whether there exist also bound entangled states without hidden nonlocality. Even though we could not prove the existence of such states, we found a bipartite bound entangled state with a local model for POVMs in the so-called filter normal form [27], which seems to play an important role for hidden nonlocality. We think, therefore, that this state is a good candidate to show bound entanglement without hidden nonlocality. For further details, see the Supplemental Material [36]. In the future, one should investigate the potential of bound entangled states in the superactivation or even in the asymptotic scenario. Even 20 years after the Peres conjecture [21], we still learn what bound entangled states are useful for. In the spirit of these developments it seems to be well motivated to state an "inverse Peres conjecture": all bound entangled states are nonlocal resources in the asymptotic case, see Fig. 1.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 675662 and support from the Federal Ministry of Education and Research (BMBF).

*Lucas.Tendick@hhu.de

[1] J. S. Bell, On the Einstein-Podolsky-Rosen paradox, Physics **1**, 195 (1964).
[2] N. Brunner, D. Cavalcanti, S. Pironio, V. Scarani, and S. Wehner, Bell nonlocality, Rev. Mod. Phys. **86**, 419 (2014).
[3] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, Quantum entanglement, Rev. Mod. Phys. **81**, 865 (2009).
[4] A. Einstein, B. Podolsky, and N. Rosen, Can quantum-mechanical description of physical reality be considered complete?, Phys. Rev. **47**, 777 (1935).
[5] N. Gisin, Bell's inequality holds for all non-product states, Phys. Lett. **154A**, 201 (1991).
[6] S. Popescu and D. Rohrlich, Generic quantum nonlocality, Phys. Lett. **166A**, 293 (1992).
[7] R. F. Werner, Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model, Phys. Rev. A **40**, 4277 (1989).
[8] J. Barrett, Nonsequential positive-operator-valued measurements on entangled mixed states do not always violate a Bell inequality, Phys. Rev. A **65**, 042302 (2002).
[9] S. Popescu, Bell's Inequalities and Density Matrices: Revealing "Hidden" Nonlocality, Phys. Rev. Lett. **74**, 2619 (1995).
[10] F. Hirsch, M. Quintino, J. Bowles, and N. Brunner, Genuine Hidden Quantum Nonlocality, Phys. Rev. Lett. **111**, 160402 (2013).
[11] F. Hirsch, M. Quintino, J. Bowles, T. Vértesi, and N. Brunner, Entanglement without hidden nonlocality, New J. Phys. **18**, 113019 (2016).
[12] C. Palazuelos, Superactivation of Quantum Nonlocality, Phys. Rev. Lett. **109**, 190401 (2012).
[13] D. Cavalcanti, A. Acín, N. Brunner, and T. Vértesi, All quantum states useful for teleportation are nonlocal resources, Phys. Rev. A **87**, 042104 (2013).
[14] L. Masanes, Asymptotic Violation of Bell Inequalities and Distillability, Phys. Rev. Lett. **97**, 050503 (2006).
[15] M. Horodecki, P. Horodecki, and R. Horodecki, Mixed-State Entanglement and Distillation: Is There a Bound Entanglement in Nature?, Phys. Rev. Lett. **80**, 5239 (1998).
[16] P. Horodecki, M. Horodecki, and R. Horodecki, Bound Entanglement Can Be Activated, Phys. Rev. Lett. **82**, 1056 (1999).
[17] P. W. Shor, J. A. Smolin, and A. V. Thapliyal, Superactivation of Bound Entanglement, Phys. Rev. Lett. **90**, 107901 (2003).
[18] J. T. Barreiro, P. Schindler, O. Gühne, T. Monz, M. Chwalla, C. F. Roos, M. Hennrich, and R. Blatt, Experimental multiparticle entanglement dynamics induced by decoherence, Nat. Phys. **6**, 943 (2010).
[19] J. Zhang, Continuous-variable multipartite unlockable bound entangled Gaussian states, Phys. Rev. A **83**, 052327 (2011).
[20] X. Jia, J. Zhang, Y. Wang, Y. Zhao, C. Xie, and K. Peng, Superactivation of Multipartite Unlockable Bound Entanglement, Phys. Rev. Lett. **108**, 190501 (2012).
[21] A. Peres, All the Bell inequalities, Found Phys. **29**, 589 (1999).
[22] T. Vértesi and N. Brunner, Disproving the Peres conjecture: Bell nonlocality from bipartite bound entanglement, Nat. Commun. **5**, 5297 (2014).
[23] T. Vértesi and N. Brunner, Quantum Nonlocality Does Not Imply Entanglement Distillability, Phys. Rev. Lett. **108**, 030403 (2012).
[24] Y. C. Liang, L. Masanes, and D. Rosset, All entangled states display some hidden nonlocality, Phys. Rev. A **86**, 052115 (2012).
[25] D. Cavalcanti and P. Skrzypczyk, Quantum steering: A review with focus on semidefinite programming, Rep. Prog. Phys. **80**, 024001 (2017).
[26] M. Quintino, T. Vértesi, D. Cavalcanti, R. Augusiak, M. Demianowicz, A. Acín, and N. Brunner, Inequivalence of entanglement, steering, and Bell nonlocality for general measurements, Phys. Rev. A **92**, 032107 (2015).
[27] F. Verstraete, J. Dehaene, and B. DeMoor, Local filtering operations on two qubits, Phys. Rev. A **64**, 010101 (2001).
[28] L. Vandenberghe and S. Boyd, Semidefinite programming, SIAM Rev. **38**, 49 (1996).
[29] R. F. Werner and M. M. Wolf, Bell inequalities and entanglement, arXiv:quant-ph/0107093.
[30] D. Cavalcanti, L. Guerini, R. Rabelo, and P. Skrzypczyk, General Method for Constructing Local Hidden Variable Models for Entangled Quantum States, Phys. Rev. Lett. **117**, 190401 (2016).
[31] F. Hirsch, M. Quintino, T. Vértesi, M.F. Pusey, and N. Brunner, Algorithmic Construction of Local Hidden Variable Models for Entangled Quantum States, Phys. Rev. Lett. **117**, 190402 (2016).
[32] A. Peres, Separability Criterion for Density Matrices, Phys. Rev. Lett. **77**, 1413 (1996).

[33] B. Kraus, J. I. Cirac, S. Karnas, and M. Lewenstein, Separability in $2 \times n$ composite quantum systems, Phys. Rev. A **61**, 062302 (2000).

[34] C. Śliwa, Symmetries of the Bell correlation inequalities, Phys. Lett. A **317**, 165 (2003).

[35] T. Moroder, J. D. Bancal, Y. C. Liang, M. Hofmann, and O. Gühne, Device-Independent Entanglement Quantification and Related Applications, Phys. Rev. Lett. **111**, 030501 (2013).

[36] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.050401 for more details on the local state in the main text, as well as for the local bipartite bound entangled state in the filter normal form. This includes Ref. [37].

[37] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri, Complete family of separability criteria, Phys. Rev. A **69**, 022308 (2004).

[38] G. M. D'Ariano, P. Lo Presti, and P. Perinotti, Classical randomness in quantum measurements, J. Phys. A **38**, 5979 (2005).

[39] MATLAB, Version 8.2.0.701 (R2013b) (The MathWorks Inc., Natick, Massachusetts, 2010).

[40] Jos F. Sturm, Using Sedumi 1.02, a MATLAB toolbox for optimization over symmetric cones, Optimi. Methods and Softw. **11**, 625 1999.

[41] MOSEK ApS, The MOSEK optimization toolbox for MATLAB manual, Version 8.0.0.60, 2019, https://docs.mosek.com/8.0/toolbox/index.html.

[42] J. Löfberg, Yalmip: A toolbox for modeling and optimization in MATLAB, in *Proceedings of the CACSD Conference, Taipei, Taiwan* (2004).

[43] G. Tóth and T. Vértesi, Quantum States with a Positive Partial Transpose are Useful for Metrology, Phys. Rev. Lett. **120**, 020506 (2018).

Supplemental Material for "Activation of nonlocality in bound entanglement "

September 21, 2019

Details on the local state ρ_L.—In order to give a useful representation of the local state ρ_L from (16) in the main text, one has to understand how to obtain this state. Naturally, there is no hint which states one should investigate in order to try to prove their locality or whether they possess genuine hidden nonlocality. However, it becomes immediately clear when one inverts the problem and tries to find a local state after we applied local filters on a nonlocal state. Since we choose the filters to be invertible, we can easily find filters which map the local state onto the nonlocal state. The nonlocal state obtained by the see-saw algorithm has by construction a high amount of symmetry, which we decrease by the local filters and then apply the SDP techniques to find an LHS. Afterwards, the inverted filters increase the symmetry of the state again. Therefore, ρ_L is simply given by

$$\rho_L = \frac{G_A \otimes G_B \otimes G_C \; \rho_{NL} \; G_A^\dagger \otimes G_B^\dagger \otimes G_C^\dagger}{\mathrm{Tr}(G_A \otimes G_B \otimes G_C \; \rho_{NL} \; G_A^\dagger \otimes G_B^\dagger \otimes G_C^\dagger)}, \tag{S1}$$

with the local invertible filters

$$G_A = \begin{bmatrix} 0.7291 & 0.2971 \\ 0.2488 & 0.4310 \end{bmatrix},$$
$$G_B = \begin{bmatrix} 0.8688 & 0.0808 \\ 0.3664 & 0.0342 \end{bmatrix},$$
$$G_C = \begin{bmatrix} 0.6440 & 0.1873 \\ 0.1773 & 0.3268 \end{bmatrix}.$$

and the nonlocal state ρ_{NL} defined in Eq. (14) in the main text.

Local bound entanglement in the filter normal form.—Here, we want to extend our outlook by presenting a bipartite bound entangled state which admits an LHS for POVMs and is a good candidate to show bound entanglement without hidden nonlocality, as we will argue below. An important feature of this state is that the state is already in the filter normal form [1], which means all single-party reduced density matricies are maximally mixed. The filter normal form does play an important role when it comes to hidden nonlocality. For example, the filter normal form does maximize the violation of the CHSH inequality for two-qubits, as well as entanglement monotones [1]. Further, in [2] it was shown that certain Werner states admit an LHS model, even after arbitrary local filtering. Werner states are also already in the filter normal form.

Intuitively, there is no obvious reason why local filters would still be able to activate the nonlocality of such states because they cannot distinguish the *useful* part of a state from white noise. Consider the state, in filter normal form given by

$$\sigma = \frac{\mathbb{1}}{d_A d_B} + \sum_{k=1}^{d_A^2-1} \xi_A H_k^A \otimes H_k^B \tag{S2}$$

with $d_A = 2$, $d_B = 4$, the coefficients ξ_k, and the traceless mutually orthonormal matricies H_k^A, H_k^B. Specifically, we choose

$$\xi_1 = \xi_2 = 1.3219,\ \xi_3 = 1.1348,$$

and the matricies

$$H_1^A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},\ H_2^A = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix},$$
$$H_3^A = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

for Alice's subsystem, as well as

$$H_1^B = \begin{pmatrix} 0 & 0 & 0 & -0.0983 \\ -0.6393 & 0 & 0 & 0 \\ 0 & -0.4158 & 0 & 0 \\ 0 & 0 & -0.6393 & 0 \end{pmatrix},$$
$$H_2^B = \begin{pmatrix} 0 & 0.6393 & 0 & 0 \\ 0 & 0 & 0.4158 & 0 \\ 0 & 0 & 0 & 0.6393 \\ 0.0983 & 0 & 0 & 0 \end{pmatrix},$$
$$H_3^B = \begin{pmatrix} -0.4859 & 0 & 0 & 0 \\ 0 & -0.5137 & 0 & 0 \\ 0 & 0 & 0.5137 & 0 \\ 0 & 0 & 0 & 0.4859 \end{pmatrix},$$

for Bob's side. As one can quickly verify, σ is a PPT state. Nevertheless, it can be shown to be entangled by the SDP techniques presented in [3]. With the methods described in the main text, we were able to show that σ does admit an LHS model for general POVMs on Alice's side.

As argued above, this state is a good candidate to show bound entanglement without hidden nonlocality. However, it is quite complicated to prove our conjecture, due to the fact that many degrees of freedom are involved. If our conjecture turns out to be true, other scenarios like the superactivation or the asymptotic scenario have to be considered. If it turns out that σ can show hidden nonlocality, it would be the first example of a nonlocal bound entangled state in the lowest possible dimension for two parties. So far the smallest dimension for examples of nonlocal bound entangled states is 3×3 [4].

References

[1] F. Verstraete, J. Dehaene, and B. DeMoor. Local filtering operations on two qubits. *Phys. Rev. A*, 64:010101, Jun 2001.

[2] F. Hirsch, M. Quintino, J. Bowles, T. Vértesi, and N. Brunner. Entanglement without hidden nonlocality. *New Journal of Physics*, 18(11):113019, 2016.

[3] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri. Complete family of separability criteria. *Phys. Rev. A*, 69:022308, Feb 2004.

[4] T. Vértesi and N. Brunner. Disproving the Peres conjecture: Bell nonlocality from bipartite bound entanglement. *Nature Communications 5, 5297 (2014)*, 2014.

B

Training variational quantum algorithms is NP-hard

Title:	Training variational quantum algorithms is NP-hard
Authors:	Lennart Bittel, Martin Kliesch
Journal:	Physical Review Letters
Date of submission:	24 April 2021
Publication status:	Published

This publication corresponds to the article [86]. The summary of the results is presented in section 4.1.1.
Contribution: My contribution was in deriving the main reductions and proofs of the paper. The paper was written in close discussion with MK.

C

Fast gradient estimation for variational quantum algorithms

Title:	Fast gradient estimation for variational quantum algorithms
Authors:	Lennart Bittel, Jens Watty, Martin Kliesch
Journal:	TBD

This publication corresponds to the article [87]. The summary of the results is presented in section 4.2.2.

Contribution: The work was inspired from numerical observations that resulted from JW's bachelor thesis. My main contributions were in deriving the estimation framework and proving most analytical results presented in the paper. I also helped in the numerical analysis, for which JW was mostly responsible.

Fast gradient estimation for variational quantum algorithms

Lennart Bittel[1], Jens N. Watty[1], and Martin Kliesch[1,2]

[1]Institute for Theoretical Physics, Heinrich Heine University Düsseldorf, Germany

[2]Institute for Quantum-Inspired and Quantum Optimization, Hamburg University of Technology, Germany

Many optimization methods for training variational quantum algorithms are based on estimating gradients of the cost function. Due to the statistical nature of quantum measurements, this estimation requires many circuit evaluations, which is a crucial bottleneck of the whole approach. We propose a new gradient estimation method to mitigate this measurement challenge and reduce the required measurement rounds. Within a Bayesian framework and based on the generalized parameter shift rule, we use prior information about the circuit to find an estimation strategy that minimizes expected statistical and systematic errors simultaneously. We demonstrate that this approach can significantly outperform traditional gradient estimation methods, reducing the required measurement rounds by up to an order of magnitude for a common QAOA setup. Our analysis also shows that an estimation via finite differences can outperform the parameter shift rule in terms of gradient accuracy for small and moderate measurement budgets.

1 Introduction

It has been demonstrated that quantum devices can outperform classical computers on computational problems specifically tailored to the hardware [1, 2]. While this has been an important milestone, the ultimate goal is a *useful quantum advantage*, i.e. a similar speedup for a problem with relevant applications. The central practical challenge is that only noisy and intermediate scale quantum (NISQ) hardware is available for the foreseeable future [3]. This restriction means that quantum devices have limited qubit numbers and can only run short quantum circuits, as the quantum computation must be finished before noise effects become too dominant. For this reason, great efforts are being made to design quantum algorithms in a NISQ-friendly way. One central idea in this effort is to trade an increased number of circuit evaluations and additional classical computation for reduced qubit numbers and lower circuit depths.

Lennart Bittel: lennart.bittel@uni-duesseldorf.de
Jens N. Watty: jens.schneider@uni-duesseldorf.de
Martin Kliesch: martin.kliesch@tuhh.de

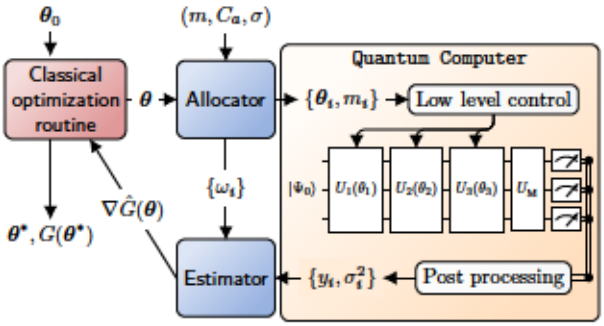

Figure 1: Sketch of the gradient estimation for VQA optimization routines as shown in algorithm 1.

One of the leading approaches toward achieving useful quantum advantages is given by variational quantum algorithms (VQAs). They address the problem of estimating the ground state energy of a quantum many-body Hamiltonian via a variational optimization, as follows. The quantum part of VQAs is implemented using parametrized quantum circuits (PQCs), which are used to prepare the variational quantum states. In order to interface it with a classical computer, the energy and the energy gradient w.r.t. the variational parameters are typically estimated. Then a classical computer, which typically runs some gradient descent-based algorithm, is used to minimize the energy via repeated parameter updates and estimations of the energy functional.

Several challenges occur in this approach. First, on the classical computation side, the optimization might reach a barren plateau for the objective function [4] or get stuck in local minima [5]. Barren plateaus can sometimes be avoided by using smart initializations for the parameters [6]. Moreover, sophisticated constructions of the quantum circuit family can help to bypass such problematic regions in the parameter space [7, 8]. Local minima can at least partially be avoided using natural gradients [9]. Second, the measurement effort of the quantum computer can pose a critical bottleneck for VQAs. The reason is that

(i) many iterations steps are done in the classical optimization,
(ii) several partial derivatives are needed for each gradient update step,
(iii) multiple measurement settings might be needed for the estimation of observables such as local

Hamiltonians,

(iv) quantum measurements are probabilistic, requiring $O(1/\epsilon^2)$ measurement rounds for ϵ accuracy

and this can add up to a large number of total number of measurements rounds. Since quantum measurements are destructive, one also needs to prepare the entire variational state from scratch for each measurement round.

In this work, we develop a new gradient estimation algorithm that balances statistical and systematic errors which achieve a better gradient estimate with fewer measurement rounds than conventional estimators. Specifically, we first characterize both the statistical and systematic error that arise in the estimation procedure, where for the systematic error, we introduce a Bayesian framework using prior information and assumptions about the system to estimate it. Then, we develop allocator methods, which for a given measurement budget determine an optimal strategy, namely what and how often we want to measure each circuit configuration, in order to minimize the total error. The estimator then returns a gradient estimate based on the measurement outcomes. A sketch of the procedure is shown in figure 1.

The Bayesian approach takes advantage, but also requires prior knowledge about the system. We develop strategies to obtain this prior information depending on the circuit depth:

(i) For short circuit, we use experimental or numerical/analytical observations.

(ii) For higher depths, we use unitary 2-design properties of random circuits.

(iii) In the intermediate regime, we use an interpolation of the two.

For the analysis in this paper we neglect all error sources arising from imperfect quantum hardware and only focus on noise due to finite measurements (i.e. shot noise). Additionally, we assume time periodic unitaries, meaning that, without loss of generality, the eigenenergies of the generators can be assumed to be integers.

Finally, we demonstrate numerically that the Bayesian approach outperforms previous parameter shift rule (PSR) approaches in terms of gradient estimation and VQA optimization accuracy.

1.1 Related work

There are several approaches to estimating gradients in VQAs, notably the PSR [10, 11] is able to obtain unbiased gradient estimates for generators with only two distinct eingenvalues. Further generalizations were made in Ref. [12] for a wider class of Hamiltonians. This approach often requires ancillary qubits or unitaries generated by commuting generators with two eigenvalues. There are also generalizations proposed for non-commuting generators [13] in a stochastic framework. These approaches generally require measurement settings not contained in the VQA-ansatz, but which can be assumed to be feasible for real hardware. There are also unbiased estimators for arbitrary periodic unitaries [14–17], where all measurements are contained in the ansatz class.

There are also strategies [18–21] that replace the actual observable underlying the gradient estimation with some surrogate observables. Another research direction has been to find efficient estimation schemes for the whole gradient [22, 23] instead of its individual partial derivatives and thus reducing the required measurement resources.

We approach the gradient estimation differently. Namely, we use that VQAs due to their setup experience some typical behavior, which can be analyzed in advance and during the VQA optimization. This allows us to also evaluate the performance of biased estimation strategies, which under very reasonable assumptions on measurement budget, can significantly outperform their unbiased counterparts. We use the general framework of periodic parametrized quantum gates [15] but believe that a similar Bayesian reasoning can also benefit other VQA ansatz classes and gradient estimation strategies. It should therefore not be regarded as a competitor to existing methods, but as a complementary strategy to further reduce the measurement effort of gradient estimation strategies.

1.2 Notation

We use the notation $[n] := \{1, \dots, n\}$. The Pauli matrices are denoted by X, Y and Z. An operator O acting on subsystem j of a larger quantum system is denoted by O_j, e.g. X_1 is the Pauli-X-matrix acting on subsystem 1. ℓ_p-norms including $p = 0$ are denoted by $\|\cdot\|_p$. We use several symbols that are summarized in section 8.

2 Variational quantum algorithms

In a VQA the goal is to find parameters $\boldsymbol{\theta}$ such that a cost function is minimized. In general, this cost function is given by

$$G(\boldsymbol{\theta}) := \langle \Psi_0 | \left[\prod_{\alpha=1}^{L} U_\alpha(\theta_\alpha) \right]^\dagger O \left[\prod_{\alpha=1}^{L} U_\alpha(\theta_\alpha) \right] | \Psi_0 \rangle \,, \tag{1}$$

where $|\Psi_0\rangle$ is the initial state, $U_\alpha(\theta_\alpha) = \mathrm{e}^{-\mathrm{i}\theta_\alpha H_\alpha}$ are the unitary gates generated by H_α and O is the observable encoding the optimization problem. In this work, we are considering unitaries that are T-periodic in θ_α (w.l.o.g. $T = 2\pi$), which implies that all eigenvalues of H_α are integers.

Estimating the gradient of $G(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ is an important task, as most optimization algorithms are gradient descent based and thus require an efficient ap-

proximation of the gradient. Our strategy estimates the gradient by the functions partial derivatives. For this it is convenient to define the cost function at a point shifted by a value of x in the parameter θ_l

$$\begin{aligned} F_l(x) &\coloneqq G(\boldsymbol{\theta} + x\,\boldsymbol{e}_l) \\ &= \langle \Psi' | U_l^\dagger(x) O' U_l(x) | \Psi' \rangle \,, \end{aligned} \tag{2}$$

where $\boldsymbol{e}_l$ is the l-th canonical basis vector and the other layers are absorbed into the observable as O' and the initial state as $|\Psi'\rangle$. Furthermore, we used that $U_l(\theta_l + x) = U_l(\theta_l)U_l(x)$. For later reference, the modified state and observable are

$$\begin{aligned} |\Psi'\rangle &= \left[\prod_{\alpha=1}^{l-1} U_\alpha(\theta_\alpha)\right] |\Psi_0\rangle \,, \text{ and} \\ O' &= \left[\prod_{\alpha=l}^{L} U_\alpha(\theta_\alpha)\right]^\dagger O \left[\prod_{\alpha=l}^{L} U_\alpha(\theta_\alpha)\right]. \end{aligned} \tag{3}$$

The evaluation at point $\boldsymbol{\theta}$ is therefore just $G(\boldsymbol{\theta}) = F_l(0)$ and $\frac{\mathrm{d}G(\boldsymbol{\theta})}{\mathrm{d}\theta_l} = F_l'(0)$. We will henceforth focus only on the estimation of a single partial derivative w.r.t. a parameter θ_l. In the interest of readability we write U, and F instead of U_l, and F_l, as well as $|\Psi\rangle$ and O instead of $|\Psi'\rangle$ and O'.

In this restricted view, we are now going to examine the structure of the function $F(x)$ more closely. The parametrized unitary that defines this function has the form

$$U(\theta) = \mathrm{e}^{-\mathrm{i}\theta H} = \sum_{i=1}^{n_\lambda} P_i \mathrm{e}^{-\mathrm{i}\lambda_i \theta} \,, \tag{4}$$

where H is a Hermitian generator and the P_i are the projectors onto the eigenspaces corresponding to the eigenvalues $\{\lambda_1, \ldots, \lambda_{n_\lambda}\}$ of H in ascending order.

Using this notation, we obtain

$$\begin{aligned} F(x) &= \langle \Psi | U^\dagger(x) O U(x) | \Psi \rangle \\ &= \sum_{i,j=1}^{n_\lambda} \mathrm{e}^{\mathrm{i}(\lambda_i - \lambda_j)x} \langle \Psi | P_i O P_j | \Psi \rangle \,, \end{aligned} \tag{5}$$

where each $c_{ij} \coloneqq \langle \Psi | P_i O P_j | \Psi \rangle$ is just a scalar. This lets us rewrite the function as

$$\begin{aligned} F(x) &= \sum_{i,j=1}^{n_\lambda} \mathrm{e}^{\mathrm{i}(\lambda_i - \lambda_j)x} c_{ij} = \sum_{k=1}^{n_\mu} c_k \mathrm{e}^{\mathrm{i}\mu_k x} + c_k^* \mathrm{e}^{-\mathrm{i}\mu_k x} \\ &= \sum_{k=1}^{n_\mu} \big(a_k \sin(\mu_k x) + b_k \cos(\mu_k x) \big) \,, \end{aligned} \tag{6}$$

where $\mu_k \in \{|\lambda_i - \lambda_j|\}$ are all possible eigenvalue differences of the generator H,

$$c_k = \sum_{i,j:\lambda_i - \lambda_j = \mu_k} c_{ij} \tag{7}$$

and $\boldsymbol{c} = \frac{\boldsymbol{b} + \mathrm{i}\boldsymbol{a}}{2}$ with $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n_\mu}$ are Fourier coefficients. For the total number of frequencies, it follows $n_\mu = |\{\mu_k\}| \leq \binom{n_\lambda}{2}$. The derivative at $x = 0$ is

$$\delta \coloneqq F'(0) = \sum_{k=1}^{n_\mu} \mu_k a_k \,. \tag{8}$$

For generators with two eigenvalues, where we can set w.l.o.g. $\mu_k \in \{0, \nu\}$, it has been shown that an unbiased estimate for the partial derivative can be obtained via

$$\delta = \nu \frac{F(\frac{\pi}{2\nu}) - F(-\frac{\pi}{2\nu})}{2} \,, \tag{9}$$

which is known as the PSR [12].

In essence, we are going to generalize this method. A helpful tool for this task is the antisymmetric projection

$$f(x) = \frac{F(x) - F(-x)}{2} = \sum_{k=1}^{n_\mu} a_k \sin(\mu_k x) \,. \tag{10}$$

We are only considering symmetric measurement schemes as symmetrizing an estimation method will not make the prediction worse [15]. As such, we refer only to the positive measurement positions x of $f(x)$, knowing that estimates of $F(+x)$ and $F(-x)$ are required to determine it. We will also omit the $\mu = 0$ frequency, since it does not affect the derivative. Additionally, $\nu \coloneqq \|\boldsymbol{\mu}\|_\infty$ refers to the spectral width of the generator, meaning $\boldsymbol{\mu} \subset [\nu]$, since the periodicity of $U(x)$ implies that the entries of $\boldsymbol{\mu}$ are positive integers.

3 Gradient estimation approach

The *allocator* decides the measurement resource allocation and how to generate the estimate. In particular, we use a symmetric linear estimator of the derivative which for a set of measurement positions $\boldsymbol{x} \in [0, \pi)^{n_x}$ and number of measurements for each position $\boldsymbol{m} \in \mathbb{N}^{n_x}$ returns a gradient estimate

$$\hat{\delta} = \sum_{i=1}^{n_x} w_i y_i \,, \tag{11}$$

where y_i are the empirical estimates of $f(x_i)$ using m_i measurement rounds. Since each x_i requires 2 measurement settings, the total number of settings is $2n_x$. In the following we develop strategies to find optimal $\boldsymbol{x}, \boldsymbol{m}$ and $\boldsymbol{w}$.

The error

$$\epsilon_{\text{tot}} \coloneqq \hat{\delta} - \delta \tag{12}$$

between our estimator guess $\hat{\delta}$ and the true derivative δ is an important metric that we are going to use as

a figure of merit for choosing our estimation parameters. In practice, imperfections of current quantum hardware and the lack of quantum error correction will cause a significant noise level when evaluating the cost function on the quantum device. However, even on a perfect device, the measurement process introduces a shot noise error as the estimates are determined by sampling from the underlying multinomial distribution. We denote the expectation value over the shot noise by $\langle\cdot\rangle_s$.

The expected error under shot noise can then be written as

$$\begin{aligned}\langle\epsilon_{\text{tot}}\rangle_s &= \sum_{t=1}^{n_x} w_t\langle y_t\rangle_s - \delta && (13)\\ &= \sum_{t=1}^{n_x} w_t\left(\sum_{k=1}^{n_\mu} a_k\sin(\mu_k x_t)\right) - \sum_{k=1}^{n_\mu}\mu_k a_k && (14)\\ &=: (S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})^T\boldsymbol{a}\,, && (15)\end{aligned}$$

where we have used the definition (Eq. (8)) of δ, the unbiased nature of the estimate $\langle y_t\rangle_s = f(x_t)$ and Eq. (10) for f. We also defined the matrix $S^{\boldsymbol{x}}$ with entries $S^{\boldsymbol{x}}_{kt} := \sin(\mu_k x_t)$.

For the mean squared error this means

$$\begin{aligned}\langle\epsilon^2_{\text{tot}}\rangle_s &= \left\langle\left(\sum_{t=1}^{n_x} w_t y_t - \delta\right)^2\right\rangle_s\\ &= \left[(S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})^T\boldsymbol{a}\right]^2 + \sum_t w_t^2\left(\langle y_t^2\rangle_s - \langle y_t\rangle_s^2\right)\\ &=: \epsilon^2_{\text{sys}} + \langle\epsilon^2_{\text{stat}}\rangle_s\,, && (16)\end{aligned}$$

where the first term describes the systematic error resulting from the method not accurately determining the derivative even for exact measurements and the second term describes the statistical error arising from measurement shot noise.

3.1 Estimating the statistical error

For the statistical error, each term $\langle y_t^2\rangle_s - \langle y_t\rangle_s^2$ is the variance for the measurement position x_t resulting from shot-noise errors. If the single shot variance at position x_t is σ_t^2, we find the expression

$$\langle y_t^2\rangle_s - \langle y_t\rangle_s^2 = \frac{\sigma_t^2}{m_t} \tag{17}$$

where m_t is the number of measurement rounds performed for x_t. For a fixed measurement budget $m = \sum_{t=1}^{n_x} m_t$, the optimal measurement allocation is given by

$$\langle\epsilon^2_{\text{stat}}\rangle_s = \min_{\boldsymbol{m}:\|\boldsymbol{m}\|_1=m}\sum_{t=1}^{n_x} w_t^2\frac{\sigma_t^2}{m_t} = \frac{\left(\sum_{t=1}^{n_x}|w_t|\sigma_t\right)^2}{m}\,,$$

with $m_t = m\frac{|w_i|\sigma_i}{\sum_{j=1}^{n_x}|w_j|\sigma_j}$. If one assumes constant shot noise $\sigma_t \equiv \sigma$ regardless of the measurement position which we will hence force do, this simplifies to

$$\langle\epsilon^2_{\text{stat}}\rangle_s = \frac{\sigma^2}{m}\|\boldsymbol{w}\|_1^2 \quad\text{with}\quad m_t = \frac{|w_t|}{\|\boldsymbol{w}\|_1}m\,. \tag{18}$$

While σ_t or σ are unknown a priori, it is generally possible to give a rough estimate of σ beforehand and to determine an estimate of σ_t after only a few measurement rounds are performed. For this reason, we assume that σ is a known quantity in the following sections.

3.2 Estimating the systematic error through a Bayesian approach

Determining the systematic error is more challenging because it depends explicitly on the Fourier coefficients $\boldsymbol{a}$ which are not known. For our analysis we assume that the estimator will be used for an ensemble of multiple different positions $\boldsymbol{\theta}$, as one expects to occur during a full gradient descent optimization routine. Therefore, instead of finding the minimum for a particular instance, we want to find a strategy where the average total error over the entire ensemble is minimized. The benefit of this approach is that the average only requires knowledge of the general behavior of the Fourier coefficients, not specific values of the particular realization.

Formally, we assume that a distribution $\mathcal{D}_{\boldsymbol{\theta}}$ over the relevant positions $\boldsymbol{\theta}$ induces a distribution of the Fourier coefficients $\boldsymbol{a}$. One natural distribution $\mathcal{D}_{\boldsymbol{\theta}}$ is the uniform distribution over all parameter points $\boldsymbol{\theta}\in[-\pi,\pi)^L$, which can be motivated as a model for the case of random initialization $\boldsymbol{\theta}_0$. For an expectation value over the distribution $\mathcal{D}_{\boldsymbol{\theta}}$, we write $\langle\cdot\rangle_{\boldsymbol{\theta}}$. In this framework, taking the expectation value for $\mathcal{D}_{\boldsymbol{\theta}}$ yields

$$\begin{aligned}\langle\epsilon^2_{\text{sys}}\rangle_{\boldsymbol{\theta}} &= \left\langle\left[(S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})^T\boldsymbol{a}\right]^2\right\rangle_{\boldsymbol{\theta}}\\ &= (S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})^T\langle\boldsymbol{a}\boldsymbol{a}^T\rangle_{\boldsymbol{\theta}}(S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})\\ &=: (S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})^T C_a(S^{\boldsymbol{x}}\boldsymbol{w}-\boldsymbol{\mu})\,,\end{aligned} \tag{19}$$

meaning that the estimation of the expected squared systematic error requires knowledge of the second moment matrix

$$C_a := \langle\boldsymbol{a}\boldsymbol{a}^T\rangle_{\boldsymbol{\theta}} \in \mathbb{R}^{n_\mu\times n_\mu}\,. \tag{20}$$

In the following, we derive properties of the of C_a assuming the uniform distribution over $\boldsymbol{\theta}$.

First, we note that for a shift $\boldsymbol{\theta}\to\boldsymbol{\theta}+\boldsymbol{e}_l z$ in the layer l, the complex Fourier coefficients transform as $c_k \to c_k e^{\mathrm{i}\mu_k z}$, see Eq. (6). As $\mathcal{D}_{\boldsymbol{\theta}}$ is invariant under such a shift, we have

$$\langle c_k\rangle_{\boldsymbol{\theta}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\langle c_k\rangle_{\boldsymbol{\theta}}\mathrm{d}z = \frac{1}{2\pi}\int_{-\pi}^{\pi}\langle c_k e^{\mathrm{i}\mu_k z}\rangle_{\boldsymbol{\theta}}\mathrm{d}z = 0$$

for all $\mu_k \neq 0$, implying that c_k is centered around $c_k = 0$ in expectation over $\boldsymbol{\theta}$. Similarly for the second

moment, we compute

$$\langle c_k c_p\rangle_{\boldsymbol{\theta}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\langle c_k c_p\rangle_{\boldsymbol{\theta}} e^{\mathrm{i}(\mu_k+\mu_p)z}\mathrm{d}z = 0\,,$$
$$\langle c_k^* c_p\rangle_{\boldsymbol{\theta}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\langle c_k^* c_p\rangle_{\boldsymbol{\theta}} e^{\mathrm{i}(\mu_k-\mu_p)z}\mathrm{d}z = \langle |c_k|^2\rangle_{\boldsymbol{\theta}}\delta_{kp}\,,$$

where δ_{kp} is the Kronecker delta.

From the Fourier expansion Eq. (6) and $a_k = 2\,\mathrm{Im}(c_k)$, it follows that

$$\begin{aligned}\langle a_k\rangle_{\boldsymbol{\theta}} &= 0\\ \langle a_k a_p\rangle_{\boldsymbol{\theta}} &= 2\,\delta_{kp}\langle |c_k|^2\rangle_{\boldsymbol{\theta}}\,,\end{aligned} \tag{21}$$

meaning that C_a is a diagonal matrix with entries $\langle a_k^2\rangle_{\boldsymbol{\theta}} = 2\langle |c_k|^2\rangle_{\boldsymbol{\theta}}$ given by the expected squares of the Fourier coefficients.

What remains is determining $\langle a_k^2\rangle_{\boldsymbol{\theta}}$. It is worth pointing out that while underestimating $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ can lead to suboptimal results, even significantly overestimating the amplitudes will still outperform methods, where no prior assumptions are made, meaning rough estimates of $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ are already sufficient for good performance. One way of estimating them is to use already existing empirical measurement data from previous optimization rounds or initial calibration. The coefficients can be estimated using a Fourier fit. Another strategy involves numerically simulating smaller system sizes and extrapolating to the actual size used in the VQA.

If the applied unitaries in the VQA are known, $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ can sometimes be derived theoretically. For instance in appendix D.2, we derive analytically exact results for a VQA with a single layer ($L = 1$). For a small constant circuit depth, $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ can be computed efficiently, using Monte-Carlo sampling algorithms, even for large system sizes. This is convenient, as the case for deep circuits, under certain assumptions, $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ can be approximated again using only the spectral composition of the generator. This is shown in the following.

Ergodic limit – barren plateaus

VQA optimization routines have to overcome a general phenomenon known as *barren plateaus*. This term describes the tendency of a gradient in VQAs to be exponentially suppressed in the system size with increasing circuit depth and for almost all parameters $\boldsymbol{\theta}$. This phenomenon has been extensively studied and while mitigation techniques have been proposed [6–8], it appears to be unavoidable, at least in the general setup.

For the rigorous analysis of barren plateaus, it is beneficial to use the language of unitary t-designs. Effectively, with increasing circuit depth, the overall applied gate will appear more and more like a Haar-random unitary, with respect to which the derivative is suppressed by the Hilbert space dimension. This assumes that the underlying generators describe a universal gate set. For this effect to occur, we do not need convergence to the Haar measure but convergence to a unitary 2-design, which is significantly quicker [24, 25]. For our purposes, such approximate unitary 2-designs are sufficient since $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ can be expressed as a polynomial in $U, U^\dagger$ of degree $(2,2)$. If this condition is met, we can replace the expectation value over all angles by the expectation value over all unitaries. Hence, this condition can be summarized by the *ergodic assumption*

$$\langle \Gamma(U(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta}} = \int \Gamma(U)\mathrm{d}U\,, \tag{22}$$

which holds for any polynomial $\Gamma(U)$ of degree at most $(2,2)$ and where the integral is taken w.r.t. the Haar probability measure on the unitary group.

Under this assumption we derive in appendix C that

$$\langle a_k^2\rangle_{\boldsymbol{\theta}} = \xi_d\frac{\sigma_O^2}{d}\sum_{i\geq j:\mu_k=\lambda_i-\lambda_j}\mathrm{Tr}[P_i/d]\,\mathrm{Tr}[P_j/d] \tag{23}$$

with $\sigma_O^2 := \mathrm{Tr}[O^2/d] - \mathrm{Tr}[O/d]^2$, where d is the Hilbert space dimension and ξ_d is a constant close to 1 that depends only on d. $\mathrm{Tr}[P_i/d]$ is the relative multiplicity of the eigenvalue λ_i. Notably, the factor of $\frac{1}{d}$ shows the exponential suppression of the derivative in the system size, meaning that gate sets drawn from a 2-design experience barren plateaus. For $L\to\infty$ this result confirms the assumption that the relative multiplicity of an eigenvalue difference μ_k in the spectrum of the generator determines the expected size of its respective Fourier coefficient. We will analyze the strengths and limitations of this approach with an example in section 5.1.

4 Allocation methods

In this section, we derive several allocation methods for a given measurement budget. A Python implementation of these methods is available on GitHub [26].

The estimation algorithm requires the values $\boldsymbol{w}, \boldsymbol{m}$ and $\boldsymbol{x}$. We have already seen that making assumptions about the ensemble of configurations allows us to estimate the error using the second-moment matrix C_a from Eq. (20) and an a priori shot noise estimate σ^2. In the following, we devise explicit measurement procedures by making use of the knowledge of these quantities. In section 4.1, we show that using convex optimization procedures, one can derive an optimal measurement strategy, which we call Bayesian linear gradient estimator (BLGE).

In section 4.2, we then consider the case where the number of total measurements goes to infinity. We call this method unbiased linear gradient estimator (ULGE), as it does not require access to the estimates of C_a and σ^2 and yields an estimate with-

out systematic error. ULGE is an equivalent formulation of a known method in literature [15]. In section 4.2.1, we show strong similarity between ULGE and another popular generalized PSR found in the literature [12]. Finally, in section 4.3, we restrict the number of measurement positions n_x to 1 and derive a strategy that is optimal under this constraint. We call this method single Bayesian linear gradient estimator (SLGE). We note that this solution coincides with the result obtained for the non-restricted problem in the case where only very few total measurements are available for the gradient estimation.

4.1 Bayesian linear gradient estimator

For the linear estimator of a partial derivative we want to minimize the expected squared error $\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}$ by finding suitable measurement positions $\boldsymbol{x} \in [0,\pi)^{n_x}$ with weights $\boldsymbol{w} \in \mathbb{R}^{n_x}$ of our linear estimator, i.e. we wish to find the optimal solution of

$$(\boldsymbol{w}^*, \boldsymbol{x}^*) = \arg\min_{\boldsymbol{w},\boldsymbol{x}} \left\{ \langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{x}) \right\} \tag{24}$$

with

$$\begin{aligned}\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}} &= (S^{\boldsymbol{x}}\boldsymbol{w} - \boldsymbol{\mu})^T C_a (S^{\boldsymbol{x}}\boldsymbol{w} - \boldsymbol{\mu}) + \frac{\sigma^2}{m}\|\boldsymbol{w}\|_1^2 \\ &= \sum_{k=1}^{n_\mu} \langle a_k^2\rangle_{\boldsymbol{\theta}} \left(\sum_{t=1}^{n_x} w_t \sin(\mu_k x_t) \right)^2 + \frac{\sigma^2}{m}\|\boldsymbol{w}\|_1^2\end{aligned}$$

from Eq. (19) and Eq. (18). Since the cost function is non-convex in x, a direct approach may not reach the optimal solution. In appendix A.1 we show that instead, we can perform an equivalent maximization problem which arises as and effective dual problem after adding certain constraints:

$$\langle\epsilon^2_{\text{tot}}\rangle^*_{s,\boldsymbol{\theta}} = \max_{\boldsymbol{\kappa}\in\mathbb{R}^{n_\mu}} g(\boldsymbol{\kappa}) \tag{25}$$

with

$$g(\boldsymbol{\kappa}) = \sum_{k=1}^{n_\mu} \left(2\kappa_k\mu_k - \frac{\kappa_k^2}{\langle a_k^2\rangle_{\boldsymbol{\theta}}} \right) - \frac{m}{\sigma^2} \left\| \sum_{k=1}^{n_\mu} \kappa_k \sin(\mu_k(\,\cdot\,)) \right\|_\infty^2 ,$$

where the last term refers to the L^∞-space norm. Each dual variable $\{\kappa_k\}$ has the interpretation as the systematic error w.r.t. only one frequency component

$$\langle\epsilon^2_{\text{sys},k}\rangle_{\boldsymbol{\theta}} = \left((S^{\boldsymbol{x}}\boldsymbol{w})_k - \mu_k\right)^2 \langle a_k^2\rangle_{\boldsymbol{\theta}} = \frac{\kappa_k^2}{\langle a_k^2\rangle_{\boldsymbol{\theta}}} . \tag{26}$$

Due to complementary slackness between the primal and the dual problem, the global maxima positions of the function

$$\rho_{\boldsymbol{\kappa}}(x) = \left| \sum_{k=1}^{n_\mu} \kappa_k \sin(\mu_k x) \right| \tag{27}$$

are the set of optimal measurement positions $\boldsymbol{x}^*$. Having determined those, we can proceed determining the weights $\boldsymbol{w}^*$ by solving the convex problem Eq. (24) with the fixed positions $\boldsymbol{x}^*$ and obtain the measurement budget $\boldsymbol{m}$ via Eq. (18). We also allow for basic post-processing, where after the measurements are performed, we obtain updated weights $\boldsymbol{w}$ by replacing the statistical error in Eq. (24) by one using the empirically determined shot-noise variances.

$$\epsilon^2_{\text{stat}} = \sum_t w_t^2 \sigma^2_{\text{emp},t} , \tag{28}$$

where $\boldsymbol{\sigma}^2_{\text{emp}}$ refers to the empirical estimate of the variance. All the steps are shown in algorithm 1, also including the two other methods we consider in the following section.

Algorithm 1 Gradient estimation

Allocator and estimator procedure for the three outlined strategies. $\mathcal{M}_m(x)$ refers to performing physical measurement at position x using m measurement rounds and estimating the expectation value and variance.

procedure ALLOCATOR($\boldsymbol{\mu}$,C_a,σ,m)
  if Bayesian **then**
    $\boldsymbol{\kappa}^* \leftarrow \arg\max[g(\boldsymbol{\kappa})]$ ▷ (25)
    $\boldsymbol{x}^* \leftarrow \arg\max[\rho_{\boldsymbol{\kappa}^*}(\boldsymbol{x})]$ ▷ (27)
    $\boldsymbol{w}^* \leftarrow \arg\min\left[\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{x}^*)\right]$ ▷ (24)
  if Unbiased **then**
    $\boldsymbol{x}^* \leftarrow \frac{\pi}{\nu}([\nu] + \frac{1}{2})$
    $\boldsymbol{w}^* \leftarrow \frac{(-1)^i}{2\nu \sin^2(\frac{\pi}{2\nu}([\nu]+\frac{1}{2}))}$ ▷ (36)
  if Single **then**
    $x^* \leftarrow \arg\min\left[\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}(x)\right]$ ▷ (40)
    $w^* \leftarrow w(x^*)$
  $\boldsymbol{m} \leftarrow \text{ROUND}\left(m\frac{\boldsymbol{w}}{2\|\boldsymbol{w}\|_1}\right)$
  return $\boldsymbol{x}^*, \boldsymbol{w}^*, \boldsymbol{m}$
procedure ESTIMATOR($\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{m}$)
  for $i \in \{1,\dots,n_x\}$ **do**
    $y_{(\text{emp},t,+)}, \sigma^2_{(\text{emp},t,-)} \leftarrow \mathcal{M}_{m_i}(x_t)$
    $y_{(\text{emp},t,-)}, \sigma^2_{(\text{emp},t,-)} \leftarrow \mathcal{M}_{m_i}(-x_t)$
    $y_{(\text{emp},t)} \leftarrow \frac{y_{(\text{emp},i,+)} - y_{(\text{emp},i,-)}}{2}$
    $\sigma^2_{(\text{emp},t)} \leftarrow \frac{\sigma^2_{(\text{emp},i,+)} + \sigma^2_{(\text{emp},i,-)}}{2}$
  if postprocess **then**
    $\boldsymbol{w}^* \leftarrow \arg\min\left[\langle\epsilon^2_{\text{tot}}\rangle_{\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{x}^*)|_{\boldsymbol{\sigma}^2_{\text{emp}}}\right]$ ▷ (28)
  $\hat{\delta} \leftarrow \boldsymbol{y}^T_{\text{emp}}\boldsymbol{w}^*$;
  return $\hat{\delta}$

We also show that the number of measurement settings is restricted by the number of frequency differences in the generator spectrum.

Theorem 1. *For every optimization problem as defined in Eq. (24), there exists an optimal BLGE strategy that requires at most $n_x = n_\mu$ unique (positive)*

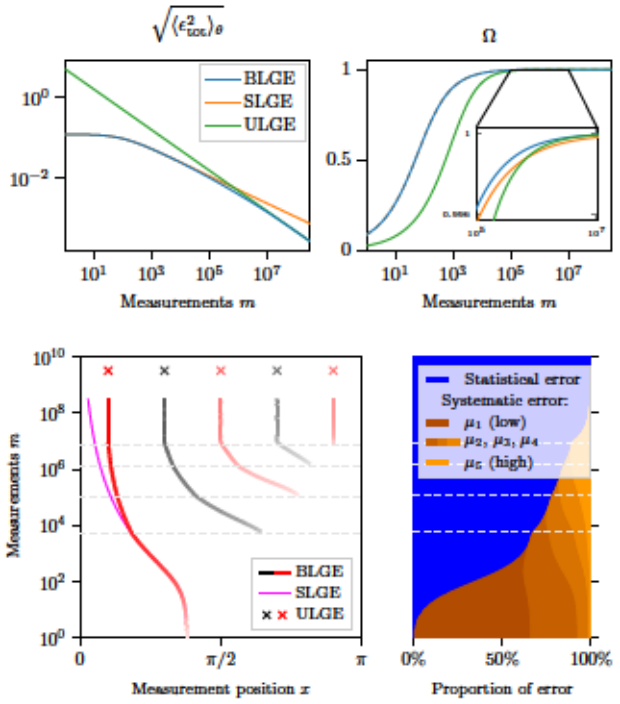

Figure 2: **Top left:** The theoretical mean error $\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}$ for different measurement budgets for all three methods. **Top right:** The theoretical relative correlation Ω for all methods. **Left:** The measurement positions chosen by the different methods for given total measurement budgets. The color of the lines indicates the sign of the coefficient w_i applied to the value (red - positive; black - negative); the fainter the color, the lower the absolute value of the coefficient w_i. ULGE is only shown once, since its allocation is independent of m. **Right:** The proportion of the total error that is due to the statistical error (blue) or the systematic error (shades of orange) in the case of the BLGE. Used values throughout: $\boldsymbol{\mu} = (1, 2, 3, 4, 5)$, $\sigma^2 = 1$, $\langle a_k^2\rangle_{\boldsymbol{\theta}} = 0.1 \times 10^{-\mu_k}$.

measurement positions. (Meaning $2n_\mu$ positions in total)

In appendix A.2 we show that this follows directly from the existence of a sparse solution for an ℓ_1-norm minimization with a linear constraint.

As a quality measure, we define the *relative correlation* Ω as

$$\Omega^2 := \frac{\langle\hat{\delta}\delta\rangle^2_{s,\boldsymbol{\theta}}}{\langle\delta^2\rangle_{\boldsymbol{\theta}}\langle\hat{\delta}^2\rangle_{s,\boldsymbol{\theta}}}\,, \tag{29}$$

which is the covariance between the real and the estimated derivative scaled onto the interval $\Omega \in [0, 1]$. Using the quadratic inequality we can find a relationship between the error and the relative correlation

$$\begin{aligned}\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}} &= \langle\delta^2\rangle_{s,\boldsymbol{\theta}} + \langle\hat{\delta}^2\rangle_{s,\boldsymbol{\theta}} - 2\langle\hat{\delta}\delta\rangle_{s,\boldsymbol{\theta}} \\ &\geq \langle\delta^2\rangle_{s,\boldsymbol{\theta}} - \frac{\langle\hat{\delta}\delta\rangle^2_{s,\boldsymbol{\theta}}}{\langle\hat{\delta}^2\rangle_{s,\boldsymbol{\theta}}} \\ &= \langle\delta^2\rangle_{s,\boldsymbol{\theta}}\left(1 - \Omega^2\right),\end{aligned} \tag{30}$$

where the lower bound is saturated by an optimal rescaling of $\boldsymbol{w}$. This allows us to think of Ω as an expected relative accuracy.

Figure 2 shows the result of this method for $n_\mu = 5$ and prior estimates $(\langle a_k^2\rangle_{\boldsymbol{\theta}}, \sigma^2)$ inspired by the model VQA defined in section 5, where small frequencies contribute significantly more to the overall gradient than large frequencies. We see that in the beginning, the error (blue line, top left) plateaus and only starts dropping once more than 10^3 measurement rounds are performed. This occurs as the method returns a very small guess to avoid being wrong, meaning the error is just the expected size of the derivative. For large measurement budgets, we have the expected $\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}} \propto \frac{1}{m}$ behavior. Similarly, for the relative correlation plot (top right), Ω is increasing from a value close to $\Omega = 0$ for very few measurement rounds to $\Omega = 1$ for $m \sim 10^4$ measurement rounds.

The bottom left plot show the measurement positions used in the BLGE. The y-axis denotes the measurement budget that the allocation method has available. We see that n_μ many positions only occur for very large measurement budgets (10^7), while only a single position (2 when also considering the negative measurement position) is returned for $m \leq 10^4$ measurements. For increasing m, the measurement positions move further to the left with new ones being added on the right.

On the bottom right plot, the composition of the error into statistical error (blue) as well as the different systematic errors based on the different frequency components (shades of orange), as defined in Eq. (26), are shown for different measurement budgets m. For few measurements the largest contribution to the systematic error error dominates with most of the error coming from the Fourier coefficient a_1, as this is the most significant component of f. As m increases, the statistical error starts becomes dominant while the systematic error, first the small frequencies, vanishes. This shows that as one might expect, the estimator becomes more unbiased, as the measurement budget increases.

In the next sections we analyze the behavior in the two limits, the central differences behavior in small m and the unbiased equidistant measurement strategy for large m.

4.2 Unbiased linear gradient estimator

If we let $m \to \infty$, the method will use all resources to make the systematic error vanish exactly, meaning that the remaining minimization of the statistical error in Eq. (24) simplifies to the convex optimization problem

$$\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}} = \frac{\sigma^2}{m} \min_{\boldsymbol{w}:\, S^{\boldsymbol{x}}\boldsymbol{w} = \boldsymbol{\mu}} \|\boldsymbol{w}\|_1^2\,, \tag{31}$$

where we recall $\boldsymbol{\mu} \subset [\nu]$ with ν the spectral width. This method is equivalent to one without a system-

atic error, which has been studied before in more depth [15]. Its behavior can be summarized in the following theorem.

Theorem 2. *For constant shot noise variance σ^2, any unbiased gradient estimation method given by Eq. (31) has an error $\langle\epsilon_{\text{tot}}^2\rangle_{s,\boldsymbol{\theta}} \geq \frac{\sigma^2\nu^2}{m}$ and $\Omega^2 \leq \frac{\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\langle\delta^2\rangle_{\boldsymbol{\theta}}+\frac{\nu^2\sigma^2}{m}}$, where tightness can be achieved with at most $n_{\boldsymbol{x}} = n_\mu$ measurement positions.*

Proof. As the method should be unbiased for all functions anti-symmetric function $f(x)$ with the allowed frequencies, we can choose $f(x) = \sin(\nu x)$ to find an upper-bound.:

$$\delta = \nu = \sum_i w_i \langle y_i\rangle_s = \sum_i w_i f(x_i)\,. \tag{32}$$

using $|f(x_i)| \leq 1$ and the triangle equality it follows that $\nu \leq \|\boldsymbol{w}\|_1$. This means that the error for m measurements is

$$\langle\epsilon_{\text{tot}}^2\rangle_{s,\boldsymbol{\theta}} = \frac{\sigma^2}{m}\|\boldsymbol{w}\|_1^2 \geq \frac{\sigma^2\nu^2}{m}\,. \tag{33}$$

Since there is no systematic error we also obtain

$$\begin{aligned}\Omega^2 &= \frac{\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\langle\delta^2\rangle_{\boldsymbol{\theta}} + \frac{\nu^2\sigma^2}{m}} \\ &\overset{m\to 0}{=} \frac{m\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\sigma^2} \times \frac{1}{\nu^2} + O(m^2)\,.\end{aligned} \tag{34}$$

What remains to show is that there exists a closed form solution for Eq. (31) which reaches this bound. For this we choose the measurement positions $x_i = \frac{\pi}{\nu}(i+\frac{1}{2})$ with $i \in \{0,\ldots,\nu-1\}$ yielding

$$S_{ik} = \sin\left(\frac{\pi}{\nu}\left(i+\frac{1}{2}\right)\mu_k\right). \tag{35}$$

This describes the discrete sine transform (DST-II), which by inversion yields $\boldsymbol{w} = S^{-1}\boldsymbol{\mu}$

$$w_i = \frac{(-1)^i}{2\nu\sin^2\left(\frac{\pi}{2\nu}\left(i+\frac{1}{2}\right)\right)}\,. \tag{36}$$

We note that these coefficients were already derived in literature [15] using Dirichlet kernels. One can verify that $\|\boldsymbol{w}\|_1 = \nu$, i.e. that our choice of coefficients saturates the lower bound we derived. We note that while this closed form solution requires ν many measurement positions, since it is in the limit of the general Bayesian method, there always exists a strategy with at most n_μ many positions as shown in theorem 1. □

As can be seen in figure 2, especially for small m, BLGE performs significantly better than ULGE for both the expected error and Ω. This is because ULGE, in order to be unbiased, is very dependent on shot noise, which leads to very significant statistical errors.

4.2.1 Comparison with PSR for sums of commuting 2-level generators

Even though the PSR in its simplest form is only valid for unitaries with two-level generators, it is straightforward to extend it to generators H which are the sum of commuting two-level generators H_i, i.e.

$$H = \sum_{i=1}^{n_\zeta} \zeta_i H_i\,, \tag{37}$$

where with without loss of generality the eigenvalues of all H_i are $\lambda \in \{0,1\}$ and n_ζ is the number of generators.

As was shown by Ref. [12], PSR (Eq. (9)) together with the product rule of differentiation yields an unbiased estimate of the derivative

$$\delta = \sum_{i=1}^{n_\zeta} \zeta_i \frac{F_{i|+\frac{\pi}{2\zeta_i}} - F_{i|-\frac{\pi}{2\zeta_i}}}{2}\,, \tag{38}$$

where $F_{i|+x}$ refers to applying an additional unitary $e^{ix\zeta_i H_i}$ during state preparation.

While these operations may create quantum states outside the ansatz class of the VQA, most physical implementations can facilitate them. In total $2n_\zeta$ expressions are evaluated. If we assume shot noise that is uniform over the parameter space with a variance given by $\sigma_{i\pm}^2 = \frac{\sigma^2}{m_{i\pm}}$, where $m_{i\pm}$ is the number of measurements at one position with the total number of measurements $m = \sum_{i=1}^{n_\mu} m_{i+} + m_{i-}$, the optimal choice of measurement distribution is such that $2m_{i+} = 2m_{i-} := m_i \propto |\zeta_i|$. This yields a total error of

$$\begin{aligned}\langle\epsilon_{\text{tot}}^2\rangle_{s,\boldsymbol{\theta}} &= \sum_{i=1}^{n_\zeta} \frac{\zeta_i^2}{4}\left(2\frac{\sigma^2}{m_i/2}\right) \\ &\xrightarrow{m_i\propto|\zeta_i|} \frac{\sigma^2\|\boldsymbol{\zeta}\|_1^2}{m} \gtrsim \frac{\sigma^2\nu^2}{m}\,,\end{aligned} \tag{39}$$

where the second step optimizes over the measurement budget $\boldsymbol{m}$ and $\|\boldsymbol{\zeta}\|_1$ is an upper bound to the spectral width ν^2. This is the same scaling as for ULGE, which is also true for Ω, as this method is also unbiased.

4.3 Single Bayesian linear gradient estimator

For $m \to 0$, $\boldsymbol{w}$ will be chosen small to minimize the statistical error. As the ℓ_1 norm term heavily penalizes multiple measurements, this leads to a strategy with only a single measurement position method, similar to finite differences. The method where the measurement settings is restricted to $n_{\boldsymbol{x}} = 1$,we call SLGE. The optimization problem Eq. (24) for SLGE simplifies to

$$\langle\epsilon_{\text{tot}}^2\rangle_{s,\boldsymbol{\theta}}(w,x) = \sum_{k=1}^{n_\mu}\langle a_k^2\rangle_{\boldsymbol{\theta}}(w\sin(\mu_k x) - \mu_k)^2 + \frac{\sigma^2}{m}w^2\,, \tag{40}$$

	$\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}$ $(m\to\infty)$	Ω^2 $(m\to 0)$	$\langle\epsilon^2_{\text{stat}}\rangle_s/\langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}$	#Meas. Pos. $(2n_x)$	AC?	Priors?
BLGE	$\frac{\sigma^2\nu^2}{m}$	$\frac{m\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\sigma^2}\times\frac{1}{\nu^2_{\text{eff}}}$	$0\to 1$	$2\to 2n_\mu$	Yes	Yes
SLGE	$\propto m^{-2/3}$	$\frac{m\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\sigma^2}\times\frac{1}{\nu^2_{\text{eff}}}$	$0\to\frac{2}{3}$	2	Yes	Yes
ULGE	$\frac{\sigma^2\nu^2}{m}$	$\frac{m\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\sigma^2}\times\frac{1}{\nu^2}$	1	$2n_\mu$	Yes	No
PSR	$\frac{\sigma^2\nu^2}{m}$	$\frac{m\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\sigma^2}\times\frac{1}{\nu^2}$	1	$2n_\zeta$	No	No

Table 1: Summary of the analyzed methods. The total error refers to the limit of many measurements, while the expression for Ω is valid in the setting of very few measurements. The notation $a\to b$ indicates that the value a is valid for small measurement budgets and b is valid for large measurement budgets. AC: Ansatz Class - "yes" indicates that all expressions to be evaluated are in the original ansatz class of the VQA.

which describes a quadratic polynomial in w and a trigonometric polynomial in x, which can be minimized to numerical precision by finding roots of a polynomial of degree 3ν. We show in appendix B.3, that for $m\to\infty$, SLGE scales as

$$\langle\epsilon^2\rangle_{s,\boldsymbol{\theta}}\propto m^{-2/3}\,, \tag{41}$$

which is outperformed by the $\propto m^{-1}$ scaling of the previous methods. This shows that for the best performance for a massive measurement budget, multiple measurement positions are required. In contrast for a small measurement budget, we derive the following theorem in appendix B.4

Theorem 3. *For an optimal SLGE (and therefore BLGE) strategy, in the limit of small m, the relative correlation (Ω) can be lower-bounded by*

$$\Omega^2\geq\frac{m\langle\delta^2\rangle_{\boldsymbol{\theta}}}{\sigma^2}\times\frac{1}{\nu^2_{\text{eff}}}+O(m^2) \tag{42}$$

with the effective spectral width

$$\nu_{\text{eff}}:=\sqrt{\langle\Delta^2\rangle_{\boldsymbol{\theta}}/\langle\delta^2\rangle_{\boldsymbol{\theta}}}\in[1,\nu]\,, \tag{43}$$

the expected ratio between the second derivative Δ and first derivative δ of $F(x)$.

This theorem shows the strength of the strategy, as ν_{eff} also takes into account the relative significance of the individual eigenvalue differences. Notably, this can be significantly smaller than the spectral width when large eigenvalues are very rare, as is the case in an exponential or Gaussian like distributions with a long but negligible tail. Thus we expect that in the regime of small m, PSR and ULGE require a factor of $\frac{\nu^2}{\nu^2_{\text{eff}}}$ more measurement rounds for the same quality. For large m, one can ask when SLGE will be outperformed by ULGE type methods. In the example of used in figure 2, the crossover occurred only after $\sim 10^5$ measurements and $\Omega\geq 0.996$, which is only visible on the top right plot Ω after extensive magnification. This means that the regime where ULGE takes over is only for very large m where the exact derivative is basically known already. This behavior is not specific to the selected priors, but actually holds for any prior distributions.

Theorem 4. *For any distribution of frequency amplitudes $\mathcal{D}_{\boldsymbol{\theta}}$, single shot noise variance σ^2 and a measurement budget m, there exists an optimal SLGE strategy with a relative correlation of at least 99% that of an unbiased method. (i.e. $\Omega_{\text{S}}\geq 0.99\,\Omega_{\text{U}}$)*

The proof is shown in appendix B.5. We show this by constructing an explicit SLGE strategy which achieves this bound for all possible distributions. There we also proof the following corollary which shows that even a deterministic SLGE method only dependent on the spectral width ν is already competitive

Corollary 1. *An SLGE algorithm measuring at position $x=\frac{\pi}{2\nu}$ has a relative correlation that is at least 97.5% the relative correlation of ULGE, regardless of the underlying distribution ($\mathcal{D}_{\boldsymbol{\theta}}$, σ^2, m). (i.e. $\Omega_{\text{S}}\geq 0.975\,\Omega_{\text{U}}$)*

The theorem and corollary show quantitatively that the expected gains from using unbiased estimation methods with multiple measurement positions w.r.t. the relative correlation (Ω) are small, even for large measurement budgets. It is also worth pointing out that the theorem and corollary do not use the periodicity of the unitary explicitly, meaning they also hold for *non-periodic* unitaries with a generator of spectral width ν.

4.4 Summary of the measurement budget allocation methods

Table 1 summarizes the main methods discussed above by comparing their error scalings, the number of expressions that need to be evaluated, whether all expressions are within the ansatz class of the VQA, and whether prior estimates of the shot noise σ^2 and the second-moment matrix C_a are needed.

5 Application to QAOAs

For our numerics, we use a popular quantum approximate optimization algorithm (QAOA) setup [27]. Here, the ground state of the problem Hamiltonian

H_c encodes the solution to the MaxCut problem on a graph $\mathcal{G} = ([N], E)$ with vertex set $[N]$ and edge set E. The MaxCutproblem is the problem of finding a labelling of the vertex set that maximizes the number of so-called *cut edges*. The allowed labels are 0 and 1 and an edge is *cut* if it connects two vertices that have different labels. We identify the computational basis states of our qubits with the two labels. H_c contains terms for every edge in the graph, which are valued at -1 if the edge is cut and 0 otherwise:

$$H_c = \frac{1}{2} \sum_{(i,j)\in E} (Z_i Z_j - \mathbb{1}), \quad E \subseteq [N] \times [N]. \tag{44}$$

The energy of the system described by H_c is minimized by any state that corresponds to the maximal number of edges being cut. We denote this maximal number of cut edges by MaxCut($\mathcal{G}$).

In our numerical experiments the edge set E was randomly generated by selecting a subset of $2N$ edges from the set of all possible edges on N vertices.

The QAOA cost function is defined as

$$F(\boldsymbol{\theta}) = \langle \boldsymbol{\theta} | H_c | \boldsymbol{\theta} \rangle, \tag{45}$$

where the state $|\boldsymbol{\theta}\rangle$ is prepared using the parametrized circuit

$$|\boldsymbol{\theta}\rangle = \left[\prod_{\alpha=1}^{L} U_b(\theta_{2\alpha-1}) U_c(\theta_{2\alpha}) \right] |+\rangle^{\otimes N} \tag{46}$$

with $U_c(\theta) = \exp(\mathrm{i}\theta H_c)$ an evolution under H_c and $U_b(\theta) = \exp(\mathrm{i}\theta H_b)$ an evolution under a mixing Hamiltonian H_b. In our setup the Hamiltonian is given by

$$H_b = -\frac{1}{2} \sum_{i=1}^{N} X_i. \tag{47}$$

The number of times each unitary appears is referred to as the circuit depth and labeled L and one pair of unitaries U_b and U_c is referred to as one layer in the following. The initial state that these unitaries are applied to is $|+\rangle^{\otimes N}$, the ground state of H_b. In order to effectively compare the performance of QAOA on different graph instances $\mathcal{G}$ and $\mathcal{G}'$ with MaxCut($\mathcal{G}$) $\neq$ MaxCut($\mathcal{G}'$), we introduce the approximation ratio

$$r = -\frac{F(\boldsymbol{\theta}_{\text{alg}})}{\text{MaxCut}(\mathcal{G})} = \frac{F(\boldsymbol{\theta}_{\text{alg}})}{\lambda_{\min}(H_c)}, \tag{48}$$

where $\boldsymbol{\theta}_{\text{alg}}$ is some (possibly intermediate) parameter point determined by some optimization algorithm and $\lambda_{\min}(\,\cdot\,)$ evaluates to the smallest eigenvalue of its argument.

5.1 Deriving prior estimates

In this example, H_c is dependent on the graph instance, hence the $\langle a_k^2 \rangle_{\boldsymbol{\theta}}$ are too. Finding the exact coefficients for a particular instance can be assumed to be difficult, since determining the spectral width of H_c is already an NP-hard task in general. It is possible, however, to sample from graph bipartitions to find an approximate spectral distribution numerically.

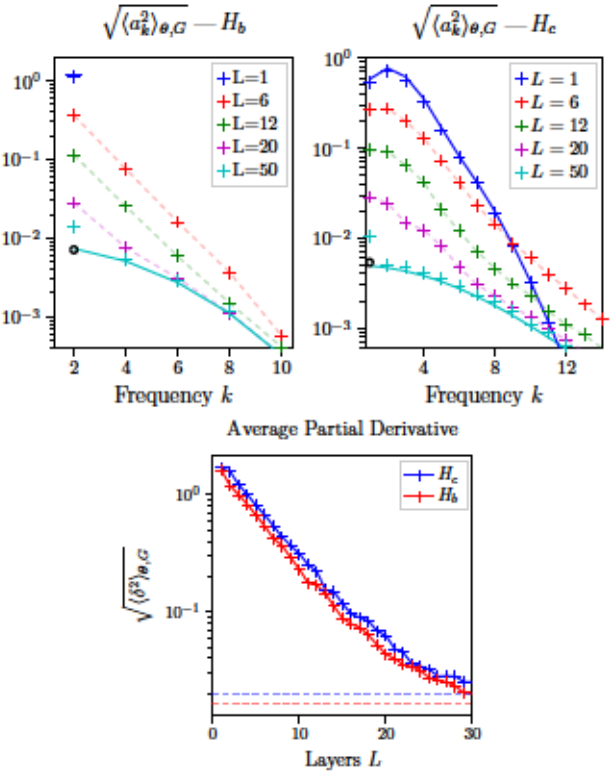

Figure 3: **Top:** The dependence of $\langle a_k^2 \rangle_{\boldsymbol{\theta}}$ on the frequency μ_k for various circuit depths L and a system size $N = 16$ with $M = 32$ edges as well as a uniform distribution over $\boldsymbol{\theta}$ and all graphs for 2000 randomly sampled instances. The Fourier decomposition is taken w.r.t. a layer in the bulk of the circuit, specifically the $\lceil \frac{L}{2} \rceil$-th layer. The decomposition is for the H_b gate (left) and the H_c gate (right). The solid lines for $L = 1, 50$ represent analytical estimates. The black circles show the estimate when we restrict the distribution to graphs which are fully connected and exhibit no non-trivial graph automorphism. **Bottom:** The root of mean square-amplitude of the partial derivative in dependence of the circuit depth. The dashed line shows the Barren plateau limit.

To derive analytical estimates, we are extending our distribution to also include all considered graph instances

$$\langle \cdot \rangle_{s,\boldsymbol{\theta}} \to \langle \cdot \rangle_{s,\boldsymbol{\theta},G}, \tag{49}$$

where G indicates drawing the samples from the set of all graphs with M edges and N vertices. This generalization makes it possible to estimate the priors analytically, which we do in appendix D. The results are shown in figure 3. The plotted points represent a mean over 2,000 randomly chosen points and graph instances. For $L = 1$ we derive analytical expressions for both Hamiltonians (dark blue line) in appendix D.2. While this is a tedious problem, it is

efficiently solvable. For U_b, the specific structure of the VQA means that only the Fourier coefficient for $\mu = 2$ does not vanish and in general, only even coefficients contribute. For $L \to \infty$ we obtain an analytical estimate (cyan line) using the 2-design assumption in appendix D.1. While this estimate faithfully reproduces the bulk of the frequencies at $L = 50$ layers, the empirical estimate for the first frequency is significantly larger than the theoretical prediction. This discrepancy arises from graph instances which correspond to non-universal gate-set VQAs instances. In particular, this is the case when the graph is not fully connected or has a non-trivial automorphism. The additional black circles for the first frequency in the figures shows the empirical estimate obtained when we restrict the set G to only graphs which do not exhibit these properties. As can be seen these instances faithfully reproduce the 2-design prediction.

Figure 3c shows this convergence of the expected derivative. Numerically, the convergence appears to be close to completed at $L \sim 30$ layers, which is consistent with known 2-design convergence results in literature, where depth scales roughly on the order of the number of qubits [28]. The dotted lines show the barren plateau limit when the ergodic argument is used. Again due to the existence of graphs with symmetries, the convergence is not exactly to the theoretical 2-design limit. For the intermediate layers, we can argue (particularly for the case regarding H_c) that the overall amplitude of the coefficients decays exponentially while the relative values transition from a behavior like $L = 1$ to one more closely related to $L \to \infty$. It is our belief that a more rigorous understanding of t-design convergence within quantum circuits could help to make a more quantitative assessment than we are able to make at present. We note that while quantifying the priors may be a worthwhile theoretical pursuit, for practical applications it often suffices to have a rough estimate of the amplitudes, as even with significant overestimation of amplitudes our method will still outperform traditional non-Bayesian schemes. Also, since real implementations attempt to avoid the barren plateau regimes, the gradient along the optimization path may be significantly larger and therefore not fully representative of the behavior of the ensemble average.

5.2 Numerical results

In this section, we first test the estimation accuracy for the different gradient estimation strategies for randomly selected angles $\boldsymbol{\theta}$. Afterwards, we demonstrate their usefulness for gradient descent based VQA optimization.

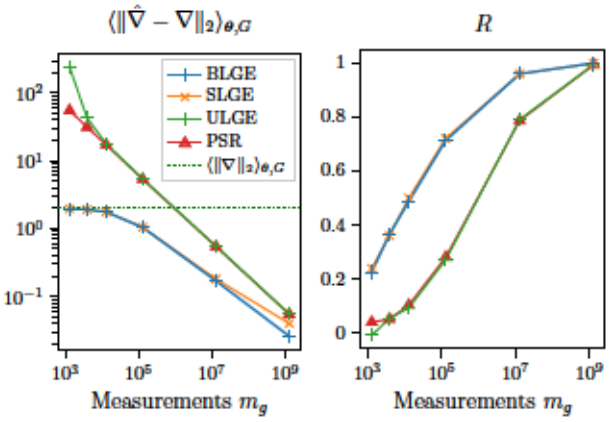

Figure 4: A comparison of the different gradient estimation methods for different measurement budget for the entire gradient (m_g). The data was generated with a system size of $N = 18$, graph instances with $M = 36$ edges, and a circuit depth of $L = 12$ with $\langle a_k^2\rangle_{\boldsymbol{\theta}}$ estimates as stated in Eq. (50). Each data point represents an empirical mean over 300 total samples drawn from 30 different graph instances. 10 random parameter points were evaluated for each instance. The measurement budget is distributed in such a way that the H_c-layers get twice the measurements of the H_b-layers. **Left:** The average 2-norm distance between the exact gradient and the estimate using the different methods. **Right:** The average relative slope R.

5.2.1 Gradient quality for the different allocation methods

Figure 4 shows the quality of the estimated gradients for the different optimization routines and a range of measurement budgets. The size of the measurement budget (m_g) is the entire budget, i.e. jointly for all partial derivatives. The first value of $m_g = 1296$ is just big enough to allow PSR and ULGE to have exactly one measurement round per expression to be evaluated, i.e. these methods cannot be run as intended with a smaller measurement budget. For the priors of the Fourier coefficients, we select

$$\begin{aligned} \langle a_k^2\rangle_{\boldsymbol{\theta}}\big|_b &= 10^{-0.3k-1.1} \times \delta_{k\in 2\mathbb{Z}} \\ \langle a_k^2\rangle_{\boldsymbol{\theta}}\big|_c &= 10^{-0.3k-1.6} \end{aligned} \tag{50}$$

resulting from a rough exponential fit obtained from figure 3. Figure 4a depicts the dependence of the 2-norm difference between the exact and the estimated gradient for different budget m_g

$$\|\hat{\nabla} - \nabla\|_2\,, \tag{51}$$

where ∇ is the exact gradient and $\hat{\nabla}$ the estimate generated by performing the estimation routine for every component of the gradient. We note that for very few measurements ULGE performs significantly worse than PSR. This is because the required positive integer rounding for the individual number of measurements at each site implies a far from optimal allocation for ULGE, while PSR does not require any

rounding. Besides this we see good agreement between the numerical results and what was predicted from figure 2. Explicitly enforcing the condition that only two position are to be evaluated as in SLGE (section 4.3) only starts to make a significant difference compared to BLGE at around 10^7 measurements which may be already infeasible in a practical experiment. This shows that while the asymptotic behavior is significantly worse, for practical purposes, a correctly chosen finite differences model performs remarkably well. For fewer measurements, ULGE and PSR require already $m_g = 10^5$ measurements to outperform an estimator which returns the all zero vector.

For the purpose of gradient descent, one can argue that the direction of the gradient is actually more important than its magnitude. In order to quantify this notion we investigate the *relative slope* R, the ratio between the slope in the direction of the actual gradient and the slope in the direction of the estimated gradient

$$R := \left\langle \frac{\hat{\nabla}^T \nabla}{\|\hat{\nabla}\|_2 \|\nabla\|_2} \right\rangle_{\boldsymbol{\theta}} = \left\langle \frac{\mathrm{desc}(\hat{\nabla})}{\mathrm{desc}(\nabla)} \right\rangle_{\boldsymbol{\theta}}, \tag{52}$$

where

$$\mathrm{desc}(\boldsymbol{g}) = \frac{\mathrm{d}}{\mathrm{d}x}\left[F\left(\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}x\right)\right], \tag{53}$$

which indicates the slope in the direction of $\boldsymbol{g}$. A value of $R = 0$ would indicate that the estimated gradient is orthogonal to the actual gradient. A perfect estimator would achieve a value of $R = 1$. The relative slope R is similar in nature to Ω defined in Eq. (29) but differs in the way that the ensemble averages are taken. Numerically, the manner in which the averages are taken does not qualitatively change our results. As with the 2-norm error, we see good agreement with the behavior of R and the behavior predicted in figure 2 for Ω. This also justifies why Ω is indeed a good quality parameter for the estimation. For the fewest number of measurements considered, PSR and ULGE struggle to find any decreasing direction, while BLGE reliably has a $R = 20\%$ overlap. The other methods require nearly 100 times the number of measurements for the same quality. Similarly to what we saw for Ω, BLGE and SLGE show basically identical performance with regard to the quality measure R.

5.2.2 Parameter Optimization

We also simulated a complete parameter optimization routine. Following the proposal from Zhou et al. [6], we chose the initial parameters $\boldsymbol{\theta}$ to resemble an approximate linear annealing ramp, as this can significantly improve performance compared to random initialization. The exact initialization we used was (even index H_b, odd H_c)

$$\theta_i^{(0)} = \frac{\pi}{20}\left(\frac{4 - 5\delta_{i\in 2\mathbb{Z}}}{L-1} \times i + 4\right). \tag{54}$$

Since the VQA starts from a specific initialization, the uniformity assumption of our assumed distribution $\mathcal{D}_{\boldsymbol{\theta}}$ might no longer be valid, meaning this also tests the applicability of our approach outside of the idealized conditions we assumed. For the update step, we use a basic gradient descent routine

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta^{(t)}\hat{\nabla}^{(t)}. \tag{55}$$

Since the different methods return gradients with massively different norms, a fixed step size will skew the result heavily. To compensate, we use a backtracking line-search routine to find a good $\eta^{(t)}$. It starts with an initial step size that is significantly too large. Then it repeatably measures the observable at the proposed new point $\boldsymbol{\theta}^{(t)} - \eta^{(t)}\hat{\nabla}^{(t)}$ and either accepts the step size if the estimate decreased compared to the current position $\boldsymbol{\theta}^{(t)}$ or halves the step size $\eta^{(t)} \to \eta^{(t)}/2$ and repeats. The measurement budget allocated for the line-search estimate is set to be the same as for the gradient estimation. This line search removes the dependence on the size of the returned gradient without the algorithm becoming a full sweeping algorithm.

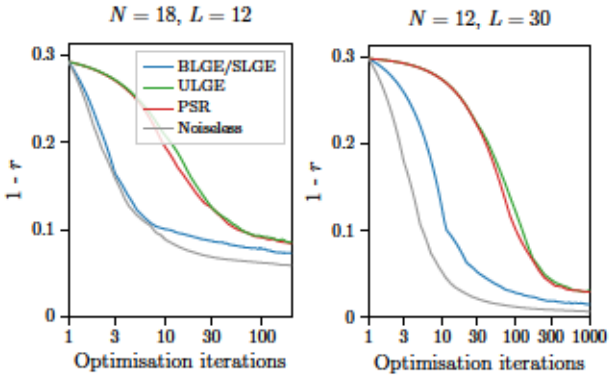

Figure 5: The empirical average approximation ratio r over the iterations of a gradient descent optimization with line search for the different optimization routines. The results are averaged over 23 different problem instances. We note that in the considered measurement regimes BLGE and SLGE are identical. **Left:** A shallow circuit with $N = 18$, $M = 36$, $L = 12$ with the total measurement for each gradient being $m_g = 3888$, which corresponds to 6 measurements for each generator in the PSR. The priors were taken as an exponential ansatz as given in Eq. (50). **Right:** A deeper circuit with $N = 12$, $M = 24$, $L = 30$, $m_g = 6480$, which also corresponds to 6 measurements for each generator. The priors were taken as the barren plateau estimates Eq. (150) and Eq. (154).

Figure 5 shows the numerical results of the optimization routine using each of the gradient estimation

routines. The cost function shown here is the approximation ratio r as defined in Eq. (48), which is a common metric for QAOA numerics. We consider both a case with a shallow circuit ($L = 12$) with $N = 18$ and a deep circuit ($L = 30$) in the barren plateau regime, but for a smaller system size of $N = 12$. In both these cases we consider an overall measurement budget so that each measurement setup for PSR is performed 3 times. This translates to $m_g = 3888$ for the shallow circuit case and $m_g = 6480$ for the deep circuit. It is also worth noting that with such a small measurement budget, BLGE and SLGE are identical. Additionally, we plot the case for the exact gradient with an infinite measurement budget in gray.

Figure 5 also shows that BLGE significantly outperforms the unbiased methods with convergence being nearly an order of magnitude faster and also reaching a better minimal value.

6 Conclusion and outlook

We have shown that using a Bayesian approach in a generalized PSR setting can significantly improve the resulting quality of the estimation, which ultimately improves the overall run time and results of the VQA. In particular, when dealing with barren plateaus, this estimation tool may prove crucial for performance in practical implementations. Our study also makes a strong argument that central difference methods with a reasonably chosen step size is a very good first strategy that is only outperformed by the unbiased PSR for large measurement budgets.

Our work opens up several new research quests.

- In future work, we aim to formally extend our framework to non-periodic unitaries.
- Understanding the second moment matrix and its convergence into the barren plateau regime might allow us to develop strategies to mitigate its effects allowing VQAs to be effective for more ansatz classes. Such improved priors might come from studies of approximate unitary 2-designs.
- Improved optimization methods, such as natural gradient estimation [9, 29, 30] or higher order optimization procedures requiring second order derivatives, may also benefit from introducing prior assumptions into their estimation routines.
- In particular, for practical applications where quantum computation is expensive and classical computation is cheap, using an adaptive estimator may be advantageous, i.e. updating the Fourier priors along the optimization path may also help to improve the gradient quality further.

7 Aknowledgement

We thank Raphael Brieger, Lucas Tendick, Markus Heinrich, Juan Henning, Michael J. Hartmann, Nathan McMahon, and Samuel Wilkinson, for fruitful discussions. This work has been funded by the German Federal Ministry of Education and Research (BMBF) within the funding program "quantum technologies – from basic research to market" via the joint project MANIQU (grant number 13N15578) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the grant number 441423094 within the Emmy Noether Program.

8 Acronyms and list of symbols

N: number of qubits/vertices (QAOA)
L: number of layers of the VQA
$\boldsymbol{\lambda}$: eigenvalues of the gate generator
$\boldsymbol{\mu}$: eigenvalue differences of the gate generator
a_k, b_k, c_k: Fourier coefficients
C_a: Second moment matrix of the Fourier coefficient a_k
σ: single shot shot noise
w_t: weights of linear estimator
$\boldsymbol{\theta}$: VQA parameters
$\theta = x_t$: measurement position for given $\theta \in \boldsymbol{\theta}$
n_x: number of (positive) measurement positions
m_t: number of measurement rounds for x_t
m: total number $m = \sum m_t$
ν: largest difference between two eigenvalues of the unitary generator, i.e. $\max_k\{\mu_k\}$
δ: derivative of the cost function w.r.t. one parameter θ
Δ: second derivative of the cost function w.r.t. one parameter θ

References

[1] F. Arute et al., *Quantum supremacy using a programmable superconducting processor*, Nature **574**, 505 (2019), arXiv:1910.11333 [quant-ph].
[2] H. S. Zhong et al., *Quantum computational advantage using photons*, Science **370**, 1460 (2020), arXiv:2012.01625 [quant-ph].

[3] J. Preskill, *Quantum computing in the NISQ era and beyond*, Quantum **2**, 10.22331/q-2018-08-06-79 (2018), arXiv:1801.00862 [quant-ph].

[4] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, *Barren plateaus in quantum neural network training landscapes*, Nat. Commun. **9**, 4812 (2018), arXiv:1803.11173 [quant-ph].

[5] L. Bittel and M. Kliesch, *Training variational quantum algorithms is NP-hard*, Phys. Rev. Lett. **127**, 120502 (2021), arXiv:2101.07267 [quant-ph].

[6] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, *Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices*, Phys. Rev. X **10**, 021067 (2020), arXiv:1812.01041 [quant-ph].

[7] H. R. Grimsley, S. E. Economou, E. Barnes, and N. J. Mayhall, *An adaptive variational algorithm for exact molecular simulations on a quantum computer*, Nat. Commun. **10**, 3007 (2019), arXiv:1812.11173 [quant-ph].

[8] H. R. Grimsley, G. S. Barron, E. Barnes, S. E. Economou, and N. J. Mayhall, *ADAPT-VQE is insensitive to rough parameter landscapes and barren plateaus*, arXiv:2204.07179 [quant-ph] (2022).

[9] D. Wierichs, C. Gogolin, and M. Kastoryano, *Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer*, Phys. Rev. Research **2**, 043246 (2020), arXiv:2004.14666 [quant-ph].

[10] J. Li, X. Yang, X. Peng, and C.-P. Sun, *Hybrid quantum-classical approach to quantum optimal control*, Phys. Rev. Lett. **118**, 150503 (2017), arXiv:1608.00677 [quant-ph].

[11] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, *Quantum circuit learning*, Phys. Rev. A **98**, 032309 (2018), arXiv:1803.00745 [quant-ph].

[12] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, *Evaluating analytic gradients on quantum hardware*, Phys. Rev. A **99**, 032331 (2019), arXiv:1811.11184 [quant-ph].

[13] L. Banchi and G. E. Crooks, *Measuring analytic gradients of general quantum evolution with the stochastic parameter shift rule*, Quantum **5**, 386 (2021), arXiv:2005.10299 [quant-ph].

[14] O. Kyriienko and V. E. Elfving, *Generalized quantum circuit differentiation rules*, Phys. Rev. A **104**, 052417 (2021), arXiv:2108.01218 [quant-ph].

[15] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, *General parameter-shift rules for quantum gradients*, Quantum **6**, 677 (2022), arXiv:2107.12390 [quant-ph].

[16] J. Gil Vidal and D. O. Theis, *Calculus on parameterized quantum circuits*, arXiv:1812.06323 [quant-ph].

[17] D. O. Theis, *Optimality of finite-support parameter shift rules for derivatives of variational quantum circuits*, arXiv:2112.14669 [quant-ph].

[18] A. F. Izmaylov, R. A. Lang, and T.-C. Yen, *Analytic gradients in variational quantum algorithms: Algebraic extensions of the parameter-shift rule to general unitary transformations*, Phys. Rev. A **104**, 062443 (2021), arXiv:2107.08131 [quant-ph].

[19] A. Harrow and J. Napp, *Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms*, Phys. Rev. Lett. **126**, 140502 (2021), arXiv:1901.05374 [quant-ph].

[20] G. E. Crooks, *Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition*, arXiv:1905.13311 [quant-ph].

[21] J. S. Kottmann, A. Anand, and A. Aspuru-Guzik, *A feasible approach for automatically differentiable unitary coupled-cluster on quantum computers*, Chemical Science **12**, 3497–3508 (2021), arXiv:2011.05938.

[22] A. Gilyén, S. Arunachalam, and N. Wiebe, *Optimizing quantum optimization algorithms via faster quantum gradient computation*, in *Proc. 2019 Ann. ACM-SIAM Symp. Discrete Algorithms (SODA)* (2019) pp. 1425–1444, arXiv:1711.00465 [quant-ph] .

[23] C. Cade, L. Mineh, A. Montanaro, and S. Stanisic, *Strategies for solving the Fermi-Hubbard model on near-term quantum computers*, Phys. Rev. B **102**, 235122 (2020), arXiv:1912.06007 [quant-ph].

[24] A. W. Harrow and R. A. Low, *Random quantum circuits are approximate 2-designs*, Commun. Math. Phys. **291**, 257 (2009), arXiv:0802.1919 [quant-ph].

[25] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, *Local random quantum circuits are approximate polynomial-designs*, Commun. Math. Phys. **346**, 397 (2016), arXiv:1208.0692.

[26] L. Bittel, J. Watty, and M. Kliesch, *Fast gradient estimation for VQAs*, `https://github.com/lennartbittel/GradientEstimationVQA` (2022), [Online; accessed 11-October-2022].

[27] E. Farhi, J. Goldstone, and S. Gutmann, *A quantum approximate optimization algorithm*, arXiv:1411.4028 [quant-ph] (2014).

[28] N. Hunter-Jones, *Unitary designs from statistical mechanics in random quantum circuits*, arXiv:1905.12053 [quant-ph].

[29] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, *Quantum natural gradient*, Quantum **4**, 269 (2020), arXiv:1909.02108 [quant-ph].

[30] N. Yamamoto, *On the natural gradient for variational quantum eigensolver*, arXiv:1909.05074 [quant-ph].

[31] M. Kliesch and I. Roth, *Theory of quantum system certification*, PRX Quantum **2**, 010201 (2021), tutorial, arXiv:2010.05925 [quant-ph].

[32] J. Emerson, R. Alicki, and K. Życzkowski, *Scalable noise estimation with random unitary operators*, J. Opt. B **7**, S347 (2005), arXiv:quant-ph/0503243.

Appendices

A Solving the Bayesian allocation problem

Here we derive the theoretic underpinnings of the Bayesian allocation method. Namely, we derive the effective dual that is used for the numerical implementation and proof theorem 1, which concerns the sparsity of the solution.

A.1 Finding the dual problem

In this section, we derive the dual formulation, we use to find the optimal measurement positions as explained in the main text. The minimization (Eq. (24)) we are interested in can be rewritten into a constraint problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^{n_x},\boldsymbol{w}\in\mathbb{R}^{n_x}} \langle\epsilon^2_{\text{tot}}\rangle_{s,\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{w}) = \min_{\boldsymbol{x}\in\mathbb{R}^{n_x},\boldsymbol{w}\in\mathbb{R}^{n_x}} \sum_{k=1}^{n_\mu}\langle a_k^2\rangle_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n_x} w_i\sin(x_i\mu_k)-\mu_k\right)^2 + \frac{\sigma^2}{m}\|\boldsymbol{w}\|_1^2 \tag{56}$$

$$= \min_{\boldsymbol{x}\in\mathbb{R}^{n_x},\boldsymbol{w}\in\mathbb{R}^{n_x},\boldsymbol{z}\in\mathbb{R}^{n_\mu},l\in\mathbb{R}} \sum_{k=1}^{n_\mu}\langle a_k^2\rangle_{\boldsymbol{\theta}} z_k^2 + \frac{\sigma^2}{m}l^2 \tag{57}$$

$$\text{s.t.} \quad \sum_{i=1}^{n_x} w_i\sin(x_i\mu_k)-\mu_k = z_k\,,\ \|\boldsymbol{w}\|_1\le l\,, \tag{58}$$

where we introduced the variables $\boldsymbol{z}$ and l. If we keep $\boldsymbol{x}$ fixed (it can be assumed to describe a fine grid covering a complete period of the function), this describes a convex optimization problem. We proceed to derive the dual g.

$$g(\boldsymbol{x};\boldsymbol{z},\boldsymbol{w},l;\boldsymbol{\kappa},\tau) = \sum_{k=1}^{n_\mu}\langle a_k^2\rangle_{\boldsymbol{\theta}} z_k^2 + \frac{\sigma^2}{m}l^2 + \sum_{k=1}^{n_\mu} 2\kappa_k\left(\sum_{i=1}^{n_x} w_i\sin(x_i\mu_k)-\mu_k-z_k\right) - 2\tau(l-\|\boldsymbol{w}\|_1) \tag{59}$$

$$\frac{\partial g}{\partial z_k} = 2\langle a_k^2\rangle_{\boldsymbol{\theta}} z_k - 2\kappa_k \to z_k = \frac{\kappa_k}{\langle a_k^2\rangle_{\boldsymbol{\theta}}} \tag{60}$$

$$\frac{\partial g}{\partial l} = 2\frac{\sigma^2}{m}l - 2\tau \to l = \tau\frac{m}{\sigma^2} \tag{61}$$

$$\frac{\partial g}{\partial w_i} = 2\tau\,\text{sgn}(w_i) + 2\sum_{k=1}^{n_\mu}\kappa_k\sin(x_i\mu_k) \to \tau \ge \left|\sum_{k=1}^{n_\mu}\kappa_k\sin(x_i\mu_k)\right|, \tag{62}$$

where the last step follows as g is affine w.r.t. w_i for the positive and negative axis. Equation (62) also implies that either $w_i = 0$ or $\tau = |\sum_{k=1}^{n_\mu}\kappa_k\sin(x_i\mu_k)|$, also known as complementary slackness.

$$g(\boldsymbol{x};\boldsymbol{\kappa},\tau) = -\sum_{k=1}^{n_\mu}\frac{\kappa_k^2}{\langle a_k^2\rangle_{\boldsymbol{\theta}}} - \frac{m}{\sigma^2}\tau^2 - 2\sum_{k=1}^{n_\mu}\kappa_k\mu_k \tag{63}$$

$$\text{s.t.} \quad \forall x_i: \quad \tau \ge |\sum_{k=1}^{n_\mu}\kappa_k\sin(x_i\mu_k)| \tag{64}$$

Since $\boldsymbol{x}$ wants to minimize this expression, τ needs to be maximized. As all measurement positions are allowed and we impose no bound on n_x meaning $\boldsymbol{x}$ can be distributed arbitrarily dense, it follows

$$\max_{i\in[n_x]}|\sum_{k=1}^{n_\mu}\kappa_k\sin(x_i\mu_k)| \to \|\sum_{k=1}^{n_\mu}\kappa_k\sin(\mu_k(\cdot))\|_\infty\,, \tag{65}$$

where the infinity norm refers the absolute value maximum over the domain of the function, and we get $\tau = \|\sum_{k=1}^{n_\mu}\kappa_k\sin(\mu_k(\cdot))\|_\infty$, which leads to

$$g(\boldsymbol{\kappa}) = -\sum_{k=1}^{n_\mu}\frac{\kappa_k^2}{\langle a_k^2\rangle_{\boldsymbol{\theta}}} - \frac{m}{\sigma^2}\|\sum_{k=1}^{n_\mu}\kappa_k\sin(\mu_k(\cdot))\|_\infty^2 - 2\sum_{k=1}^{n_\mu}\kappa_k\mu_k\,. \tag{66}$$

We conclude that

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}^* = \max_{\boldsymbol{\kappa} \in \mathbb{R}^{n_\mu}} g(\boldsymbol{\kappa}) \,, \tag{67}$$

since the minimization of the primal is convex for fixed $\boldsymbol{x}$. Solving this requires maximizing a concave problem. g has the same solution as minimizing

$$\bar{g}(\boldsymbol{\kappa}) = \sum_{k=1}^{n_\mu} \frac{\sigma^2}{m \langle a_k^2 \rangle_{\boldsymbol{\theta}}} \kappa_k^2 + \| \sum_{k=1}^{n_\mu} \kappa_k \sin(\mu_k(\cdot)) \|_\infty^2 - 2 \sum_{k=1}^{n_\mu} \kappa_k \mu_k \tag{68}$$

with $\bar{g} = -\frac{m}{\sigma^2} g$ and $\kappa^* \to -\frac{\sigma^2}{m} \kappa^*$ which we find has better numerical stability, especially for large measurement budgets m. Via complementary slackness, the final measurement positions $\boldsymbol{x}^*$ are just the global maxima positions of $\rho_{\boldsymbol{\kappa}}(x) = |\sum_{k=1}^{n_\mu} \kappa_k^* \sin(x \mu_k)|$, which can be obtained by solving a trigonometric polynomial, which has a most ν maxima. To obtain $\boldsymbol{w}^*$, we solve the original problem for the now fixed measurement positions x^*.

A.2 Proof of theorem 1

Here we proof that the optimization problem has a sparse solution. Explicitly that $n_x = n_\mu$ positions suffices. For this we assume that we have set of measurement positions $\boldsymbol{x} \in [0, \pi)^{n_x}$. For the problem we assume to have

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}} = (S^{\boldsymbol{x}} \boldsymbol{w} - \boldsymbol{\mu})^T C_a (S^{\boldsymbol{x}} \boldsymbol{w} - \boldsymbol{\mu}) + \frac{\sigma^2}{m} \|\boldsymbol{w}\|_1^2 \,, \tag{69}$$

and an optimal solution $\boldsymbol{w}^*$. As such we can define $\langle \epsilon_{\text{sys}}^2 \rangle_{s,\boldsymbol{\theta}}^* \coloneqq (S^{\boldsymbol{x}} \boldsymbol{w}^* - \boldsymbol{\mu})^T C_a (S^{\boldsymbol{x}} \boldsymbol{w}^* - \boldsymbol{\mu})$ and $y^* = S^{\boldsymbol{x}} \boldsymbol{w}^*$ This means the problem is equivalent as

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}} = \langle \epsilon_{\text{sys}}^2 \rangle_{s,\boldsymbol{\theta}}^* + \frac{\sigma^2}{m} \left(\min_{\boldsymbol{w}: \boldsymbol{y}^* = S^{\boldsymbol{x}} \boldsymbol{w}} \|\boldsymbol{w}\|_1 \right)^2 , \tag{70}$$

The latter term constrains an ℓ_1-norm optimization with n_μ linear constraints. It is well known that there exists an optimal sparse solution where the number of non-vanishing entries is at most the number of constraints, which are known as *basic feasible solutions* in LP literature. As this holds regardless of the actual value $\boldsymbol{y}^*$, there also exists a sparse solution for the entire problem which proofs the theorem. □

The same also holds for the unbiased case, where $\boldsymbol{y}^* = \boldsymbol{\mu}$ a strict requirement.

B Single Bayesian linear gradient estimator (SLGE)

In the following section we derive various properties of our single measurement estimation strategy. In appendix B.1, we derive the optimal coefficient w for our estimator, which we use to rewrite the expected total error of our estimator as a function of only the measurement position x. Next in appendix B.3, we proof the scaling in the limits of many measurements from Eq. (41). In appendix B.4, we prove theorem 3, which is valid for very few measurements. Finally, in appendix B.5 we prove theorem 4 and corollary 1 from our main text, which are about the relative performance of SLGE and ULGE.

B.1 Preliminaries

Recall that we can express the expected total error as

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}(w, x) = \sum_{k=1}^{n_\mu} \langle a_k^2 \rangle_{\boldsymbol{\theta}} (w \sin(\mu_k x) - \mu_k)^2 + \frac{\sigma^2}{m} w^2 \tag{71}$$

$$=: [\![(w \sin(\mu x) - \mu)^2]\!] + \frac{\sigma^2}{m} w^2 \,, \tag{72}$$

where we introduced the shorthand for the ensemble average

$$[\![g(\mu)]\!] \coloneqq \sum_{k}^{n_\mu} \langle a_k^2 \rangle_{\boldsymbol{\theta}} g(\mu_k) \,, \tag{73}$$

for an arbitrary function g.

Given x, the optimal coefficient w can be determined to be

$$w^* = \frac{[\![\mu \sin(\mu x)]\!]}{[\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}} \,. \tag{74}$$

Plugging w^* into Eq. (72) yields

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}(x) = [\![\mu^2]\!] - \frac{[\![\mu \sin(\mu x)]\!]^2}{[\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}} = \frac{[\![\mu^2]\!][\![\sin^2(\mu x)]\!] - [\![\mu \sin(\mu x)]\!]^2 + [\![\mu^2]\!]\frac{\sigma^2}{m}}{[\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}} \tag{75}$$

B.2 Numerical optimization

We are now going to outline how to numerically choose the optimal value of x. For the numerical minimization we use Eq. (75). Here we assume that estimates of $\langle a_k^2 \rangle_{\boldsymbol{\theta}}$ and σ are known.

The derivative of Eq. (75) w.r.t. x is

$$\begin{aligned} \partial_x \langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}(x) &= -\frac{2[\![\mu \sin(\mu x)]\!][\![\mu^2 \cos(\mu x)]\!]\left([\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}\right) - [\![\mu \sin(\mu x)]\!]^2 [\![2\mu \cos(\mu x)\sin(\mu x)]\!]}{\left([\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}\right)^2} \\ &= -\frac{2[\![\mu \sin(\mu x)]\!]}{\left([\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}\right)^2} \times \underbrace{\left([\![\mu^2 \cos(\mu x)]\!]\left([\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}\right) - [\![\mu \sin(\mu x)]\!][\![\mu \cos(\mu x)\sin(\mu x)]\!]\right)}_{=:h(x)} \,. \end{aligned} \tag{76}$$

The relevant minima candidates are therefore the roots of the second factor - labeled $h(x)$ - since roots of the first factor always yield a maximum of $\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}(x)$ (as the first factor appears in Eq. (75) with a negative sign).

We can rewrite $h(x)$ explicitly as

$$\begin{aligned} h(x) &= [\![\mu^2 \cos(\mu x)]\!]\left([\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}\right) - \frac{1}{2}[\![\mu \sin(\mu x)]\!][\![\mu \sin(2\mu x)]\!] \\ &= \sum_{k,l} \langle a_k^2 \rangle_{\boldsymbol{\theta}} \langle a_l^2 \rangle_{\boldsymbol{\theta}} \mu_k^2 \cos(\mu_k x)\sin^2(\mu_l x) - \frac{1}{2}\sum_{k,l} \langle a_k^2 \rangle_{\boldsymbol{\theta}} \langle a_l^2 \rangle_{\boldsymbol{\theta}} \mu_k \mu_l \sin(\mu_k x)\sin(2\mu_l x) + \frac{\sigma^2}{m}\sum_k \langle a_k^2 \rangle_{\boldsymbol{\theta}} \mu_k^2 \cos(\mu_k x) \,, \end{aligned} \tag{77}$$

which can be solved efficiently using standard solvers as it only requires finding the roots of a trigonometric polynomial of degree 3ν. From these candidate solutions, we can find the numerical exact optimal solution.

B.3 Limit for many measurements

In this section we determine the scaling of the expected total error in the limit of many measurements, i.e. high measurement accuracy, as stated in table 1.

For $m \to \infty$, the optimal measurement position becomes $x \to 0$, since noiseless measurements lead to a finite difference approximation which becomes exact when the measurement position approaches 0. This observation justifies the use of a Taylor series expansion

$$[\![\mu^2]\!][\![\sin^2(\mu x)]\!] - [\![\mu \sin(\mu x)]\!]^2 = \frac{\xi}{2}x^6 + O(x^8) \,, \tag{78}$$

where $\xi = \frac{[\![\mu^2]\!][\![\mu^6]\!] - [\![\mu^4]\!][\![\mu^4]\!]}{18}$, which is non-negative. Plugging this expression into Eq. (75) yields

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}(x) = \frac{\frac{\xi}{2}x^6 + [\![\mu^2]\!]\frac{\sigma^2}{m} + O(x^8)}{[\![\mu^2]\!]x^2 + \frac{\sigma^2}{m} + O(x^4)} \tag{79}$$

$$= \frac{\frac{\xi}{2}x^6 + [\![\mu^2]\!]\frac{\sigma^2}{m}}{[\![\mu^2]\!]x^2} + O(m^{-1}, x^6) \tag{80}$$

The expression is minimized by

$$x^* = \frac{[\![\mu^2]\!]^{1/6}}{\xi^{1/6}}\left(\frac{\sigma^2}{m}\right)^{1/6} , \tag{81}$$

$$\langle \epsilon_{\text{tot}}^2 \rangle_{s,\boldsymbol{\theta}}(x^*) \to \frac{3\xi^{1/3}}{2[\![\mu^2]\!]^{1/3}}\left(\frac{\sigma^2}{m}\right)^{2/3} \propto m^{-2/3} \,. \tag{82}$$

For $m \to \infty$, taking the limit of Eq. (74), the single measurement scheme will converge to a standard central differences method with coefficient $w^* \sim \frac{1}{x^*}$. Therefore, the expected statistical error $\langle \epsilon^2_{\text{stat}} \rangle_{s,\boldsymbol{\theta}} = w^2 \frac{\sigma^2}{m}$ becomes

$$\langle \epsilon^2_{\text{stat}} \rangle_{s,\boldsymbol{\theta}} \approx \frac{\sigma^2}{m x^{*2}} = \frac{2}{3} \langle \epsilon^2_{\text{tot}} \rangle_{s,\boldsymbol{\theta}} \,, \tag{83}$$

meaning that for $m \to \infty$ the statistical error makes up $2/3$ of the total error.

B.4 Limit for few measurements – proof of theorem 3

We are now going to turn our attention toward the regime of few measurements. In this scenario statements about the scaling of the total expected error do not make much sense since the SLGE strategy tends to return very small gradient estimates in the case of low measurement accuracy. This means that the total expected error becomes the expected magnitude of the partial derivative ($\langle \epsilon^2_{\text{tot}} \rangle_{s,\boldsymbol{\theta}} = \langle \delta^2 \rangle_{\boldsymbol{\theta}}$), where it plateaus, as can be seen in figure 4. Instead, the relative correlation describes a useful quantity in this regime. One can straightforwardly derive these simple expressions.

$$\langle \delta^2 \rangle_{\boldsymbol{\theta}} = [\![\mu^2]\!] \,, \tag{84}$$

$$\langle \hat{\delta} \delta \rangle_{s,\boldsymbol{\theta}} = w [\![\mu \sin(\mu x)]\!] \,, \tag{85}$$

$$\langle \hat{\delta}^2 \rangle_{s,\boldsymbol{\theta}} = w^2 [\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m} w^2 \,. \tag{86}$$

Also, for the second derivative of the cost function F, which we denote as Δ, we get

$$\langle \Delta^2 \rangle_{\boldsymbol{\theta}} := \langle (\delta^2_x F(0))^2 \rangle_{\boldsymbol{\theta}} = [\![\mu^4]\!] \,, \tag{87}$$

which uses that $\langle a_k^2 \rangle_{\boldsymbol{\theta}} = \langle b_k^2 \rangle_{\boldsymbol{\theta}}$ resulting from the shift invariance assumption of $\mathcal{D}_{\boldsymbol{\theta}}$.

To proof theorem 3, proceed to derive a lower bound to the relative correlation (Ω) as defined in Eq. (29). For $m \ll \sigma^2 / [\![\sin^2(\mu x)]\!]$, we can approximate Ω^2 as

$$\Omega^2 = \frac{[\![\mu \sin(\mu x^*)]\!]^2}{[\![\mu^2]\!] \left([\![\sin^2(\mu x^*)]\!] + \frac{\sigma^2}{m} \right)} \tag{88}$$

$$= m \frac{[\![\mu \sin(\mu x^*)]\!]^2}{[\![\mu^2]\!] \sigma^2} + O(m^2) \,. \tag{89}$$

using the following lemma 1 for the last step.

$$\Omega^2 \geq m \frac{[\![\mu^2]\!]^2}{\sigma^2 [\![\mu^4]\!]} + O(m^2) \tag{90}$$

$$= m \frac{\langle \delta^2 \rangle^2_{\boldsymbol{\theta}}}{\sigma^2 \langle \Delta^2 \rangle_{\boldsymbol{\theta}}} + O(m^2) \,, \tag{91}$$

Here we also inserted the definitions Eq. (84) and Eq. (87) □

Lemma 1. *For any ensemble average over $\mu \geq 0$ with $[\![\mu]\!] \neq 0$, the following statement holds*

$$\max_{x \geq 0} [\![\mu \sin(\mu x)]\!] \geq \sqrt{\frac{[\![\mu^2]\!]^3}{[\![\mu^4]\!]}} \,. \tag{92}$$

Proof. In order to prove that the maximum satisfies this inequality, it suffices to show that the inequality is satisfied at some point x^*. We choose to show this for $x^* = \frac{\pi}{2} \sqrt{\frac{[\![\mu^2]\!]}{[\![\mu^4]\!]}}$. To this end, we demonstrate that for $x \in [0, x^*]$

$$[\![\mu \sin(\mu x)]\!] \geq \sqrt{\frac{[\![\mu^2]\!]^3}{[\![\mu^4]\!]}} \sin\left(\sqrt{\frac{[\![\mu^4]\!]}{[\![\mu^2]\!]}} x \right) \tag{93}$$

holds, which gives the desired result for $x = x^*$. As the inequality is satisfied for $x = 0$, we will show the general case by verifying that the left-hand side has a larger derivative than the right-hand side everywhere in this interval which implies the original claim.

Therefore, it remains to show that

$$\frac{[\![\mu^2 \cos(\mu x)]\!]}{[\![\mu^2]\!]} \geq \cos\left(\sqrt{\frac{[\![\mu^4]\!]}{[\![\mu^2]\!]}} x\right) \geq 0\,. \tag{94}$$

By defining $p_k := \frac{\langle a_k^2 \rangle_{\boldsymbol{\theta}} \mu_k^2}{\sum_k \langle a_k^2 \rangle_{\boldsymbol{\theta}} \mu_k^2} = \frac{\langle a_k^2 \rangle_{\boldsymbol{\theta}} \mu_k^2}{[\![\mu^2]\!]}$ this simplifies to

$$\sum_t p_t \cos(\mu_t x) \geq \cos\left(\sqrt{\sum p_k \mu_k^2}\, x\right) \geq 0 \tag{95}$$

for $x \in \left(0, \frac{\pi}{2\sqrt{\sum_k p_k \mu_k^2}}\right]$. Further substituting $x' := x\sqrt{\sum p_k \mu_k^2}$ as well as $\mu_t' := \frac{\mu_t}{\sqrt{\sum p_k \mu_k^2}}$ yields

$$\sum_t p_t \cos(\mu_t' x') \geq \cos(x') \geq 0 \tag{96}$$

for $x' \in [0, \frac{\pi}{2}]$ with $\sum p_k \mu_k'^2 = 1$. This inequality needs to hold for all $\boldsymbol{p} > \mathbf{0}$, $\boldsymbol{\mu'} > \mathbf{0}$ with $\sum_k p_k = 1$. In the interest of readability, we are going to drop the primes again from here on out.

To prove this final inequality, we reformulate the left-hand side as a minimization problem

$$F(\boldsymbol{\mu}) = \sum_t p_t \cos(\mu_t x) + \frac{x^2}{2}\lambda\left(\sum_t p_t \mu_t^2 - 1\right) \tag{97}$$

with Lagrange multiplier $\frac{x^2}{2}\lambda$ for the constraint $\sum_t p_t \mu_t^2 = 1$. This gives the partial derivatives

$$\frac{\partial F}{\partial \mu_t} = -p_t x \sin(\mu_t x) + \lambda x^2 p_t \mu_t\,, \tag{98}$$

$$\frac{\partial^2 F}{\partial \mu_t^2} = -p_t x^2 \cos(\mu_t x) + \lambda x^2 p_t\,, \tag{99}$$

meaning that an extremal point satisfies

$$\lambda = \frac{x \sum_t p_t \sin(\mu_t x)}{x \sum_t p_t \mu_t x} < 1\,. \tag{100}$$

as well as either

$$\mu_t = 0, \text{ or} \tag{101}$$

$$\operatorname{sinc}(\mu_t x) = \lambda\,. \tag{102}$$

Since $\left.\frac{\partial^2 F}{\partial \mu_t^2}\right|_{\mu_t = 0} = p_t x^2(\lambda - 1)$ is negative, $\mu_t = 0$ does not describe a local minimum. Therefore, $\mu_t > 0$ holds. Since $\sum_t p_t \mu_t^2 = 1$ and $\sum_t p_t = 1$, there exists a $\mu_\alpha \leq 1$, meaning $\lambda = \operatorname{sinc}(\mu_\alpha x) \geq \operatorname{sinc}(\frac{\pi}{2}) = \frac{2}{\pi}$, which implies that all μ_t have the same value since sinc is injective on the considered interval. With the constraint, this yields $\mu_t \equiv 1$.

The second derivative at this point is

$$\left.\frac{\partial^2 F}{\partial \mu_t^2}\right|_{\boldsymbol{\mu}=\mathbf{1}} = p_t x^2(\lambda - \cos(x)) = p_t(x \sin(x) - x^2 \cos(x)) > 0 \tag{103}$$

for $0 < x \leq \frac{\pi}{2}$, which - as the Hessian is diagonal - means that $\boldsymbol{\mu} = \mathbf{1}$ is indeed the global minimum of $F(\boldsymbol{\mu})$ in the considered interval.

Since this minimal $\boldsymbol{\mu}$ satisfies Eq. (96), this concludes the proof of the lemma. □

B.5 Proof of theorem 4 and corollary 1

We derive a bound on the optimal $\frac{\Omega_S^2}{\Omega_{UB}^2}$. This expression can be written as

$$\frac{\Omega_S^2}{\Omega_{UB}^2} = \frac{[\![\mu \sin(\mu x)]\!]^2}{[\![\mu^2]\!]\left([\![\sin^2(\mu x)]\!] + \frac{\sigma^2}{m}\right)} \times \frac{[\![\mu^2]\!] + \frac{\nu^2\sigma^2}{m}}{[\![\mu^2]\!]} \tag{104}$$

$$= \frac{\frac{[\![\mu \sin(\mu x)]\!]^2}{[\![\mu^2]\!]^2}\left(1 + \frac{\nu^2\sigma^2}{m[\![\mu^2]\!]}\right)}{\frac{[\![\sin^2(\mu x)]\!]}{[\![\mu^2]\!]} + \frac{\sigma^2}{m[\![\mu^2]\!]}}\,. \tag{105}$$

By defining a probability vector $p_k = \langle a_k^2 \rangle_{\boldsymbol{\theta}} \frac{\mu_k^2}{[\mu^2]}$ and $\alpha = \frac{\nu^2 \sigma^2}{m[\mu^2]}$ we get

$$\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2} = \frac{\left(\sum_k p_k \frac{\sin(x\mu_k)}{\mu_k}\right)^2 (1+\alpha)}{\sum_k p_k \left(\frac{\sin(x\mu_k)}{\mu_k}\right)^2 + \frac{\alpha}{\nu^2}} \tag{106}$$

$$= \frac{\left(\sum_k p_k \frac{\nu \sin(x\mu_k)}{\mu_k}\right)^2 (1+\alpha)}{\sum_k p_k \left(\frac{\nu \sin(x\mu_k)}{\mu_k}\right)^2 + \alpha} \tag{107}$$

by substituting $\mu_k \to \mu_k \nu$ and $x \to x\nu$, we have the requirement that $\alpha \geq 0$ and $\mu_k \in [0,1]$. and

$$\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2} = \frac{\left(\sum_k p_k \frac{\sin(x\mu_k)}{\mu_k}\right)^2 (1+\alpha)}{\sum_k p_k \left(\frac{\sin(x\mu_k)}{\mu_k}\right)^2 + \alpha} \tag{108}$$

$$\tag{109}$$

or written as a problem we want to find the distribution for the worst relative correlation ratio using the best SLGE strategy.

$$\left(\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2}\right)^* = \min_{\substack{\alpha \geq 0,\, \boldsymbol{\mu} \in [0,1]^{n_\mu} \\ \boldsymbol{p} \in [0,1]^{n_\mu},\, \sum_i p_i = 1}} \left(\max_{x(\alpha, \boldsymbol{\mu}, \boldsymbol{p})} \left(\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2} \right) \right) \tag{110}$$

$$\geq \min_{\alpha \geq 0} \left(\max_{x(\alpha)} \left(\min_{\substack{\boldsymbol{\mu} \in [0,1]^{n_\mu} \\ \boldsymbol{p} \in [0,1]^{n_\mu},\, \sum_i p_i = 1}} \left(\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2} \right) \right) \right), \tag{111}$$

where we changed to order of optimization to find a lower bound. To solve the innermost bracket, we observe that by defining $\tau(\mu) = \frac{\sin(x\mu)}{\mu}$ as a random variable that $\sum_k p_k \frac{\sin(x\mu_k)}{\mu_k} = \langle \tau \rangle_{\boldsymbol{p}}$ and $\sum_k p_k \left(\frac{\sin(x\mu_k)}{\mu_k}\right)^2 = \langle \tau^2 \rangle_{\boldsymbol{p}}$ are the first and second moment of τ with $\langle \cdot \rangle_{\boldsymbol{p}}$ the expectation value of the distribution spanned by $\boldsymbol{p}$. This means for the expression

$$\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2} = \frac{\langle \tau \rangle_{\boldsymbol{p}}^2 (1+\alpha)}{\langle \tau^2 \rangle_{\boldsymbol{p}} + \alpha} \tag{112}$$

$$\tag{113}$$

We also note that if we restrict $x \in [0, \pi/2]$, $\tau \in [\sin(x), x]$, meaning τ is a bounded variable. We note that for a given expectation value $\langle \tau \rangle_{\boldsymbol{p}}$, the expression is minimized for a distribution with the largest possible second moment. This is described by a distribution only on the boundary of the parameter space.

$$\langle \tau \rangle_{\boldsymbol{p}} = q \sin(x) + (1-q)x \tag{114}$$

$$\langle \tau^2 \rangle_{\boldsymbol{p}} = q \sin^2(x) + (1-q)x^2 \tag{115}$$

for $q \in [0,1]$.

$$\left(\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2}\right) \geq \min_{\alpha > 0} \max_{x(\alpha) \in [0, \pi/2]} \min_{q \in [0,1]} \frac{(q \sin(x) + (1-q)x)^2 (1+\alpha)}{q \sin^2(x) + (1-q)x^2 + \alpha} \tag{116}$$

if this expression is minimized for q, one obtains

$$\left(\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2}\right)^* (\alpha, x) \geq \begin{cases} 4(1+\alpha)\frac{x \sin(x) - \alpha}{(x + \sin(x))^2} & 2\alpha \leq \sin(x)x - \sin^2(x) \\ \frac{\sin^2(x)(1+\alpha)}{\alpha + \sin^2(x)} & \text{else} \end{cases}, \tag{117}$$

where the latter is always at least 1 for $x = \pi/2$. Meaning that for $\alpha \geq \frac{\pi - 2}{4}$, Ω_{S} is always at least as large as Ω_{UB}. For $\alpha < \frac{\pi-2}{4}$, we find a lower bound by choosing $x(\alpha) = 2\alpha + 1$, where $x \in [1, \frac{\pi}{2}]$. This returns

$$\left(\frac{\Omega_{\mathrm{S}}^2}{\Omega_{\mathrm{UB}}^2}\right)^* \geq \min_{\alpha \in [0, \frac{\pi-2}{4}]} 4(1+\alpha)\frac{(2\alpha+1)\sin(2\alpha+1) - \alpha}{(2\alpha + 1 + \sin(2\alpha+1))^2} \geq 0.984\,, \tag{118}$$

which means $\Omega_S \geq 0.99\Omega_{UB}$ showing theorem 4. □

To proof ccorollary 1 we use Eq. (117), but keep a fixed $x = \pi/2$. This corresponds to a simple central differences strategy at $x = \frac{\pi}{2r}$ in the original case. Here we get

$$\left(\frac{\Omega_S^2}{\Omega_{UB}^2}\right) \geq 4(1+\alpha)\frac{\pi/2-\alpha}{(\pi/2+1)^2} \geq 2\frac{\pi}{(\pi/2+1)^2} \geq 0.95 \tag{119}$$

regardless of the underlying distribution $(\mathcal{D}_\theta, \sigma)$. □

C Barren plateau and 2-design calculations

To find estimates of the size of the Fourier coefficients, we need to estimate $\langle c_{ij}c_{kl}^*\rangle_{\boldsymbol{\theta}}$ given by

$$\langle c_{ij}c_{kl}^*\rangle_{\boldsymbol{\theta}} = \int \langle\Psi|U^\dagger P_i V^\dagger O V P_j U|\Psi\rangle\langle\Psi|U^\dagger P_l V^\dagger O V P_k U|\Psi\rangle \mathrm{d}V\,\mathrm{d}U\,. \tag{120}$$

Here U describes the ensemble of unitaries that are applied in the VQA before the layer of interest, V the unitaries after the layers, but before the measurement. We assume that both U and V describe 2-designs. This is useful because it allows us to use the identity for Haar random unitaries

$$\int U^\dagger A U \rho U^\dagger B U \mathrm{d}U = \frac{\mathbb{1}\,\mathrm{Tr}(\rho)}{d}\left(\frac{d\,\mathrm{Tr}(AB)}{d^2-1} - \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d^2-1}\right) + \rho\left(\frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d^2-1} - \frac{\mathrm{Tr}(AB)}{d(d^2-1)}\right), \tag{121}$$

where d is the Hilbert space dimension. We set

$$\rho_{ij} = P_i U|\Psi\rangle\langle\Psi|U^\dagger P_j \tag{122}$$

and use the identity Eq. (121) to obtain

$$\begin{aligned}
\langle c_{ij}c_{kl}^*\rangle_{\boldsymbol{\theta}} &= \int \mathrm{Tr}[V^\dagger O V \rho_{jl} V^\dagger O V\, \rho_{ki}]\,\mathrm{d}V\,\mathrm{d}U && (123)\\
&= \frac{1}{d^2-1}\int \mathrm{Tr}\left[\frac{\mathbb{1}\,\mathrm{Tr}[\rho_{jl}]}{d}\left(d\,\mathrm{Tr}[O^2] - \mathrm{Tr}[O]^2\right)\rho_{ki} + \rho_{jl}\left(\mathrm{Tr}[O]^2 - \mathrm{Tr}[O^2]/d\right)\rho_{ki}\right]\mathrm{d}U && (124)\\
&= \frac{d\,\mathrm{Tr}[O^2] - \mathrm{Tr}[O]^2}{d(d^2-1)}\int \mathrm{Tr}[\rho_{ki}]\,\mathrm{Tr}[\rho_{jl}]\,\mathrm{d}U + \frac{\mathrm{Tr}[O]^2 - \mathrm{Tr}[O^2]/d}{d^2-1}\int \mathrm{Tr}[\rho_{ki}\rho_{jl}]\,\mathrm{d}U && (125)
\end{aligned}$$

For the relevant terms, i.e. the ones with $i \neq j$, it follows that $\mathrm{Tr}[\rho_{ki}\rho_{jl}] = 0$. For the first term to not vanish, we require $i = k$ and $j = l$. With $\mathrm{Tr}[|\Psi\rangle\langle\Psi|] = 1$ and $\mathrm{Tr}[P_iP_j] = 0$,

$$\begin{aligned}
\int \mathrm{Tr}[\rho_{ii}]\,\mathrm{Tr}[\rho_{jj}]\,\mathrm{d}U &= \int \langle\Psi|U^\dagger P_i U|\Psi\rangle\langle\Psi|U^\dagger P_j U|\Psi\rangle\,\mathrm{d}U && (126)\\
&= \frac{1}{d}\left(\frac{d\,\mathrm{Tr}[P_iP_j]}{d^2-1} - \frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d^2-1}\right) + \left(\frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d^2-1} - \frac{\mathrm{Tr}[P_iP_j]}{d(d^2-1)}\right) && (127)\\
&= \frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d^2-1}\left(1-\frac{1}{d}\right) && (128)\\
&= \frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d(d+1)}\,. && (129)
\end{aligned}$$

This leads to

$$\begin{aligned}
\langle|c_{ij}|^2\rangle_{\mathcal{D}} &= \frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d(d+1)}\left(\frac{d\,\mathrm{Tr}[O^2] - \mathrm{Tr}[O]^2}{d(d^2-1)}\right) && (130)\\
&= \frac{1}{d}\frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d^2}\left(\mathrm{Tr}[O^2/d] - \mathrm{Tr}[O/d]^2\right)\frac{d^3}{(d+1)(d^2-1)} && (131)\\
&= \frac{\xi_d}{d}\frac{\mathrm{Tr}[P_i]\,\mathrm{Tr}[P_j]}{d^2}\left(\mathrm{Tr}[O^2/d] - \mathrm{Tr}[O/d]^2\right) && (132)\\
&= \frac{\xi_d}{d}\,\mathrm{Tr}[P_i/d]\,\mathrm{Tr}[P_j/d]\sigma_O^2 && (133)\\
& && (134)
\end{aligned}$$

with $\xi_d := \frac{d^3}{(d+1)(d^2-1)}$, $\sigma_O^2 := \mathrm{Tr}[O^2/d] - \mathrm{Tr}[O/d]^2$ being the expected variance with respect to the maximally mixed state and $\mathrm{Tr}[P_\ell]$ is the multiplicity of the eigenvalue λ_ℓ. The final estimate is therefore

$$\langle a_k^2 \rangle_{\boldsymbol{\theta}} = \sum_{i \geq j : \mu_k = \lambda_i - \lambda_j} \langle |c_{ij}|^2 \rangle_{\mathcal{D}} \tag{135}$$

$$= \xi_d \frac{\sigma_O^2}{d} \sum_{i \geq j : \mu_k = \lambda_i - \lambda_j} \mathrm{Tr}[P_i/d] \, \mathrm{Tr}[P_j/d] \tag{136}$$

Similarly, one can obtain a shot noise estimate when measuring in the Hamiltonian eigenbasis by the one design property

$$\sigma^2 = \int \langle \Psi | U^\dagger O^2 U | \Psi \rangle \mathrm{d}U - \left(\int \langle \Psi | U^\dagger O U | \Psi \rangle \mathrm{d}U \right)^2 \tag{137}$$

$$= \mathrm{Tr}[O^2/d] - \mathrm{Tr}[O/d]^2 = \sigma_O^2 \,. \tag{138}$$

We note that when O is not directly measured in its eigenbasis, the real shot noise variance might be significantly large, as typically multiple different measurement setting are required.

C.1 Twirls of linear maps on operators with unitary 2-design

Given a linear map M on the vector space of operators and a probability distribution on the unitary group one can define the *twirl* of M as T where

$$T(X) := \mathbb{E}[U^\dagger M(UXU^\dagger)U] \,. \tag{139}$$

The distribution of unitaries is called a *unitary 2-design* if the expectation value above yields the same as the similar one for the Haar measure (see, e.g. the tutorial [31] for details). In this case, T satisfies the invariance condition $T(X) = U^\dagger T(UXU^\dagger)U$ for all operators X and the expectation value can be evaluated as in the following lemma,

Lemma 2 (Twirl of maps on operators [32, Appendix]). *Let $T : \mathrm{L}(\mathbb{C}^d) \to \mathrm{L}(\mathbb{C}^d)$ be a linear map that satisfies the invariance $T(X) = U^\dagger T(UXU^\dagger)U$ for all $X \in \mathrm{L}(\mathbb{C}^d)$ and unitaries $U \in \mathrm{U}(d)$. Then*

$$T(X) = \frac{\mathrm{Tr}[T(\mathbb{1})] - \mathrm{Tr}[T]/d}{d^2-1} \mathrm{Tr}[X] \, \mathbb{1} + \frac{\mathrm{Tr}[T] - \mathrm{Tr}[T(\mathbb{1})]/d}{d^2-1} X \,. \tag{140}$$

In particular,

$$\int U^\dagger A U \rho U^\dagger B U \mathrm{d}U = \frac{\mathbb{1}\,\mathrm{Tr}(\rho)}{d} \left(\frac{d\,\mathrm{Tr}(AB)}{d^2-1} - \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d^2-1} \right) + \rho \left(\frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d^2-1} - \frac{\mathrm{Tr}(AB)}{d(d^2-1)} \right) . \tag{141}$$

Proof. We denote the identity map by id and the completely dephasing channel by $R(X) := \mathrm{Tr}[X]\,\mathbb{1}/d$. Due to a standard argument [32, Appendix] relying on Schur's lemma one can write

$$T = aR + b\,\mathrm{id} \tag{142}$$

for some coefficients $a, b \in \mathbb{C}$. Taking the trace of this equation and of $T(\mathbb{1}) = aR(\mathbb{1}) + b\,\mathrm{id}(\mathbb{1})$ results in

$$\mathrm{Tr}[T] = a + bd^2 \,, \tag{143}$$

$$\mathrm{Tr}[T(\mathbb{1})] = ad + bd \,. \tag{144}$$

Solving for a and b and inserting these coefficients in the ansatz Eq. (142) yields the statement Eq. (140).

Next, we take T to be the RHS of Eq. Eq. (141). Then T satisfies the lemma's invariance condition and one can show that $\mathrm{Tr}[T] = \mathrm{Tr}[A]\,\mathrm{Tr}[B]$ and $\mathrm{Tr}[T(\mathbb{1})] = \mathrm{Tr}[AB]$, which results in Eq. (141).

See also [31, Theorem 51] for more explicit calculations relying on the second moment operator of a unitary 2-design. □

D Analytical estimates for the Fourier coefficients of the QAOA Hamiltonian

As is mentioned in the main text, to obtain quantitative results for correlation matrix of the QAOA, we are considering the expectation value over all graph instances

$$\langle \cdot \rangle_{s,\boldsymbol{\theta}} \to \langle \cdot \rangle_{s,\boldsymbol{\theta},G} \,. \tag{145}$$

the ensemble is randomly chosen graphs with N vertices and M edges.

D.1 Case for $L \to \infty$

For $L \to \infty$ using the derived quantity for the 2-design Eq. (23), we need to calculate the expression

$$\langle a_k^2 \rangle_{\boldsymbol{\theta}} = \langle \frac{\sigma_O^2}{d} \sum_{i \geq j : \mu_k = \lambda_i - \lambda_j} \mathrm{Tr}[P_i/d] \, \mathrm{Tr}[P_j/d] \rangle_G \tag{146}$$

over all graph instances. Since $O = H_c = \frac{1}{2}\sum_{(i,j)\in E}(Z_iZ_j - \mathbb{1})$, It follows with $\mathrm{Tr}(Z_iZ_j) = 0$ that $\mathrm{Tr}(O/d) = -M/2$. Equally, by only counting the terms which are proportional to the identity, we get $\mathrm{Tr}(O^2/d) = \frac{M^2}{4} + \frac{M}{4}$, meaning $\sigma_O^2 = \frac{M}{4}$. This quantity is independent of the particular graph taken from the distribution.

We note that the VQA is invariant under the parity symmetry in the X-basis

$$\Pi_x = \sigma_x \otimes \cdots \otimes \sigma_x \tag{147}$$

the state only lives in the even number of ones sector, which also halves the Hilbert space dimension $d = 2^{N-1}$.

To find $\langle a_k^2 \rangle_{\boldsymbol{\theta}}\big|_b$, we first note that H_b is independent of the particular graph instance chosen. The eigenvalues of H_b describe spin sectors with $\lambda_i = -\frac{N}{2} + i$ and $\mathrm{Tr}(P_i^B) = \binom{N}{i}$. Introducing the symmetry reduced subspace means that

$$\mathrm{Tr}(P_i^b) = \begin{cases} \binom{N}{i}, & i \text{ even} \\ 0, & \text{else} \end{cases}, \tag{148}$$

which gives the final expression

$$\langle a_k^2 \rangle_{\boldsymbol{\theta}}\big|_b = \frac{M}{4 \cdot 8^{N-1}} \sum_{j \in \{0,2,\ldots N-k\}} \binom{N}{j+k}\binom{N}{j} \tag{149}$$

$$= \frac{M}{4 \cdot 8^{N-1}} \frac{(2N)!}{(N-k)!(N+k)!}. \tag{150}$$

For H_c, determining the expression requires to take a distribution over the graph instances. Qualitatively, it would already be sufficient to calculate $\langle \mathrm{Tr}(P_i^c) \rangle_G$ individually, but for the sake of completeness we will derive the full correlation between the two terms.

The relevant expression is

$$\zeta_k = \sum_{i=0}^{M} \langle \mathrm{Tr}(P_k^c) \, \mathrm{Tr}(P_{k+i}^c) \rangle_G / d^2, \tag{151}$$

which can be calculated by solving a combinatorial problem. For this, we reformulate the question as asking what the likelihood is that for two different partitions the fist cuts k edges and the second $k+j$ edges. This is because by design $\mathrm{Tr}(P_k^c)/d$, describes the likelihood of a random partition having k edges cut. To find this quantity, we classify the vertices according to the labels they receive in the to different bipartitions. This creates the four sectors $\{(0,0),(0,1),(1,0),(1,1)\}$. The number of vertices in each sector is $\boldsymbol{s} = (s_{00}, s_{01}, s_{10}, s_{11})$ with the probability for a particular $\boldsymbol{s}$ being

$$p(\boldsymbol{s}) = \binom{n}{s_{00}, s_{01}, s_{10}, s_{11}} \frac{1}{4^n}, \tag{152}$$

as the process of choosing vertex labels is described by a multinomial distribution where each sector has probability $p = 1/4$.

Next, we sample the edges: From $|\Gamma| = \binom{n}{2}$ possible edges, $E_1 = s_{00}s_{10} + s_{01}s_{11}$ edges are cut only by the first partition and $E_2 = s_{00}s_{01} + s_{10}s_{11}$ are cut only by the second. We do not consider the cases where an edge is cut by neither or both partitions as this will not affect the overall result. As the particular edges are chosen randomly from Γ, the process is described by a multi-hypergeometric distribution

$$\zeta_k = \sum_{\boldsymbol{s} \in [n]^4, i \in [M]} p(\boldsymbol{s}) \frac{\binom{E_1(\boldsymbol{s})}{i}\binom{E_2(\boldsymbol{s})}{i+k}\binom{|\Gamma| - E_1(\boldsymbol{s}) - E_2(\boldsymbol{s})}{M-2i-k}}{\binom{|\Gamma|}{M}}. \tag{153}$$

This yields

$$\langle \langle a_k^2 \rangle_{\boldsymbol{\theta}} \rangle_G\big|_c = \frac{M}{2^{N-3}} \zeta_k \tag{154}$$

as the final result for H_c.

D.2 Case for one layer $L = 1$

In the following section, we derive a single layer ($L = 1$) estimate by relating the underlying problem to a graph problem. We note that the rigorous analysis performed here is excessive because, for practical purposes simply sampling the derived graph problem for the particular graph instance will already give a reliable estimate. However, for completeness, we derive exact analytic results for the QAOA. We begin by defining three observables

$$O_{zz} = \frac{1}{2}\sum_{ij\in E} Z_i Z_j \tag{155}$$

$$O_{yy} = \frac{1}{2}\sum_{ij\in E} Y_i Y_j \tag{156}$$

$$O_{yz} = \frac{1}{2}\sum_{ij\in E} Y_i Z_j + Z_i Y_j \,. \tag{157}$$

Note that the derivative is not affected by replacing $H_c \to O_{zz}$ as this only describes a constant offset. The cost function for the first layer can be subdivided into three parts according to

$$F(\gamma,\beta) = \frac{1}{2}(1+\cos(2\beta))C_{zz}(\gamma) + \frac{1}{2}(1-\cos(2\beta))C_{yy}(\gamma) + \frac{1}{2}\sin(2\beta)C_{yz}(\gamma)\,, \tag{158}$$

where

$$C_\rho(\gamma) = \sum_{ij}\langle\Psi_i|O_\rho|\Psi_j\rangle e^{-i\gamma(\lambda_i-\lambda_j)} \tag{159}$$

$$= \sum_{k=0}^{M}\left(\sum_{i=0}^{M-k}\langle\Psi_i|O_\rho|\Psi_{i+k}\rangle e^{ik\gamma}\right) \tag{160}$$

$$=: \sum_{k=0}^{M} C_\rho^k(\gamma) \tag{161}$$

with $\rho \in \{zz, yy, yz\}$ and $|\Psi_j\rangle = P_j^c|+\rangle$. We can quickly check that

$$C_{zz}(\gamma) = \sum_{ij}\langle\Psi_i|O_{zz}|\Psi_j\rangle e^{-i\gamma(\lambda_i-\lambda_j)} \tag{162}$$

$$= \sum_i \langle\Psi_i|O_{zz}|\Psi_i\rangle = \sum_i \langle+|P_i O_{zz} P_i|+\rangle = \langle+|O_{zz}|+\rangle = 0\,, \tag{163}$$

which means the expression simplifies to

$$F(\gamma,\beta) = \frac{1}{2}(1-\cos(2\beta))C_{yy}(\gamma) + \frac{1}{2}\sin(2\beta)C_{yz}(\gamma)\,. \tag{164}$$

The evolution of H_b only has one non-vanishing frequency $\langle a_2^2\rangle_{\boldsymbol{\theta}}$. By integrating over β and γ w.r.t. the correct frequency we get

$$\langle a_2^2\rangle_{\boldsymbol{\theta}}\big|_b = \frac{\langle|C_{yy}|^2\rangle_\gamma + \langle|C_{yz}|^2\rangle_\gamma}{8} = \sum_{k=0}^{M}\frac{\langle|C_{yy}^k|^2\rangle_\gamma + \langle|C_{yz}^k|^2\rangle_\gamma}{8} \tag{165}$$

and

$$\langle a_k^2\rangle_{\boldsymbol{\theta}}\big|_c = \sum_{k=0}^{M}\frac{3\langle|C_{yy}^k|^2\rangle_\gamma + \langle|C_{yz}^k|^2\rangle_\gamma}{8}\,, \tag{166}$$

where

$$\langle|C_\rho^k|^2\rangle_\gamma = \sum_{i=0}^{M}\langle\Psi_i|O_\rho|\Psi_{i+k}\rangle \times \sum_{j=0}^{M}\langle\Psi_{j+k}|O_\rho|\Psi_j\rangle\,. \tag{167}$$

In the remainder of this section, we show how Eq. (167) can be interpreted as a graph problem and how to calculate the expression for general graphs with N vertices and M edges. Similar to appendix D.1,we interpret a computational basis state a particular bi-partition of vertices with every qubit describing one vertex. We also have to consider two partitions with the four cases labeled $\boldsymbol{s} = (s_{00}, s_{01}, s_{10}, s_{11})$ with the same probability as in Eq. (152). O_ρ describes a certain graph operation which will be performed on both partitions. Only if the number of edges cut changes by k in both partitions, will this partition pair contribute to $\langle |C_\rho^k|^2 \rangle_\gamma$.

For the action of O_ρ, with $\rho \in \{yy, yz\}$, we use $Y = -iXZ$ to rewrite $YY = -XX \times ZZ$ and $ZY + YZ = -i(\mathbb{1}X + \mathbb{1}X) \times ZZ$. Effectively, X corresponds to moving a vertex to the other side of the respective bi-partition and ZZ accumulates a sign if the edge vertices are on opposing sides. The overall sign is therefore determined by the rule

$$\text{sign}(E_1, E_2) = \begin{cases} + & \text{Both edges cut their partition or both do not.} \\ - & \text{One edges cuts its partition the other does not.} \end{cases} . \tag{168}$$

Each operator O_ρ sums over all edges which gives M^2 different edge pairs to consider. There are $\omega_{\text{same}} = M$ cases where the two edges considered are the same edge in the graph. When this is not the case, we will separate into additional cases depending on whether the two edges share a vertex or not. This gives a total of three cases:

1. The edges are the same edge: $\omega_{\text{same}} = M$
2. The edges have a common vertex: $\omega_{\text{con}} = (M^2 - M)\frac{2(N-2)}{\binom{M}{2}-1}$, where $2(N-2)$ is the number of possible edges that connect to the first edge.
3. The edges have no overlapping vertex: $\omega_{\text{sep}} = M^2 - \omega_{\text{same}} - \omega_{\text{con}}$.

We then proceed to draw the corresponding vertices depending on each of the three cases, which means drawing $2, 3$ or 4 vertices, respectively. The probability of choosing an individual vertex from each set is

$$q_{\boldsymbol{s}}(v_1 = (i,j)) = \frac{s_{ij}}{\|\boldsymbol{s}\|_1} \quad \text{for the first vertex, and} \tag{169}$$

$$q_{\boldsymbol{s}}(v_l = (i,j)) = \frac{\bar{s}_{ij}}{\|\bar{\boldsymbol{s}}\|_1}, \quad \text{where} \quad \bar{\boldsymbol{s}} = \boldsymbol{s}_{\setminus\{v_1,\dots,v_{l-1}\}} . \tag{170}$$

$s_{\setminus V}$ is the sector distribution of vertices without the vertices in the set V. Additionally, if $\rho = yz$, we also need to decide on which vertex the Y operation acts. For this we place each vertex into the edge sets $\boldsymbol{E} = (E_1^Y, E_1^Z, E_2^Y, E_2^Z) \in S_{\rho,\xi}$ which encapsulates which operation operation is performed on them according to the correct (ρ, ξ). If a vertex is shared, it enters in this edge set twice. We use the smaller indices for the first edge and the shared edges also start from v_1. $S_{\rho,\xi}$ describes all legal operations, for instance

$$S_{yy,\text{same}} = \{(\{v_1, v_2\}, \emptyset, \{v_1, v_2\}, \emptyset)\} \tag{171}$$

$$S_{yz,\text{con}} = \{(\{v_1\}, \{v_2\}, \{v_1\}, \{v_3\}), (\{v_2\}, \{v_1\}, \{v_1\}, \{v_3\}), (\{v_1\}, \{v_2\}, \{v_3\}, \{v_1\}), (\{v_2\}, \{v_1\}, \{v_3\}, \{v_1\})\} . \tag{172}$$

Now, we are able to proceed to draw the remaining edges of the graph. Notably, they only contribute to $\langle |C_\rho^k|^2 \rangle_\gamma$ if they are connected to the drawn vertices with an X operation. To summarize, the steps are:

1. Select $\rho \in \{yy, yz\}$.
2. Select a bipartitions $\boldsymbol{s}$ with the probability $p(\boldsymbol{s})$ as defined in Eq. (152).
3. Select the relationship of the edges to each other $\xi \in (\text{same}, \text{con}, \text{sep})$.
4. Select $2-4$ vertices from $\boldsymbol{s}$ according to the particular case ξ.
5. Select a vertex distribution according to the correct edge set $\boldsymbol{E} = (E_1^Y, E_1^Z, E_2^Y, E_2^Z) \in S_{\rho,\xi}$.
6. Draw the remaining edges of the graph from a hyper-geometric distribution to calculate the effect on the vertices cut.

The final equation is given as

$$\langle |C_\rho^k|^2 \rangle_{\boldsymbol{\theta}} = \sum_{\substack{\boldsymbol{s} \in \{0,\dots,N\}^4 | \sum s_i = N \\ \xi \in \{\text{same},\text{con},\text{sep}\} \\ \boldsymbol{v} \in \{0,1\}^{2 \times n_\xi} \\ \boldsymbol{E} \in S_{\rho,\xi}}} p(\boldsymbol{s}) \omega_\xi q_{\boldsymbol{s}}(v_1) \cdots q_{\bar{\boldsymbol{s}}}(v_{n_\xi}) \text{sign}(E_1, E_2) p(k|\boldsymbol{s}, \boldsymbol{E}) , \tag{173}$$

where $p(k|\boldsymbol{s}, \boldsymbol{E})$ is the probability of the changes on both bipartitions being k.

There are guaranteed changes which we label with $\boldsymbol{F}(\boldsymbol{E})$, which arise from the effect of the already drawn edges. This is only relevant if the edges share vertices on which one X operation operates, meaning $\xi = \text{con}$

and if $\rho = yz$ also $\xi = \text{same}$. The sign is determined by whether or not the other edge is cut in the bi-partition. This can be calculated to be

$$F_1(\boldsymbol{E}) = \delta_{|E_1^Y \cap E_2|,1}(-1)^{\sum_{w\in E_2} w_1} \tag{174}$$

$$F_2(\boldsymbol{E}) = \delta_{|E_1 \cap E_2^Y|,1}(-1)^{\sum_{w\in E_1} w_2}, \tag{175}$$

where w_i refers to the labeling for the i-th partition. The edges themselves are drawn from a hyper-geometric distribution. From the full edge set Γ with $|\Gamma| = \binom{N}{2}$ after the one/two edges are chosen Γ_r edges remain, with $|\Gamma_r| = \binom{N}{2} - 1$ for $\xi = \text{same}$ and $|\Gamma_r| = \binom{N}{2} - 2$ for the other cases. Similarly, the number of edges to select are $M_r = M - 1$ and $M_r = M - 2$. Any of these edges fits into one of 9 categories, either not affecting, increasing or decreasing the value of the cut after performing the operation for each of the partitions. For this we define $K_{\alpha,\beta}$ with $(\alpha,\beta) \in \{-1,0,1\}^2$. The number of edges drawn from each category is labeled $X_{\alpha,\beta}$. We can determine K with the following rules

$$K_{1-2m,0} = \sum_{i\in\{0,1\}, v\in E_1^Y \setminus E_2^Y} \bar{s}_{v_1\oplus m,i} \qquad \bar{\boldsymbol{s}} = \boldsymbol{s}_{\setminus\{E_1^Y \cup E_2^Y \cup E_1^Z\}} \tag{176}$$

$$K_{0,1-2m} = \sum_{i\in\{0,1\} v\in E_2^Y \setminus E_1^Y} \bar{s}_{i,v_2\oplus m} \qquad \bar{\boldsymbol{s}} = \boldsymbol{s}_{\setminus\{E_1^Y \cup E_2^Y \cup E_2^Z\}} \tag{177}$$

$$K_{1-2m,1-2l} = \sum_{v\in E_1^Y \cap E_2^Y} \bar{s}_{v\oplus(m,l)} + \sum_{v\in E_1^Y, w\in E_2^Y} \delta_{v_1\oplus w_2,(m,l)} \qquad \bar{\boldsymbol{s}} = \boldsymbol{s}_{\setminus\{E_1^Y \cup E_2^Y \cup E_1^Z \cup E_2^Z\}} \tag{178}$$

$$K_{0,0} = |\Gamma_r| - \sum_{\substack{\alpha,\beta\in\{-1,0,1\}^2 \\ (\alpha,\beta)\neq(0,0)}} K_{\alpha,\beta}\,. \tag{179}$$

Here $E_1 \cap E_2$ refers to vertices that are shared by the two edges. The corresponding probability distribution is then given by

$$g(k|\boldsymbol{s},\boldsymbol{E}) = g(k|\boldsymbol{K}(\boldsymbol{s},\boldsymbol{E}),\boldsymbol{F}(\boldsymbol{E})) = \frac{1}{\binom{|\Gamma_r|}{M_r}} \sum_{\substack{\boldsymbol{X}\in[\boldsymbol{K}] \\ \sum_{ij} X_{ij} = M_r \\ \sum_\beta (X_{+1,\beta} - X_{-1,\beta}) + F_1 = k \\ \sum_\alpha (X_{\alpha,+1} - X_{\alpha,-1}) + F_2 = k}} \prod_{(\alpha,\beta)\in\{-1,0,1\}^2} \binom{K_{\alpha,\beta}}{X_{\alpha,\beta}} \tag{180}$$

With this, we have all the ingredients to calculate $\langle\langle a_k^2\rangle_{\boldsymbol{\theta}}\rangle_G\big|_b$ and $\langle\langle a_k^2\rangle_{\boldsymbol{\theta}}\rangle_G\big|_c$, the result of which is plotted in figure 3.

D

Optimizing the depth of variational quantum algorithms is strongly QCMA-hard to approximate

Title:	Optimizing the depth of variational quantum algorithms is strongly QCMA-hard to approximate
Authors:	Lennart Bittel, Sevag Gharibian, Martin Kliesch
Conference:	Computational Complexity Conference (CCC)
Journal:	Theory of Computing (ToC)
Date of submission:	7 February 2023
Publication status:	Accepted (CCC), in review (ToC)

This publication corresponds to the article [88]. The summary of the results is presented in section 4.3.1.

Contribution: The article was joint work with SG and MK. I contributed significantly in all parts of the article, especially in the section for deriving the hardness of the QAOA instances.

The optimal depth of variational quantum algorithms is QCMA-hard to approximate

Lennart Bittel* Sevag Gharibian† Martin Kliesch‡

Abstract

Variational Quantum Algorithms (VQAs), such as the Quantum Approximate Optimization Algorithm (QAOA) of [Farhi, Goldstone, Gutmann, 2014], have seen intense study towards near-term applications on quantum hardware. A crucial parameter for VQAs is the *depth* of the variational "ansatz" used — the smaller the depth, the more amenable the ansatz is to near-term quantum hardware in that it gives the circuit a chance to be fully executed before the system decoheres. In this work, we show that approximating the optimal depth for a given VQA ansatz is intractable. Formally, we show that for any constant $\epsilon > 0$, it is QCMA-hard to approximate the optimal depth of a VQA ansatz within multiplicative factor $N^{1-\epsilon}$, for N denoting the encoding size of the VQA instance. (Here, Quantum Classical Merlin-Arthur (QCMA) is a quantum generalization of NP.) We then show that this hardness persists in the even "simpler" QAOA-type settings. To our knowledge, this yields the first natural QCMA-hard-to-approximate problems.

1 Introduction

In the current era of Noisy Intermediate Scale Quantum (NISQ) devices, quantum hardware is (as the name suggests) limited in size and ability. Thus, NISQ-era quantum algorithm design has largely focused on *hybrid* classical-quantum setups, which ask: What types of computational problems can a classical supercomputer, paired with a *low-depth* quantum computer, solve? This approach, typically called Variational Quantum Algorithms (VQA), has been studied intensively in recent years (see, e.g. [Cer+21; Bha+22] for reviews), with Farhi, Goldstone and Gutmann's Quantum Approximate Optimization Algorithm (QAOA) being a prominent example [FGG14].

More formally, VQAs roughly work as follows. One first chooses a variational ansatz (i.e. parameterization) over a family of quantum circuits. Then, one iterates the following two steps until a "suitably good" parameter setting is found:

1. Use a classical computer to optimize the ansatz parameters variationally[1].
2. Run the resulting parameterized quantum algorithm on a NISQ device to evaluate the "quality" of the chosen parameters (relative to the computational problem of interest).

The essential advantage of this setup over more traditional quantum algorithm design techniques (such as full Trotterization of a desired Hamiltonian evolution) is that one can attempt

*Institute for Theoretical Physics, Heinrich Heine University Düsseldorf, Germany. Email: lennart.bittel@uni-duesseldorf.de.

†Department of Computer Science, and Institute for Photonic Quantum Systems, Paderborn University, Germany. Email: sevag.gharibian@upb.de.

‡Institute for Theoretical Physics, Heinrich Heine University Düsseldorf, and Institute for Quantum-Inspired and Quantum Optimization, Hamburg University of Technology, Germany. Email: martin.kliesch@tuhh.de.

[1] In practice, this typically means heuristic optimization.

to minimize the *depth* of the ansatz used. (A formal definition of "depth" is given in Problem 1; briefly, it is the number of Hamiltonian evolutions the ansatz utilizes.) This possibility gives VQAs a potentially crucial advantage on near-term quantum hardware (i.e. noisy hardware without quantum error correction), because a NISQ device can, in principle, execute a low-depth ansatz before the system decoheres, i.e. before environmental noise destroys the "quantumness" of the computation. From an analytic perspective, low-depth ansatzes also have an important secondary benefit — VQAs of superlogarithmic depth are exceedingly difficult to analyze via worst-case complexity. Sufficiently low-depth setups, however, sometimes *can* be rigorously analyzed, with the groundbreaking QAOA work of [FGG14] for MAX-CUT being a well-known example. Thus, estimating the optimal depth for a variational quantum algorithm (VQA) appears central to its use in near-term applications.

1.1 Our results

In this work, we show that it is intractable to approximate the optimal depth for a given VQA ansatz, even within large multiplicative factors. Moreover, this hardness also holds for the restricted "simpler" case of the QAOA. To make our claim rigorous, we first define the VQA optimization problem we study. (Intuition to follow.)

Problem 1 (VQA minimization (MIN-VQA(k,l))). *For an n-qubit system:*

- *Input:*
  1. *Set $H = \{H_i\}$ of Hamiltonians[2], where H_i acts non-trivially only on a subset[3] $S_i \subseteq [n]$ of size $|S_i| = k$.*
  2. *An l-local observable M acting on a subset of l qubits.*
  3. *Integers $0 \leq m \leq m'$ representing circuit depth thresholds.*
- *Output:*
  1. *YES if there exists a list of at most m angles[4] $(\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$ and a list $(G_1, \ldots, G_m)$ of Hamiltonians from H (repetitions permitted) such that*

$$|\psi\rangle := e^{i\theta_m G_m} \cdots e^{i\theta_1 G_1} |0 \cdots 0\rangle \tag{1}$$

  satisfies $\langle\psi|M|\psi\rangle \leq 1/3$.
  2. *NO if for all lists of at most m' angles $(\theta_1, \ldots, \theta_{m'}) \in \mathbb{R}^{m'}$ and all lists $(G_1, \ldots, G_{m'})$ of Hamiltonians from H (repetitions permitted),*

$$|\psi\rangle := e^{i\theta_{m'} G_{m'}} \cdots e^{i\theta_1 G_1} |0 \cdots 0\rangle \tag{2}$$

  satisfies $\langle\psi|M|\psi\rangle \geq 2/3$.

For intuition, recall that a VQA ansatz is a parameterization over a family of quantum circuits. Above, the ansatz is parameterized by angles θ_j, and the family of quantum circuits is generated by Hamiltonians H_j. The aim is to pick a *minimum-length* sequence of Hamiltonian evolutions

[2] An n-qubit Hamiltonian H is a $2^n \times 2^n$ Hermitian matrix. Any unitary operation U on a quantum computer can be generated via an appropriate choice of Hamiltonian H and evolution time $t \geq 0$, i.e. $U = e^{iHt}$.

[3] For Theorem 1, it will suffice to take $k \in O(1)$. In principle, however, containment in QCMA holds for any $k \leq n$, so long as the H_i are sparse in the standard Hamiltonian simulation sense [AT03]. By sparse, one means that each row r of H_i contains at most r non-zero entries, which can be computed in poly-time given r.

[4] Throughout Problem 1, for clarity we assume all angles are specified to poly(n) bits.

$e^{i\theta_j G_j}$, so that the generated state $|\psi\rangle$ has (say) low overlap with the target observable, M. For clarity, throughout this work, by "depth" of a VQA ansatz, we are referring to the standard VQA notion of the number of Hamiltonian evolutions m applied[5]. (In the setting of QAOA, the "depth" is often referred to as the "level", up to a factor of 2.)

We remark for Problem 1 that we do not restrict the order in which Hamiltonians H_i are applied, and any H_i may be applied multiple times. Moreover, our results also hold if one defines the YES case to maximize overlap with M (as opposed to minimize overlap).

Our first result is the following.

Theorem 1. MIN-VQA(k, l) *is* QCMA*-complete for* $k \geq 4$, $l = 2$, *and* $m \leq \mathsf{poly}(n)$. *Moreover, for any* $\epsilon > 0$, *it is* QCMA*-hard to distinguish between the YES and NO cases of* MIN-VQA *even if* $m'/m \geq N^{1-\epsilon}$, *where* N *is the encoding size of the instance.*

Here, Quantum-Classical Merlin-Arthur (QCMA) is a quantum generalization of NP with a classical proof and quantum verifier (formal definition in Definition 1). For clarity, the *encoding size* of the instance is the number of bits required to write down a MIN-VQA instance, i.e. to encode $H = \{H_i\}$, M, m, m' (see Problem 1). Note the encoding size is typically dominated by the encoding size of H, which may be assumed to scale as $|H|$, i.e. with the number[6] of *interaction terms* H_i, which can be asymptotically larger than the number of qubits, n. Thus, simple gap amplification strategies such as taking many parallel copies of all interaction terms do *not* suffice to achieve our hardness ratio of $N^{1-\epsilon}$.

A direct consequence of Theorem 1 is that it is intractable (modulo the standard conjecture that BQP $\neq$ QCMA, which also implies P $\neq$ QCMA) to compute the optimum circuit depth within relative precision $N^{1-\epsilon}$ (proof given in Appendix A for completeness):

Corollary 2 (Depth minimization). *In Problem 1, let* m_{opt} *denote the minimum depth* m *such that* $\langle\psi|M|\psi\rangle \leq 1/3$. *Then, for any constant* $\epsilon > 0$, *computing estimate* $m_{\text{est}} \in [m_{\text{opt}}, N^{1-\epsilon} m_{\text{opt}}]$ *is* QCMA*-hard.*

On the other hand, even if a desired depth $m = m'$ is specified in advance, it is also QCMA-hard to find the minimizing angle and Hamiltonian sequences $(\theta_1, \ldots, \theta_m)$ and $(G_1, \ldots, G_m)$, respectively, which follows directly from Theorem 1:

Corollary 3 (Parameter optimization). *Consider Problem 1 with input* $m = m'$. *Then the problem of finding the angles* $(\theta_1, \ldots, \theta_m)$ *that minimize the expectation value* $\langle\psi|M|\psi\rangle$ *is* QCMA*-hard.*

We next turn to the special case of QAOAs. As detailed shortly under "Previous work", the study of QAOA ansatzes was initiated by [FGG14] in the context of *quantum* approximation algorithms for MAX CUT. In that work, a QAOA is analogous to a VQA, except there are only *two* Hamiltonians $H = \{H_b, H_c\}$ given as input and M is one of those two observables (see Problem 3 for a formal definition). For clarity, here we work with a more general definition of QAOA than [FGG14], in which neither H_b nor H_c need be diagonal in the standard basis. (In this sense, our definition is closer to the more general Quantum Alternating Operator Ansatz,

[5] Alternatively, one could consider the *circuit depth* of any simulation of the desired Hamiltonian sequence in Problem 1. The downside of this is that it would be much more difficult to analyze — one would presumably first need to convert each $e^{i\theta_j G_j}$ to a circuit U_j via a fixed choice of Hamiltonian simulation algorithm. One would then need to characterize the depth of the concatenated circuit $U_m \cdots U_1$.

[6] Indeed, in the construction in the proof of Theorem 1, $N \in O(|H|)$.

also with acronym QAOA [Had+19].) For our hardness results, it will suffice for H_b and H_c to be k-*local* Hamiltonians[7]. For QAOA, we show a matching hardness result:

Theorem 4. MIN-QAOA(k) *is* QCMA*-complete for* $k \geq 4$ *and* $m \leq \mathsf{poly}(n)$. *Moreover, for any* $\epsilon > 0$, *it is* QCMA*-hard to distinguish between the YES and NO cases of* MIN-QAOA *even if* $m'/m \geq N^{1-\epsilon}$, *where* N *is the number of strictly* k*-local terms comprising* H_b *and* H_c.

Note that in contrast to MIN-VQA, which is parameterized by k (the Hamiltonians' locality) and l (the observable's locality), MIN-QAOA is only parameterized by k. This is because in QAOA, the "cost" Hamiltonian H_c *itself* acts as the observable (in addition to helping drive the computation), which will be one of the obstacles we will need to overcome. For context, typically in applications of QAOA, H_c encodes (for example [FGG14]) a MAX CUT instance.

To the best of our knowledge, Theorem 1 and Theorem 4 yield the first natural QCMA-hard to approximate problems.

1.2 Previous work

Generally speaking, it is well-known that VQA parameters are "hard to optimize", both numerically and from a theoretical perspective. We now discuss selected works from the (vast) VQA literature, and clarify how these differ from our work.

1. Theoretical studies. As previously mentioned, in 2014, Farhi, Goldstone and Gutmann proposed the Quantum Approximate Optimization Algorithm (QAOA), a special case of VQA with only two local Hamiltonians $H = \{H_b, H_c\}$ (acting on n qubits each). They showed that level-1 of the QAOA (what we call "depth 2" in Problem 1) achieves a 0.6924-factor approximation for the NP-complete MAX CUT problem. Unfortunately, worst-case analysis of higher levels has in general proven difficult, but Bravyi, Kliesch, Koenig and Tang [Bra+20] have shown an interesting negative result — QAOA to any *constant* level/depth cannot outperform the classical Goemans-Williams algorithm for MAX CUT [GW95]. Thus, superconstant depth is *necessary* if QAOA is to have a hope of outperforming the best classical algorithms for MAX CUT. In terms of complexity theoretic hardness, Farhi and Harrow [FH16] showed that even level-1 QAOA's output distribution cannot be efficiently simulated by a classical computer.

Most relevant to this paper, however, is the work of Bittel and Kliesch [BK21], which roughly shows that finding the optimal set of rotation angles (the θ_j in Problem 1 and Problem 3) is NP-hard. Let us clearly state how the present work differs from [BK21]:

1. [BK21] fixes both the depth of the VQA and the precise sequence of Hamiltonians H_i to be applied as part of the input. It then asks: What is the complexity of computing the optimal rotation angles θ_i so as to minimize overlap with a given observable?

   In contrast, our aim here is to study the complexity of optimizing the *depth* itself. Thus, Problem 1 does not fix the depth m, nor the order/multiplicity of application of any of the Hamiltonian terms.

2. [BK21] shows that optimizing the rotation angles in QAOA is NP-hard, *even if* one is allowed to work in time polynomial in the *dimension* of the system. (Formally, this is obtained by reducing a MAX CUT instance of encoding size N to QAOA acting on $\log(N)$ qubits.)

[7] A k-local n-qubit Hamiltonian H is a quantum analogue of a MAX-k-SAT instance, and can be written $H = \sum_i H_i$, with each "quantum clause" H_i acting non-trivially on some subset of k qubits. Strictly speaking, each H_i is tensored with the identity matrix on $n - k$ qubits to ensure all operators in the sum have the correct dimension.

In contrast, we work in the standard setting of allowing only poly-time computations in the number of qubits, n, not the dimension. In return, we obtain stronger hardness results, both in that NP $\subseteq$ QCMA (and thus QCMA-hardness is a stronger statement than NP-hardness[8]), and in that we show hardness of approximation up to any multiplicative factor $N^{1-\epsilon}$.

2. Practical/numerical studies. For clarity, numerical studies are not directly related to our work. However, due to the intense practical interest in VQA for the NISQ era, for completeness we next survey some of the difficulties encountered when optimizing VQAs on the numerical side. For this, note that VQAs are typically used to solve problems which can be phrased as energy optimization problems (such as NP-complete problems like MAX CUT [FGG14]).

In this direction, two crucial problems can arise in the classical optimization part of the standard VQA setup: (i) barren plateaus [McC+18], which lead to vanishing gradients, and (ii) local minima [BK21], many of which can be highly non-optimal. Such unwanted local minima are also called *traps*. In order to counterbalance these challenges, heuristic optimization strategies have led to promising results in relevant cases but with not too many qubits. Initialization-dependent barren plateaus [McC+18] can be avoided by tailored initialization [Zho+20], and there are indications that barren plateaus are a less significant challenge than traps [AK22]. In general, the optimization can be improved using natural gradients [WGK20], multitask learning type approach [ZY20], optimization based on trigonometric model functions [KB22], neural network-based optimization methods [Riv+21], brick-layer structures of generic unitaries [SVC22], and operator pool-based methods [Gri+19; BK22]. ADAPT-VQEs [Gri+19] iteratively grow the VQA's parametrized quantum circuit (PQC) by adding operators from a pool that have led to the largest derivative in the previous step. This strategy allows one to avoid barren plateaus and even "burrow" out of some traps [Gri+22]. CoVar [BK22] is based on similar ideas complemented with estimating several properties of the variational state in parallel using classical shadows [HKP20]. The optimization strategies are of a heuristic nature, and analytic results are scarce. Finally, it has been numerically observed [TLM20; Wie+20] and analytically shown [Lar+21] that VQA-type ansätze become almost free from traps when the ansatz is overparameterized. Our work implies that these practical approaches cannot work for all instances and, therefore, provides a justification to resort to such heuristics.

1.3 Techniques

We focus on techniques for showing QCMA-hardness of approximation, as containment in QCMA is straightforward[9] for both MIN-VQA and MIN-QAOA.

To begin, recall that in a QCMA proof system (Definition 1), given a YES input, there exists a poly-length *classical* proof y causing a quantum poly-size circuit V to accept, and for a NO input, all poly-length proofs y cause V to reject. Our goal is to embed such proof systems into instances of Problem 1 and Problem 3, while maintaining a large promise gap ratio m'/m. To do so, we face three main challenges: (1) Where will hardness of approximation come from? Typically, one requires a PCP theorem [AS98; Aro+98] for such results, which remains a notorious open question for both QCMA and QMA[10] [AAV13]. (2) Problem 1 places no restrictions on which Hamiltonians are applied, in which order, and with which rotation angles.

[8] Note that for $\log(N)$-size instances of QAOA as in [BK21], one cannot hope for more than NP-hardness, since both Hamiltonians H_b and H_c have polynomial dimension, and thus can be classically simulated efficiently. Thus, such instances are verifiable in NP.

[9] The prover sends angles θ_j, and the verifier simulates each $e^{i\theta_j H_j}$ via known Hamiltonian simulation algorithms [LC17].

[10] Quantum Merlin-Arthur (QMA) is QCMA but with a quantum proof.

How can one enforce computational structure given such flexibility? In addition, MIN-QAOA presents a third challenge: (3) How to overcome the previous two challenges when we are only permitted two Hamiltonians, H_b and H_c, the latter of which must also act as the observable?

To address the first challenge, we appeal to the hardness of approximation work of Umans [Uma99]. The latter showed how to use a graph-theoretical construct, known as a *disperser*, to obtain strong hardness of approximation results for Σ_2^p (the second level of the Polynomial-Time Hierarchy). Hiding at the end of that paper is Theorem 9, which showed that the techniques therein also apply to yield hardness of approximation within factor $N^{1/5-\epsilon}$ for a rather artificial NP-complete problem. Gharibian and Kempe [GK12] then showed that [Uma99] can be extended to obtain hardness of approximation results for a quantum analogue of Σ_2^p, and also obtained QCMA-hardness of approximation within $N^{1-\epsilon}$ for an even more artificial problem, Quantum Monotone Minimum Satisfying Assignment (QMSA, Problem 2). Roughly, QMSA asks — given a quantum circuit V accepting a monotone set (Definition 2) of strings, what is the smallest Hamming weight string accepted by V? Here, our approach will be to construct many-one reductions from QMSA to MIN-VQA and MIN-QAOA, where we remark that maintaining the $N^{1-\epsilon}$ hardness ratio (i.e. making the reduction approximation-ratio-preserving) will require special attention.

1. The reduction for MIN-VQA. To reduce a given QMSA circuit $V = V_L \cdots V_1$ to a VQA instance $(\{H_i\}, M, m, m')$, we utilize a "hybrid Cook-Levin + Kitaev" circuit-to-Hamiltonian construction, coupled with a *pair* of clocks (whereas Kitaev [KSV02] requires only one clock). Here, a *non-hybrid* (i.e. standard) circuit-to-Hamiltonian construction is a quantum analogue of the Cook-Levin theorem, i.e. a map from quantum circuits V to local Hamiltonians H_V, so that there exists a proof $|\psi\rangle$ accepted by V if and only if H_V has a low-energy[11] "history state", $|\psi_{\text{hist}}\rangle$. A history state, in turn, is a quantum analogue of a Cook-Levin tableau, except that each time step of the computation is encoded in superposition via a clock construction of Feynman [Fey86]. In contrast, our construction is "hybrid" in that it uses a clock register like Kitaev, but does not produce a history state in superposition over all time steps, like Cook-Levin. A bit more formally, the Hamiltonians $\{H_i\}$ of our VQA instance act on four registers, $ABCD$, denoting proof (A), workspace (B), clock 1 (C), and clock 2 (D). To an honest prover, these Hamiltonians $\{H_i\}$ may be viewed as being partitioned into two sets: Hamiltonians for "setting proof bits", denoted P, and Hamiltonians for simulating gates from V, denoted Q. An example of a Hamiltonian in P is

$$P_j \coloneqq X_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}} \tag{3}$$

which says: If clock 1 (register C) is at time j *and* clock 2 (register D) is at time $|D|$ (more on clock 2 shortly), then flip the jth qubit of register A via a Pauli X gate. An example of a Hamiltonian in Q is

$$Q_j \coloneqq (V_j)_{AB} \otimes |01\rangle\langle 10|_{C_{|A|+j},C_{|A|+j+1}} + (V_j^\dagger)_{AB} \otimes |10\rangle\langle 01|_{C_{|A|+j},C_{|A|+j+1}}, \tag{4}$$

which allows the prover to apply gate V_j of V to registers AB, while updating clock 1 from time $|A|+j$ to $|A|+j+1$. In this first (insufficient) attempt at a reduction, the honest prover for MIN-VQA acts as follows: First, apply a subset of the P Hamiltonians to prepare the desired input y to the QMSA verifier V in register A, and then evolve Hamiltonians Q_1 through Q_L to simulate gates V_1 through V_L on registers A and B. The observable M is then defined to

[11] By "energy" of a state $|\psi\rangle$ against Hamiltonian H, one means the expectation $\langle\psi|H|\psi\rangle$, whose minimum possible value is precisely $\lambda_{\min}(H)$, i.e. the smallest eigenvalue of H.

measure the designated output qubit of B in the standard basis, conditioned on C being at time T.

The crux of this (honest prover) setup is that if we start with a YES (respectively, NO) instance of QMSA, then the Hamming weight of the optimal y is at most g (respectively, at least g'), for $g'/g \geq N_{\mathsf{QMSA}}^{1-\epsilon}$ and N_{QMSA} the encoding size of the QMSA instance. This, in turn, means that the VQA prover applies at most g Hamiltonians from P (YES case), or at least g' Hamiltonians from P (NO case). The problem is that the prover must *also* apply Hamiltonians Q_1 through Q_L in order to simulate the verifier, V, and so we have hardness ratio $m'/m = (g'+L)/(g+L) \to 1$ if $L \in \omega(g)$, as opposed to $N^{1-\epsilon}$!

To overcome this, we make flipping each bit of P "more costly" by utilizing a *2D clock setup*. This, in turn, will ensure the hardness ratio $(g'+L)/(g+L)$ becomes (roughly)

$$\frac{g'\,|D| + L}{g\,|D| + L} \approx \frac{g'}{g} \text{ for } |D| \in \omega(L), \tag{5}$$

as desired. Specifically, to flip bit A_j for any j, we force the prover to first sequentially increment the *second* clock, D, from 1 to $|D|$. By Equation (3), P_j can now flip the value of A_j — but it cannot increment time in C (i.e. we remain in time step j on clock 1). This next forces the prover to decrement D from $|D|$ back to 1, at which point a separate Hamiltonian (not displayed here) can increment clock C from j to $j+1$. The entire process then repeats itself to flip bit A_{j+1}. What is crucial for our desired approximation ratio is that we only have a single copy of register D, i.e. we re-use it to flip each bit A_j, thus effectively making CD act as a 2D clock. This ensures the added overhead to the encoding size of the VQA instance scales as $|D|$, not $|A|\,|D|$, which is what one would obtain if CD encoded a 1D clock (i.e. if each A_j had a *separate* copy of D).

Finally, to show soundness against provers deviating from the honest strategy above, we first establish that any sequence of evolutions from $\{H_i\}$ keeps us in a desired logical computation space, i.e. the span of vectors of form

$$S \coloneqq \left\{ V_{s-|A|} \cdots V_1 |y\rangle_A |0\cdots 0\rangle_B |\widetilde{s}\rangle_C |\widetilde{t}\rangle_D \;\middle|\; y \in \{0,1\}^{|A|}, s \in \{1,\ldots,|C|\}, t \in \{1,\ldots,|D|\} \right\}, \tag{6}$$

for $|y\rangle_A$ the "proof string" prepared via P-gates and $\widetilde{s}$ and $\widetilde{t}$ the unary representations of time steps s and t in clocks 1 and 2, respectively. We then show that applying too few Hamiltonian evolutions from $\{H_i\}$ results in a state with either no support on large Hamming weight strings y (meaning the verifier V must reject in the NO case), or no support on states with a fully executed verification circuit $V = V_L \cdots V_1$ (in which case we design V to reject).

2. The reduction for MIN-QAOA. At a high level, our goal is to mimic the reduction to MIN-VQA above. However, the fact that we have only two Hamiltonians at our disposal, H_b (driving Hamiltonian) and H_c (cost Hamiltonian), and no separate observable M, complicates matters. Very roughly, our aim is to *alternate* even and odd steps of the honest prover's actions from MIN-VQA, so that H_b simulates the even steps, and H_c the odd ones. To achieve this requires several steps:

1. First, we modify the MIN-VQA setup so that all the odd (respectively, even) local terms H_i pairwise commute. This ensures that the actions of $\exp(i\theta H_b)$ and $\exp(i\theta H_c)$ can be analyzed, since H_b and H_c will consist of sums of (now commuting) H_i terms.

2. In MIN-VQA, all Hamiltonians satisfied $H_i^2 = I$, which intuitively means an honest prover could use H_i to either act trivially ($\theta_i = 0$) or perform some desired action ($\theta_i = \pi$). For MIN-QAOA, we instead require a trick inspired by [BK21] — we introduce certain local

terms G_j (Equation (61)) with 3-cyclic behavior. In words, the honest prover can induce *three* logical actions from such G_j, obtained via angles $\theta_j \in \{0, \pi/3, 2\pi/3\}$, respectively.

3. We next add additional constraints to H_b to ensure its unique ground state encodes the correct start state (see Equation (57) of Problem 3). This is in contrast to MIN-VQA, where the initial state $|0\cdots0\rangle$ is fixed and independent of the H_i.

4. Finally, the observable M is added as a local term to H_c, but scaled larger than all other terms in H_c. This ensures that for any state $|\psi\rangle$, $|\langle\psi|H_c - M|\psi\rangle|$ is "small", so that measuring cost Hamiltonian H_c once the QAOA circuit finishes executing is "close" to measuring M.

As for soundness, the high-level approach is similar to MIN-VQA, in that we analyze a logical space of computation steps, akin to Equation (6), and track Hamming weights of prepared proofs in this space. The analysis, however, is more involved, as the construction itself is more intricate than for MIN-VQA. For example, a new challenge for our MIN-QAOA construction is that evolving by a Hamiltonian (specifically, H_c) does *not* necessarily preserve the logical computation space. We thus need to prove that we may "round" each intermediate state in the analysis back to the logical computation space, in which we can then track the Hamming weight of the proof y (Lemma 5).

1.4 Open questions

We have shown that the optimal depth of a VQA or QAOA ansatz is hard to approximate, even up to large multiplicative factors. A natural question is whether similar NP-*hardness* of approximation results for depth can be shown when (e.g.) the cost Hamiltonian in QAOA is classical, such as in [FGG14]? Since we aimed here to capture the strongest possible hardness result, i.e. for QCMA, our Hamiltonians were necessarily not classical/diagonal. Second, although our results are theoretical worst-case results, VQAs are of immense practical interest in the NISQ community. Can one design good heuristics for optimal depth approximation which often work well in practice? Third, can one approximate the optimal depth for QAOA on *random* instances of a computational problem? Here, for example, recent progress has been made by Basso, Gamarnik, Mei and Zhou [Bas+22], Boulebnane and Montanaro [BM22], and Anshu and Metger [AM22], which give analytical bounds on the success probability of QAOA at various levels and on random instances of various constraint satisfaction problems, for instance size n going to infinity. The bounds of [AM22], for example, show that even *superconstant* depth (i.e. scaling as $o(\log\log n)$) is insufficient for QAOA to succeed with non-negligible probability for a random spin model. On a positive note, we remark that [BM22] give numerical evidence (based on their underlying analytical bounds) that at around level 14, QAOA begins to surpass existing classical SAT solvers for the case of random 8-SAT. Fourth, we have given the first natural QCMA-hard to approximate problems. What other QCMA-complete problems can be shown hard to approximate? A natural candidate here is the *Ground State Connectivity* problem [GS15; GMV17; WBG20], whose hardness of approximation we leave as an open question. Finally, along these lines, can a PCP theorem for QCMA be shown as a first stepping stone towards a PCP theorem for QMA?

1.5 Organization

This paper is organized as follows. We begin with basic definitions and notation in Section 2. In Section 3, we show Theorem 1. Section 4 shows Theorem 4.

2 Basic definitions and notation

We begin with notation, and subsequently define QCMA.

2.1 Notation

Throughout, the relation $\coloneqq$ denotes a definition, and $[n] \coloneqq \{1, 2, \ldots, n\}$. We use $|x|$ to specify the length of a vector or string or the cardinality of set x. The term I_A denotes the identity operator/matrix on qubits with indices in register A. By $\|H\|_\infty$ we denote the spectral norm of an operator H acting on $\mathbb{C}^d$, i.e. $\max_{|\psi\rangle \in \mathbb{C}^d} \frac{\|H|\psi\rangle\|_2}{\||\psi\rangle\|_2}$, for $\|\cdot\|_2$ the standard Euclidean norm. The trace norm of an operator is denoted by $\|\cdot\|_{\mathrm{tr}}$. e_i refers to a computational basis state.

2.2 Complexity classes

Definition 1 (Quantum-classical Merlin-Arthur (QCMA)). *Let $\Pi = (\Pi_{\mathrm{yes}}, \Pi_{\mathrm{no}})$ be a promise problem. Then $\Pi \in$ QCMA if and only if there is a polynomial p such that for any $x \in \Pi$ there exists a quantum circuit V_x of size $p(|x|)$ with one designated output qubit satisfying:*

(i) If $x \in \Pi_{\mathrm{yes}}$ there exists a string $y \in \{0,1\}^{p(|x|)}$ such that $\Pr[V_x$ accepts $y] \geq 2/3$ and

(ii) if $x \in \Pi_{\mathrm{no}}$ and all strings $y \in \{0,1\}^{p(|x|)}$ it holds that $\Pr[V_x$ accepts $y] \leq 1/3$.

Often, it is helpful to separate the qubits into an a *proof register* A, which contains the classical proof $|y\rangle$, and an *ancilla/work register* B, which is initialized in the $|0\rangle$ state. Then the acceptance probability can be expressed as

$$\Pr[V_x \text{ accepts } (x,y)] = \left\langle y;0 \middle| V_x^{(n)\dagger} M^{(B_1)} V_x^{(n)} \middle| y;0 \right\rangle, \tag{7}$$

where the measurement is given by an operator $M^{(B_1)}$ acting on the first qubit of the work register B.

QCMA was first defined in [AN02], and satisfies NP $\subseteq$ QCMA $\subseteq$ QMA. QCMA-complete problems include Identity Check on Basis States (i.e. "does a quantum circuit act almost as the identity on all computational basis states?") [WJB03] and Ground State Connectivity (GSCON) (i.e. is the ground space of a local Hamiltonian "connected"?) [GS15]. The latter remains hard (specifically, $\mathsf{QCMA}_{\mathrm{EXP}}$-hard) in the 1D translation-invariant setting [WBG20].

3 QCMA-hardness of approximation for VQAs

In this section, we show Theorem 1. We begin in Section 3.1 with relevant definitions and lemmas. Section 3.2 proves Theorem 1.

3.1 Definitions and required facts

For convenience, we first restate Problem 1.

Problem 1 (VQA minimization (MIN-VQA(k,l))). *For an n-qubit system:*

- *Input:*
  1. *Set $H = \{H_i\}$ of Hamiltonians[12], where H_i acts non-trivially only on a subset[13] $S_i \subseteq [n]$ of size $|S_i| = k$.*

[12] An n-qubit Hamiltonian H is a $2^n \times 2^n$ Hermitian matrix. Any unitary operation U on a quantum computer can be generated via an appropriate choice of Hamiltonian H and evolution time $t \geq 0$, i.e. $U = e^{iHt}$.

[13] For Theorem 1, it will suffice to take $k \in O(1)$. In principle, however, containment in QCMA holds for any $k \leq n$, so long as the H_i are sparse in the standard Hamiltonian simulation sense [AT03]. By sparse, one means that each row r of H_i contains at most r non-zero entries, which can be computed in poly-time given r.

2. *An l-local observable M acting on a subset of l qubits.*
3. *Integers $0 \leq m \leq m'$ representing circuit depth thresholds.*

- *Output:*

1. *YES if there exists a list of at most m angles*[14] *$(\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$ and a list $(G_1, \ldots, G_m)$ of Hamiltonians from H (repetitions permitted) such that*

$$|\psi\rangle := e^{i\theta_m G_m} \cdots e^{i\theta_1 G_1} |0 \cdots 0\rangle \tag{1}$$

satisfies $\langle\psi|M|\psi\rangle \leq 1/3$.

2. *NO if for all lists of at most m' angles $(\theta_1, \ldots, \theta_{m'}) \in \mathbb{R}^{m'}$ and all lists $(G_1, \ldots, G_{m'})$ of Hamiltonians from H (repetitions permitted),*

$$|\psi\rangle := e^{i\theta_{m'} G_{m'}} \cdots e^{i\theta_1 G_1} |0 \cdots 0\rangle \tag{2}$$

satisfies $\langle\psi|M|\psi\rangle \geq 2/3$.

We next require definitions and a theorem from [GK12].

Definition 2 (Monotone set). *A set $S \subseteq \{0,1\}^n$ is called* monotone *if for any $x \in S$, any string obtained from x by flipping one or more zeroes in x to one is also in S.*

Definition 3 (Quantum circuit accepting monotone set). *Let V be a quantum circuit consisting of 1- and 2-qubit gates, which takes in an n-bit classical input register, m-qubit ancilla register initialized to all zeroes, and outputs a single qubit, q. For any input $x \in \{0,1\}^n$, we say V* accepts *(respectively,* rejects*) x if measuring q in the standard basis yields 1 (respectively, 0) with probability at least $1 - \epsilon_Q$ (If not specified, $\epsilon_Q = 1/3$). We say V accepts a* monotone set *if the set $S \subseteq \{0,1\}^n$ of all strings accepted by V is monotone (Definition 2).*

Problem 2 (QUANTUM MONOTONE MINIMUM SATISFYING ASSIGNMENT (QMSA)). *Given a quantum circuit V accepting a non-empty monotone set $S \subseteq \{0,1\}^n$, and integer thresholds $0 \leq g \leq g' \leq n$, output:*

- *YES if there exists an $x \in \{0,1\}^n$ of Hamming weight at most g accepted by V.*
- *NO if all $x \in \{0,1\}^n$ of Hamming weight at most g' are rejected by V.*

Theorem 5 (Gharibian and Kempe [GK12]). *QMSA is QCMA-complete, and moreover it is QCMA-hard to decide whether, given an instance of QMSA, the minimum Hamming weight string accepted by V is at most g or at least g' for $g'/g \in O(N^{1-\epsilon})$ (where $g' \geq g$).*

In words, QMSA is QCMA-hard to approximate within $N^{1-\epsilon}$ for any constant $\epsilon > 0$, where N is the encoding size of the QMSA instance.

3.2 QCMA-completeness

Theorem 1. MIN-VQA(k, l) *is* QCMA*-complete for $k \geq 4$, $l = 2$, and $m \leq \mathsf{poly}(n)$. Moreover, for any $\epsilon > 0$, it is* QCMA*-hard to distinguish between the YES and NO cases of* MIN-VQA *even if $m'/m \geq N^{1-\epsilon}$, where N is the encoding size of the instance.*

In words, it is QCMA-hard to decide whether, given an instance of MIN-VQA, the variational circuit can prepare a "good" ansatz state with at most m evolutions, or if all sequences of m' evolutions fail to prepare a "good" ansatz state, for $m'/m \in O(N^{1-\epsilon})$ (where $m' \geq m$).

[14] Throughout Problem 1, for clarity we assume all angles are specified to $\mathsf{poly}(n)$ bits.

Proof. Containment in QCMA is straightforward; the prover sends the angles θ_i and indices of Hamiltonians H_i to evolve, which the verifier then completes using standard Hamiltonian simulation techniques [Llo96; LC17]. We now show QCMA-hardness of approximation. Let $\Pi' = (V', g, g')$ be an instance of QMSA, for $V' = V'_{L'} \cdots V'_1$ a sequence of L' 2-qubit gates taking in n'_V input bits and m'_V ancilla qubits.

Preprocessing V'. To ease our soundness analysis, we would like to make two assumptions about V' without loss of generality; these can be simply ensured as follows. Suppose V' takes in n'_V input qubits in register A' and m'_V ancilla qubits in register B'. Apply each of the following modifications in the order listed.

Assumption 6. *V' only* reads *register A', but does not write to it. To achieve this, add n'_V ancilla qubits (initialized to $|0\rangle$) to B', and prepend V' with n'_V CNOT gates applied transversally to copy input x from A' to the added ancilla qubits in B'. Update any subsequent gate which acts on the original input x to instead act on its copied version in B'.*

Assumption 7. *The output qubit of V' is set to $|0\rangle$ until V'_L is applied. To achieve this, add a single ancilla qubit to B' initialized to $|0\rangle$, and treat this as the new designated output qubit. Append to the end of V' a CNOT gate from its original output wire to the new output wire.*

Call the new circuit with all modifications V. V acts on $n_V \coloneqq n'_V$ input qubits, $m_V \coloneqq m'_V + n'_V + 1$ ancilla qubits, and consists of $L \coloneqq L' + n'_V + 1$ gates.

Proof organization. The remainder of the proof is organized as follows. Section 3.2.1 constructs the MIN-VQA instance. Section 3.2.2 proves observations and lemmas required for the completeness and soundness analyses. Sections 3.2.3 and 3.2.4 show completeness and soundness, respectively. Finally, Section 3.2.5 analyzes the hardness ratio achieved by the reduction.

3.2.1 The MIN-VQA instance

We now construct our instance Π of MIN-VQA as follows. Π acts on a total of n qubits, which we partition into 4 registers: A (proof), B (workspace), C (clock 1), and D (clock 2). Register A consists of n_V qubits, B of m_V qubits, C of $L + n_V + 1$ qubits, and D of $\lceil L^{1+\delta} \rceil$ qubits for some fixed $0 < \delta < 1$ to be chosen later. Throughout, we use shorthand (e.g.) $|A|$ for the number of qubits in a register A.

Our construction will ensure C (respectively, D) always remains in the span of logical time steps, $\mathcal{T}_C \coloneqq \{|\widetilde{s}\rangle\}_{s=1}^{|C|}$ (respectively, $\mathcal{T}_D \coloneqq \{|\widetilde{t}\rangle\}_{t=1}^{|D|}$), defined as:

$$|\widetilde{s}\rangle \coloneqq |0\rangle^{\otimes s-1}|1\rangle|0\rangle^{\otimes |C|-s} \quad \text{for } 1 \le s \le |C| \tag{8}$$

$$|\widetilde{t}\rangle = |0\rangle^{\otimes t-1}|1\rangle|0\rangle^{\otimes |D|-t} \quad \text{for } 1 \le t \le |D|\,. \tag{9}$$

For example for C, $|\widetilde{1}\rangle = |1\rangle|0\rangle^{\otimes |C|-1}$, $|\widetilde{2}\rangle = |0\rangle|1\rangle|0\rangle^{\otimes |C|-2}$, $|\widetilde{3}\rangle = |00\rangle|1\rangle|0\rangle^{\otimes |C|-3}$, and so forth. Note this differs from the usual Kitaev unary clock construction, which encodes time t via $|1\rangle^{\otimes t}|0\rangle^{\otimes N-t}$ [KSV02]. This allows us to reduce the locality of our Hamiltonian.

Throughout, we use (e.g.) C_j to refer to qubit j and $C_{i,j}$ and qubits i and j of register C. All qubits not explicitly mentioned are assumed to be acted on by the identity. Define four families of Hamiltonians as follows:

- (F) For propagation of the second clock, D, define 2-local Hamiltonians as

$$F_j \coloneqq |01\rangle\langle 10|_{D_{j,j+1}} + |10\rangle\langle 01|_{D_{j,j+1}} \text{ for all } j \in \{1, \ldots, |D| - 1\}. \tag{10}$$

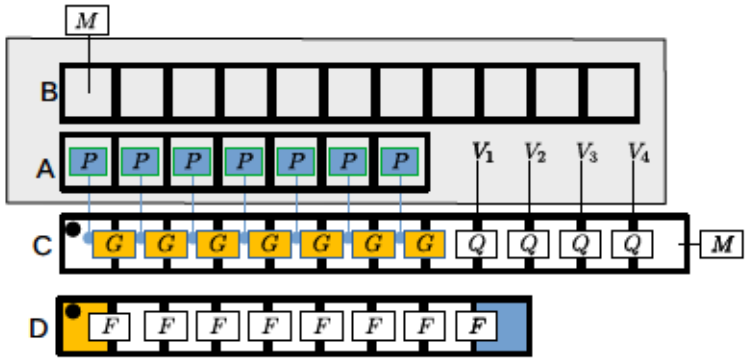

Figure 1: Sketch describing the VQA instance. A colored square (say, blue) at index j of a register means that register's jth qubit must be in $|1\rangle$ for any blue gates to act non-trivially. So, for example, the G gates increment the first clock register C, but only if the D register is in the state $|1\rangle_{D_1}$. For the initial state, C_1 and D_1 are in the $|1\rangle$ state, marked by a black dot. The gates F increment the second clock register D. The P gates are controlled operations on the C register, which perform X operations on the A register, but only if D is in the state $|1\rangle_{D_{|D|}}$. The Q gates increment the clock register C, while also applying the circuit $V_1, \ldots, V_L$ on the AB registers. The measurement operator M acts on the B_1 and $C_{|C|}$ qubit.

- (G) For propagation of the first clock, C, define 3-local Hamiltonians as

$$G_j := \left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1} \text{ for all } j \in \{1, \ldots, |A|\}. \tag{11}$$

- (P) For each qubit $j \in \{1, \ldots, |A|\}$ of A, define 3-local Hamiltonian as

$$P_j := X_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}}. \tag{12}$$

- (Q) For each gate V_k for $k \in \{1, \ldots, L\}$, let R_k denote the two qubits of AB which V_k acts on. Define 4-local Hamiltonians as

$$Q_k := (V_k)_{R_k} \otimes |01\rangle\langle 10|_{C_{|A|+k,|A|+k+1}} + (V_k^\dagger)_{R_k} \otimes |10\rangle\langle 01|_{C_{|A|+k,|A|+k+1}}. \tag{13}$$

Denote the union of these four sets of Hamiltonians as $S_{FGPQ} := F \cup G \cup P \cup Q$. Set a 2-local observable

$$M := I - |1\rangle\langle 1|_{B_1} \otimes |1\rangle\langle 1|_{C_{|C|}} \tag{14}$$

where we assume without loss of generality that V outputs its answer on qubit B_1. Set $m = g \cdot (2|D| - 1) + |A| + L$, $m' = g' \cdot (2|D| - 1) + |A| + L$. To aid the reader in the remainder of the proof, all definitions above are summarized in Figure 3.2.1.

It remains to choose our initial state. Strictly speaking, Problem 1 mandates initial state $|0 \cdots 0\rangle_{ABCD}$. However, to keep notation simple, it will be convenient to instead choose

$$|\phi\rangle := |0 \cdots 0\rangle_{AB} |10^{|C|-1}\rangle_C |10^{|D|-1}\rangle_D = |0 \cdots 0\rangle_{AB} |\widetilde{1}\rangle_C |\widetilde{1}\rangle_D, \tag{15}$$

i.e. with the two clock registers C and D initialized to their starting clock state, $|\widetilde{1}\rangle$. This is without loss of generality — we may, in fact, start with *any* standard basis state as our initial state without requiring major structural changes to our construction, as the following observation states.

Term	Description	Properties
V'	Input QMSA instance's verification circuit	$V' = V'_{L'} \cdots V'_1$
L'	Number of 1- and 2-qubit gates in V'	
n'_V	Number of proof qubits taken in by V'	
m'_V	Number of ancilla qubits taken in by V'	
g,g'	YES/NO thresholds for QMSA instance, resp.	
V	QMSA verifier obtained from V' via Assump. 6 and 7	$V = V_L \cdots V_1$
L	Number of 1- and 2-qubit gates in V	$L = L' + n'_V + 1$
n_V	Number of proof qubits taken in by V	$n_V = n'_V$
m_V	Number of ancilla qubits taken in by V	$m_V = m'_V + n'_V + 1$
A	Proof register	$\lvert A\rvert = n_V$
B	Workspace register	$\lvert B\rvert = m_V$
C	Clock 1 register	$\lvert C\rvert = L + n_V + 1$
D	Clock 2 register	$\lvert D\rvert = \lceil L^{1+\delta} \rceil$ for δ chosen in (56) to satisfy (54)
F	Propagation terms for clock 2	Act on register D, $\lvert F\rvert = \lvert D\rvert - 1$
G	Propagation terms for clock 1	Act on registers C, D, $\lvert g\rvert = \lvert A\rvert$
P	Hamiltonian terms for setting proof bits	Act on registers A, C, D, $\lvert P\rvert = \lvert A\rvert$
Q	Hamiltonian terms for simulating verifier gates, V_k	Act on registers A, B, C, $\lvert Q\rvert = L$
M	Observable for MIN-VQA instance	$M \coloneqq I - \lvert 1\rangle\langle 1\rvert_{B_1} \otimes \lvert 1\rangle\langle 1\rvert_{C_{\lvert C\rvert}}$
m,m'	YES/NO thresholds for MIN-VQA instance, resp.	$m = g \cdot (2\lvert D\rvert - 1) + \lvert A\rvert + L$, $m' = g' \cdot (2\lvert D\rvert - 1) + \lvert A\rvert + L$.

Figure 2: Terms used in the proof of Theorem 1.

Observation 8. *Fix any standard basis state* $|x\rangle_{ABCD} = \overline{X}|0\cdots 0\rangle_{ABCD}$, *for*

$$\overline{X} \coloneqq X_1^{x_1} \otimes \cdots \otimes X_N^{x_N} \tag{16}$$

with $N \coloneqq |A| + |B| + |C| + |D|$. *Consider the updated set* $S'_{FGPQ} \coloneqq \{\overline{X}H\overline{X} \mid H \in S_{FGPQ}\}$, *where for simplicity we match* $H \in S_{FGPQ}$ *with* $H' \coloneqq \overline{X}H\overline{X} \in S'_{FGPQ}$. *Then, for any* $m \in \mathbb{N}$, *and any sequence* $(H_t)_{t=1}^m$ *of Hamiltonians drawn from* S_{FGPQ},

$$e^{i\theta_m H_m} \cdots e^{i\theta_2 H_2} e^{i\theta_1 H_1} |x\rangle_{ABCD} = e^{i\theta_m H'_m} \cdots e^{i\theta_2 H'_2} e^{i\theta_1 H'_1} |0\cdots 0\rangle_{ABCD}. \tag{17}$$

Moreover, each H *and* H' *are the same locality.*

Proof. The first claim follows since $X^2 = I$, by Observation 9[15], and since $|x\rangle_{ABCD} = \overline{X}|0\cdots 0\rangle_{ABCD}$. The second claim also follows from $X^2 = I$ and the fact that $\overline{X}$ is a tensor product of operators from set $\{I, X\}$. □

This concludes the construction.

[15]Note that conjugation by $\overline{X}$ will alter the clock encoding in Equations (18)-(21), but this alternation is logically irrelevant since it is equivalent to a local change of basis applied simultaneously to all $H \in S_{FGPQ}$ and to $|0\cdots 0\rangle_{ABCD}$.

3.2.2 Helpful observations and lemmas

We next state and prove all observations and technical lemmas for the later correctness analysis of our construction.

Observation 9. *For all $\theta \in \mathbb{R}$, and all $F_j \in F$, $G_j \in G$, $P_j \in P$ and $Q_k \in Q$,*

$$e^{i\theta F_j} = \cos(\theta)(|01\rangle\langle 01| + |10\rangle\langle 10|)_{D_{j,j+1}} + i\sin(\theta)F_j + (I - |01\rangle\langle 01| - |10\rangle\langle 10|)_{D_{j,j+1}} \tag{18}$$

$$e^{i\theta G_j} = \cos(\theta)\left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1} + i\sin(\theta)G_j + \left(I - \left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1}\right) \tag{19}$$

$$e^{i\theta P_j} = (\cos(\theta)I + i\sin(\theta)X)_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}} + (I - |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}}) \tag{20}$$

$$e^{i\theta Q_k} = \cos(\theta)I_{AB} \otimes (|01\rangle\langle 01| + |10\rangle\langle 10|)_{C_{|A|+k,|A|+k+1}} + i\sin(\theta)Q_k + I_{AB} \otimes (I - |01\rangle\langle 01| - |10\rangle\langle 10|)_{C_{|A|+k,|A|+k+1}}. \tag{21}$$

For clarity, any register not explicitly listed in equations above is assumed to be acted on by identity.

Proof. Follows straightforwardly via Taylor series expansion since

$$F_j^2 = (|01\rangle\langle 01| + |10\rangle\langle 10|)_{D_{j,j+1}} \quad \text{for all } j \in \{1, \ldots, |D| - 1\}, \tag{22}$$

$$G_j^2 = (|01\rangle\langle 01| + |10\rangle\langle 10|)_{C_{j,j+1}} \otimes |1\rangle\langle 1|_{D_1} \quad \text{for all } j \in \{1, \ldots, |A|\}, \tag{23}$$

$$P_j^2 = I_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}} \quad \text{for all } j \in \{1, \ldots, |A|\}, \tag{24}$$

$$Q_k^2 = I_{AB} \otimes (|01\rangle\langle 01| + |10\rangle\langle 10|)_{C_{|A|+k,|A|+k+1}} \quad \text{for all } k \in \{1, \ldots, L\}. \tag{25}$$

□

Definition 4 (Support only on logical time steps). *We say state $|\psi\rangle_{ABCD}$ is* supported only on logical time steps *if it can be written*

$$|\psi\rangle_{ABCD} = \sum_{s=1}^{|C|} \sum_{t=1}^{|D|} \alpha_{st} |\eta_{st}\rangle_{AB} |\widetilde{s}\rangle_C |\widetilde{t}\rangle_D \tag{26}$$

for unit vectors $|\eta_{st}\rangle$ and $\sum_{st} |\alpha_{st}|^2 = 1$, and $|\widetilde{s}\rangle \in \mathcal{T}_C$ and $|\widetilde{t}\rangle \in \mathcal{T}_D$ defined as in Equation (8) and Equation (9), respectively.

Observation 10. *Recall that the initial state $|\phi\rangle = |0 \cdots 0\rangle_{AB} |\widetilde{1}\rangle_C |\widetilde{1}\rangle_D$ is supported only on logical time steps. Then, for any $m \in \mathbb{N}$ and sequence of evolutions $\exp(i\theta_j H_j)$ for $\theta_j \in \mathbb{R}$ and $H_j \in S_{FGPQ}$,*

$$e^{i\theta_m H_m} \cdots e^{i\theta_2 H_2} e^{i\theta_1 H_1} |\phi\rangle \tag{27}$$

is supported only on logical time steps.

Proof. Consider any logical time step $|\widetilde{s}\rangle_C |\widetilde{t}\rangle_D$. By Equations (18)-(21), the set of possible evolutions act as follows[16]:

- $e^{i\theta F_j}$: can map from $|\widetilde{j}\rangle_D$ to $|\widetilde{j+1}\rangle_D$ or vice versa for $j \in \{1, \ldots, |D| - 1\}$.
- $e^{i\theta G_j}$: if $|\widetilde{t}\rangle_D = |\widetilde{1}\rangle_D$, can map from $|\widetilde{j}\rangle_C$ to $|\widetilde{j+1}\rangle_C$ or vice versa for $j \in \{1, \ldots, |A|\}$.

[16] Without loss of generality, we focus on the *non-trivial* (i.e. non-identity) action of each evolution, as any trivial action immediately preserves logical time steps.

- $e^{i\theta P_j}$: acts invariantly (i.e. as identity) on C and D for all $j \in \{1, \ldots, |A|\}$.
- $e^{i\theta Q_j}$: can map from $|\widetilde{|A| + j}\rangle_C$ to $|\widetilde{|A| + j + 1}\rangle_C$ (while applying V_j to AB) or vice versa (for $V_j^\dagger$) for $j \in \{1, \ldots, L\}$. □

The following lemma tells us that any sequence of Hamiltonian evolutions $\exp(i\theta_u H_u)$ on initial state $|\phi\rangle$ remains in a certain logical computation space.

Lemma 1. *Define*

$$S \coloneqq \left\{ V_{s-|A|} \cdots V_1 |y\rangle_A |0 \cdots 0\rangle_B |\widetilde{s}\rangle_C |\widetilde{t}\rangle_D \;\middle|\; y \in \{0,1\}^{|A|}, s \in \{1, \ldots, |C|\}, t \in \{1, \ldots, |D|\} \right\}, \quad (28)$$

where we adopt the convention that the V gates are present only when $s > |A|$. Then, for any $m \in \mathbb{N}$,

$$\Pi_{u=1}^m e^{i\theta_u H_u} |\phi\rangle \in \mathrm{Span}(S) \quad (29)$$

for any angles $\theta_u \in \mathbb{R}$ and sequence of Hamiltonians $H_u \in S_{FGPQ}$.

Proof. For convenience, define

$$|\eta_{y,s,t}\rangle \coloneqq V_{s-|A|} \cdots V_1 |y\rangle_A |0\rangle_B |\widetilde{s}\rangle_C |\widetilde{t}\rangle_D \quad (30)$$

for $s \in \{1, \ldots, |C|\}$, $t \in \{1, \ldots, |D|\}$, and $y \in \{0,1\}^{|A|}$. Observe first that $|\phi\rangle = |0 \cdots 0\rangle_{AB} |\widetilde{1}\rangle_C |\widetilde{1}\rangle_D = |\eta_{0,1,1}\rangle \in S$. Thus, it suffices to prove $\mathrm{Span}(S)$ is closed under application of $e^{i\theta H}$ for any $\theta \in \mathbb{R}$ and $H \in S_{FGPQ}$.

Case 1: $H = H_j \in F$ for $j \in \{1, \ldots, |D| - 1\}$. Equations (10) and (18) immediately yield $e^{i\theta H} |\eta_{y,s,t}\rangle = |\eta_{y,s,t}\rangle$ for $t \notin \{j, j+1\}$. Consider thus $t \in \{j, j+1\}$. Restricted to this space, $e^{i\theta H}$ acts logically as

$$e^{i\theta H} = \cos(\theta) I_{ABCD} + i \sin(\theta) I_{ABC} \otimes \left(|\widetilde{j+1}\rangle\langle\widetilde{j}| + |\widetilde{j}\rangle\langle\widetilde{j+1}| \right)_D. \quad (31)$$

Thus, $e^{i\theta H}$ maps

$$|\eta_{y,s,j}\rangle \mapsto \cos(\theta) |\eta_{y,s,j}\rangle + i \sin(\theta) |\eta_{y,s,j+1}\rangle, \quad (32)$$
$$|\eta_{y,s,j+1}\rangle \mapsto i \sin(\theta) |\eta_{y,s,j}\rangle + \cos(\theta) |\eta_{y,s,j+1}\rangle. \quad (33)$$

Case 2: $H = H_j \in G$ for $j \in \{1, \ldots, |A|\}$. Equations (11) and (19) immediately yield $e^{i\theta H} |\eta_{y,s,t}\rangle = |\eta_{y,s,t}\rangle$ unless $s \in \{j, j+1\}$ and $t = 1$. Consider thus $s \in \{j, j+1\}$ and $t = 1$. Restricted to this space, $e^{i\theta H}$ acts logically as

$$e^{i\theta H} = \cos(\theta) I_{ABCD} + i \sin(\theta) I_{AB} \otimes \left(|\widetilde{j+1}\rangle\langle\widetilde{j}| + |\widetilde{j}\rangle\langle\widetilde{j+1}| \right)_C \otimes I_D. \quad (34)$$

Thus, $e^{i\theta H}$ maps

$$|\eta_{y,j,1}\rangle \mapsto \cos(\theta) |\eta_{y,j,1}\rangle + i \sin(\theta) |\eta_{y,j+1,1}\rangle, \quad (35)$$
$$|\eta_{y,j+1,1}\rangle \mapsto i \sin(\theta) |\eta_{y,j,1}\rangle + \cos(\theta) |\eta_{y,j+1,1}\rangle. \quad (36)$$

Case 3: $H = H_j \in P$ **for** $j \in \{1, \ldots, |A|\}$. By Equation (20), $e^{i\theta H}|\eta_{y,s,t}\rangle = |\eta_{y,s,t}\rangle$ unless $s = j$ and $t = |D|$. Consider thus $s = j$ and $t = |D|$. Restricted to this space, $e^{i\theta H}$ maps

$$|\eta_{y,j,|D|}\rangle \mapsto \cos(\theta)|\eta_{y,j,|D|}\rangle + i\sin(\theta)|\eta_{y',j,|D|}\rangle \tag{37}$$

for y' defined as y with its jth bit flipped. Since y in Equation (28) is not fixed, we conclude $e^{i\theta H}|\eta_{y,s,t}\rangle \in \mathrm{Span}(S)$, as claimed.

Case 4: $H = H_j \in Q$. Equations (13) and (21) immediately yield $e^{i\theta H}|\eta_{y,s,t}\rangle = |\eta_{y,s,t}\rangle$ for $s \notin \{|A| + j, |A| + j + 1\}$. Consider thus $s \in \{|A| + j, |A| + j + 1\}$. Restricted to this space, $\exp(i\theta H)$ acts logically as

$$e^{i\theta H} = \cos(\theta) I_{ABCD} + i\sin(\theta) Q_j. \tag{38}$$

Thus, $e^{i\theta H}$ maps

$$\begin{aligned} |\eta_{y,|A|+j,t}\rangle &\mapsto \cos(\theta)|\eta_{y,|A|+j,t}\rangle + i\sin(\theta)V_j|\eta_{y,|A|+j,t}\rangle \\ &= \cos(\theta)|\eta_{y,|A|+j,t}\rangle + i\sin(\theta)|\eta_{y,|A|+j+1,t}\rangle, \end{aligned} \tag{39}$$

$$\begin{aligned} |\eta_{y,|A|+j+1,t}\rangle &\mapsto i\sin(\theta)V_j^\dagger|\eta_{y,|A|+j+1,t}\rangle + \cos(\theta)|\eta_{y,|A|+j+1,t}\rangle \\ &= i\sin(\theta)|\eta_{y,|A|+j,t}\rangle + \cos(\theta)|\eta_{y,|A|+j+1,t}\rangle, \end{aligned} \tag{40}$$

where we have used Assumption 6 that gates V_j never write to the proof register A (and thus y remains unchanged under application of V_k). This yields the claim. □

Next, we relate the circuit depth of a state generated by our VQA to the Hamming weight of the proof string y.

Lemma 2. *Let $(H_u)_{u=1}^m$ be a sequence of Hamiltonians drawn from S_{FGPQ} which maps the initial state (15) to*

$$|\phi_m\rangle \coloneqq \Pi_{u=1}^m e^{i\theta_u H_u}|\phi\rangle. \tag{41}$$

Suppose $|\phi_m\rangle$ has non-zero overlap with some $|\eta_{y,s,t}\rangle$ with y of Hamming weight at least w and $s = |A| + 1$. Then, $m \geq w(2|D| - 1) + |A|$ with at least $w(2|D| - 1) + |A|$ of the H_u drawn from $F \cup G \cup P$.

Proof. By Lemma 1, $|\phi_m\rangle \in S$. Repeating the following argument for any bit of y set to 1 will yield a lower bound on the number of gates of $2w|D|$, which is almost what we want.

Consider any index $j \in \{1, \ldots, |A|\}$ such that $y_j = 1$ (equivalently, in state $|\eta_{y,s,t}\rangle$ the qubit A_j is set to 1). Since the qubit A_j of the initial state $|\phi\rangle$ is set to $|0\rangle$, there must be an evolution step $u \in \{1, \ldots, m\}$ at which A_j is mapped from[17] $|0\rangle$ to $|1\rangle$. By Observation 9, only the Hamiltonian $P_j \in P$ can induce this mapping, and P_j requires C and D to be set to $|\widetilde{j}\rangle_C|\widetilde{|D|}\rangle_D$ in order to act non-trivially. Let us analyze each of these two requirements in order.

First, to obtain $|\widetilde{j}\rangle$ in C, there are only two possibilities:

- (Case a) We are in the initial state $|\phi\rangle$ with no Hamiltonian evolutions applied yet and $j = 1$ (recall $|\phi\rangle$ has C and D set to $|\widetilde{1}\rangle_C|\widetilde{1}\rangle_D$ by definition), or
- (Case b) we are at a later evolution step at which C was updated to $|\widetilde{j}\rangle$ from either $j - 1$ or $j + 1$. Since $1 \leq j \leq |A|$, by Observation 9, only operators G_{j-1} (for $j > 1$) and G_j can effect this map. Both of these operators require D set to $|\widetilde{1}\rangle$.

[17] For clarity, throughout this proof, by "mapped from $|k\rangle$ to $|k'\rangle$", we mean $|k\rangle$ is mapped to a state with non-zero overlap with $|k'\rangle$. This suffices for Lemma 2, since it only cares about non-zero overlap with some $|\eta_{y,s,t}\rangle$.

Thus, in both cases, D is also set to $|\widetilde{1}\rangle$. So, assume one of the two cases has just occurred to update C to $|\widetilde{j}\rangle$.

The second requirement for P_j to act non-trivially was that D be set to $||\widetilde{D}|\rangle$. But since D is currently set to $|\widetilde{1}\rangle$ in the initial state, and since only operators in F can change clock D (by precisely one time step per operator), we must apply at least $|D|-1$ operators in F to obtain a state with $|\widetilde{j}\rangle_C||\widetilde{D}|\rangle_D$. (To see that C must still be set to $|\widetilde{j}\rangle_C$ at this point, observe that all operators in $S_{FGPQ} \setminus F$ act invariantly unless D equals $|\widetilde{1}\rangle$ or C is at least $||\widetilde{|A|+1}\rangle$.) Applying P_j is now necessary to flip A_j from $|0\rangle$ to $|1\rangle$. We have thus reached an intermediate state at which A_j is $|1\rangle$ and C and D are $|\widetilde{j}\rangle_C||\widetilde{D}|\rangle_D$.

Finally, either all bits of y are now set correctly and C must be updated to $||\widetilde{|A|+1}\rangle$ (due to the condition $s = |A|+1$), or we wish to repeat the argument above for the next index $j' \neq j$ for which we wish to map $A_{j'}$ from $|0\rangle$ to $|1\rangle$. In both cases, D must first be reset back to $|\widetilde{1}\rangle$ (otherwise operators in G act invariantly, and these are precisely the operators which can update C to either j' or to $|A|+1$ as needed). Running the argument above regarding F in reverse, we obtain that at least a number $|D|-1$ of F-gates are needed to return D back to $|\widetilde{1}\rangle$, and at least one G-gate is needed to update C from j to j' or to $|A|+1$.

Summing all gate costs together, for each A_j to be flipped from $|0\rangle$ to $|1\rangle$ and for C to be updated to the next value of j', we require at least $2|D|$ gates. Thus, if $|\eta_{y,s,t}\rangle$ has y with Hamming weight at least w, at least $2w|D|$ gates from $F \cup G \cup P$ are required. This is almost what we want.

The final $|A|-w$ gates required for the claim arise because one requires at least $|A|$ G-gates to map C from its initial value of 1 to $|A|+1$, and above we have only accounted for w of these G-gates (i.e. corresponding to all j with $y_j = 1$). □

Finally, the next lemma ensures that any prover applying fewer than L Hamiltonians from Q cannot satisfy the YES case's requirements for MIN-VQA.

Lemma 3. *For any $m \in \mathbb{N}$, let $(H_u)_{u=1}^m$ be any sequence of Hamiltonians drawn from S_{FGPQ} and containing strictly fewer than L Hamiltonians from Q. Then, for observable $M = I - |1\rangle\langle 1|_{B_1} \otimes |1\rangle\langle 1|_{C_{|C|}}$, the state $|\phi_m\rangle := \Pi_{u=1}^m e^{i\theta_u H_u}|0\cdots 0\rangle_{ABC}$ satisfies*

$$\langle \phi_m | M | \phi_m \rangle = 1. \tag{42}$$

Proof. By Lemma 1, $|\phi_m\rangle \in S$ for S from Equation (28). Next, by Observation 9, Hamiltonians from $F \cup P$ act invariantly on clock C, and Hamiltonians from G can only increment C from 1 (i.e. its initial value in $|\phi\rangle$) to $|A|+1$. The observable M, however, acts non-trivially only when C is set to $|C| = |A|+L+1$. The only Hamiltonians which can increment C from $|A|+1$ to $|A|+L+1$ are those from Q. Each such $H_s \in Q$ can map C from time $|A|+s$ to $|A|+s+1$ or vice versa, for $s \in \{1,\ldots,L\}$. Thus, since strictly fewer than L of the H_u chosen are from Q, it follows that $|\phi_m\rangle$ has no support on time step $|C| = |A|+L+1$, i.e. $(I_{AB} \otimes |1\rangle\langle 1|_{C_{|C|}})|\phi_m\rangle = 0$. The claim now follows since we Assumption 7 says verifier $V = V_L \cdots V_1$ has its output qubit, denoted B_1, set to $|0\rangle$ until its final gate V_L is applied. □

3.2.3 Completeness

With all observations and lemmas of Section 3.2.2 in hand, we are ready to prove completeness of the construction. Specifically, in the YES case, there exists an input $y \in \{0,1\}^{|A|}$ of Hamming weight at most g accepted with probability at least $2/3$ by V. The honest prover proceeds as follows.

- (Prepare classical proof) Prepare state (up to global phase) $|\psi_0\rangle \coloneqq |y\rangle_A|0\rangle_B|\widetilde{|A|+1}\rangle_C|\widetilde{1}\rangle_D$ as follows. Starting with $|\phi\rangle = |0\cdots 0\rangle_{AB}|\widetilde{1}\rangle_C|\widetilde{1}\rangle_D$:

  1. Set $j = 1$.
  2. If $y_j = 1$ then
     - Apply, in order, unitaries $\exp(i(\pi/2)F_1)$, $\exp(i(\pi/2)F_2)$,..., $\exp(i(\pi/2)F_{|D|-1})$. This maps registers C and D to 1 and $|D|$, respectively.
     - Apply $\exp(i(\pi/2)P_j)$, which maps A_j from 0 to 1.
     - Apply, in order, unitaries $\exp(i(\pi/2)F_{|D|})$, $\exp(i(\pi/2)F_{|D|-1})$,..., $\exp(i(\pi/2)F_1)$. This maps registers C and D back to 1 and 1, respectively.
  3. Apply unitary $\exp(i(\pi/2)G_j)$, which maps C from j to $j+1$.
  4. Set $j = j+1$.
  5. If $j < |A|+1$, return to line 2 above.

  This process applies $g(2|D|-1)+|A|$ gates.

- (Simulate verifier) Prepare the sequence of states $|\psi_j\rangle = e^{i\frac{\pi}{2}Q_j}\cdots e^{i\frac{\pi}{2}Q_1}|\psi_0\rangle$ by applying, in order, unitaries $\exp(i(\pi/2)Q_1)$, $\exp(i(\pi/2)Q_2)$,..., $\exp(i(\pi/2)Q_L)$. Since the jth step of this process applies $\exp(i(\pi/2)Q_j)$, and since the state $|\psi_0\rangle$ has clock C set to $|A|+1$, Observation 9 and Equation (13) imply that

$$e^{i\frac{\pi}{2}Q_j}|\psi_{j-1}\rangle = \left((V_j)_{R_j} \otimes |\widetilde{|A|+j+1}\rangle\langle\widetilde{|A|+j}|_C\right)|\psi_{j-1}\rangle, \tag{43}$$

  i.e. we increment the clock from $|A|+j$ to $|A|+j+1$ and apply the jth gate V_j. The final state obtained is thus $|\psi_L\rangle = (V_L\cdots V_1|y\rangle_A|0\rangle_B)\otimes|\widetilde{|A|+L+1}\rangle_C|\widetilde{1}\rangle_D$. This process applies L gates.

Since V accepts y with probability at least $2/3$, we conclude $\langle\psi_L|M|\psi_L\rangle \leq 1/3$, as desired. The number of Hamiltonians from S_{FGPQ} we needed to simulate in this case is $m = g(2|D|-1)+|A|+L$, as desired.

3.2.4 Soundness

We next show soundness. Specifically, in the NO case, for all inputs $y \in \{0,1\}^{|A|}$ of Hamming weight at most g', V accepts with probability at most $1/3$. So, consider any sequence of $m' = g'(2|D|-1)+|A|+L$ Hamiltonian evolutions producing state $|\phi_{m'}\rangle \coloneqq \Pi_{t=1}^{m'} e^{i\theta_u H_u}|0\cdots 0\rangle_{AB}|\widetilde{1}\rangle_C|\widetilde{1}\rangle_D$ for arbitrary $\theta_u \in \mathbb{R}$ and Hamiltonians $H_u \in S_{FGPQ}$. Lemma 1 says we may write

$$|\phi_{m'}\rangle = \sum_{y\in\{0,1\}^{|A|}}\sum_{s=1}^{|C|}\sum_{t=1}^{|D|}\alpha_{y,s,t}|\eta_{y,s,t}\rangle \in \mathrm{Span}(S) \tag{44}$$

with $\sum_{y,s,t}|\alpha_{y,s,t}|^2 = 1$. Now, for the observable (14) follows that

$$\langle\phi_{m'}|M|\phi_{m'}\rangle = 1 - \langle\phi_{m'}|\left(|1\rangle\langle 1|_{B_1}\otimes|1\rangle\langle 1|_{C_{|C|}}\right)|\phi_{m'}\rangle = 1 - \langle\eta||1\rangle\langle 1|_{B_1}|\eta\rangle \text{ for} \tag{45}$$

$$|\eta\rangle \coloneqq \sum_{y\in\{0,1\}^{|A|}}\sum_{t=1}^{|D|}\alpha_{y,|A|+L+1,t}V_L\cdots V_1|y\rangle_A|0\rangle_B|\widetilde{|A|+L+1}\rangle_C|\widetilde{t}\rangle_D, \tag{46}$$

where we have used Equation (44) and the fact that M projects onto time step $|C|$ in register C. Now, if we applied strictly less than L evolutions from Q, Lemma 3 says we have no weight on time step $|C|$, so that $\langle\phi_{m'}|M|\phi_{m'}\rangle = 1 \geq 2/3$, as required in the NO case. If, on the other hand, we applied at least L evolutions from Q, then we must have applied at most $g'(2|D|-1)+|A|$ evolutions from $F \cup G \cup P$ (otherwise, we have a contradiction since $m' = g'(2|D|-1)+|A|+L$). Lemma 2 hence implies the right hand side of Equation (45) equals $1 - \langle\eta_{g'}||1\rangle\langle 1|_{B_1}|\eta_{g'}\rangle$ for[18]

$$|\eta_{g'}\rangle \coloneqq \sum_{y \text{ s.t. } \mathrm{HW}(y)\leq g'} \sum_{t=1}^{|D|} \alpha_{y,|A|+L+1,t} V_L \cdots V_1 |y\rangle_A |0\rangle_B ||A|\widetilde{+L}+1\rangle_C |\widetilde{t}\rangle_D, \tag{47}$$

where $\mathrm{HW}(y)$ denotes the Hamming weight of the bitstring y. But since any input y of Hamming weight at most g' is accepted with probability at most $1/3$, we conclude $\langle\phi_{m'}|M|\phi_{m'}\rangle \geq 2/3$, as claimed.

3.2.5 Hardness ratio

Finally, we show our reduction has the desired approximation ratio. Observe

$$\frac{m'}{m} = \frac{g'(2|D|-1)+|A|+L}{g(2|D|-1)+|A|+L} = \frac{g'(2\lceil L^{1+\delta}\rceil - 1)+|A|+L}{g(2\lceil L^{1+\delta}\rceil - 1)+|A|+L}. \tag{48}$$

Since $|A| \leq L$ by definition, and since we will choose $\delta > 0$ as a small constant, this ratio scales asymptotically as g'/g. Recall now that Theorem 5 says that for any constant $\epsilon' > 0$, the QMSA instance $\Pi' = (V', g, g')$ we are reducing from is QCMA-hard to approximate within $g'/g \in O((N')^{1-\epsilon'})$, for N' the encoding size of Π'. Observe that

$$N' \geq 2L' \log(n'_V), \tag{49}$$

as L' is the number of gates comprising V', and each gate V'_i takes $O(1)$ bits to specify (assuming a constant-size universal gate set) and $2\log n'_V$ bits to indicate which pair of qubits V'_i acts on. So it remains to compare N' with the encoding size N of our MIN-VQA instance Π. For this, observe that each Hamiltonian in S_{FGPQ} may be specified using $O(\log(|A|+|B|+|C|+|D|))$ bits, since each H_i requires $O(1)$ bits to specify the 4-local matrix itself, and $4\log(|A|+|B|+|C|+|D|)$ bits to specify the (at most) 4-tuple of qubits on which H_i acts. Similarly, M is specifiable using $O(\log(|A|+|B|+|C|+|D|))$ bits. Thus, $N \in O(|S_{FGPQ}|\log(|A|+|B|+|C|+|D|))$, where we may bound

$$\begin{aligned}
|S_{FGPQ}| &= 2|A| + L + |D| - 1 && (50)\\
&= 2n'_V + (L' + n'_V + 1) + (L' + n'_V + 1)^{1+\delta} - 1 && (51)\\
&\leq 2n'_V + (L' + n'_V + 1) + (2L' + 1)^{1+\delta} - 1 && (52)\\
&\in O(L'^{1+\delta}), && (53)
\end{aligned}$$

where we have used that $n'_V \in O(L')$ (otherwise, V' does not have enough time to read all n'_V of its input bits). Recall now that to obtain the hardness ratio of our claim, we must show the following: For any desired constant $\epsilon > 0$, there exist $\epsilon' > 0$ and $\delta' > 0$ such that

$$\frac{g'}{g} \geq (N')^{1-\epsilon'} \geq N^{1-\epsilon}. \tag{54}$$

[18] Below, $\mathrm{HW}(y)$ denotes the Hamming weight of string y.

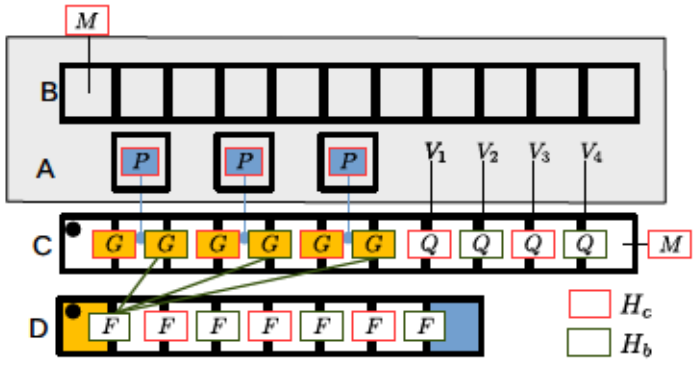

Figure 3: Figure describing the QAOA instance (see Figure 1 for further details). The border color of each gate indicates if the generator belongs to H_b or H_c. Compared to the previous VQA instance, the P now only act at even time steps in C and the even-indexed G_j and the F_1 generator are combined into one generator, denoted by the red and dark green edges.

We know that for some $c \in \Theta(1)$,

$$(N')^{1-\epsilon'} \geq (2L' \log(n_V'))^{1-\epsilon'} \text{ and } N \leq c(L')^{1+\delta} \log(|A| + |B| + |C| + |D|). \tag{55}$$

Equation (54) thus holds if

$$\frac{1-\epsilon'}{1+\delta} \geq \frac{(1-\epsilon)(\log c + \log\log(|A| + |B| + |C| + |D|)) - (1-\epsilon')\log\log(n_V^2)}{(1+\delta)\log(L')} + (1-\epsilon). \tag{56}$$

We conclude that for large enough L', for any desired $\epsilon > 0$, one can choose $\epsilon' > 0$ and $\delta > 0$ so that Equation (54) holds, as desired.

□

4 Extension of the hardness results to QAOAs

In this section, we prove Theorem 4, which is restated for convenience shortly. First, we define the optimization problem MIN-QAOA covered by the theorem.

A k-local Hamiltonian is a sum of strictly k-local terms, i.e. Hermitian operators each of which acts non-trivially on at most k qubtis. As mentioned previously, our definition of MIN-QAOA is more general than that of [FGG14], and closer to that of [Had+19].

Problem 3 (QAOA minimization (MIN-QAOA(k))). *For an n-qubit system:*

- *Input:*
  1. *A set $H = \{H_b, H_c\}$ of k-local Hamiltonians.*
  2. *A $\mathrm{poly}(n)$-size quantum circuit U_b preparing the ground state of H_b, denoted $|\mathrm{gs}_b\rangle$.*
  3. *Integers $0 \leq m \leq m'$ representing thresholds for depth.*
- *Output:*
  1. *YES if there exists a sequence of angles[19] $(\theta_i)_{i=1}^m \in \mathbb{R}^m$, such that*

$$|\psi\rangle := e^{i\theta_m H_b} e^{i\theta_{m-1} H_c} \cdots e^{i\theta_2 H_b} e^{i\theta_1 H_c} |\mathrm{gs}_b\rangle \tag{57}$$

  satisfies $\langle\psi|H_c|\psi\rangle \leq \frac{1}{3}$.

[19] Throughout Problem 3, for clarity we assume all angles are specified to $\mathrm{poly}(n)$ bits.

2. *NO if for all sequences of angles* $(\theta_i)_{i=1}^{m'} \in \mathbb{R}^{m'}$

$$|\psi\rangle := e^{i\theta_{m'}H_b}e^{i\theta_{m'-1}H_c}\cdots e^{i\theta_2 H_b}e^{i\theta_1 H_c}|\mathrm{gs}_b\rangle, \tag{58}$$

satisfies $\langle\psi|H_c|\psi\rangle \geq \frac{2}{3}$.

Just as for MIN-VQA, by "optimal depth" of a QAOA, we mean the minimum number of Hamiltonian evolutions m required above. The expectation value thresholds $\frac{1}{3}$ and $\frac{2}{3}$ are arbitrary and can be changed by rescaling and shifting H_c.

Theorem 4. MIN-QAOA(k) *is* QCMA*-complete for* $k \geq 4$ *and* $m \leq \mathsf{poly}(n)$. *Moreover, for any* $\epsilon > 0$, *it is* QCMA*-hard to distinguish between the YES and NO cases of* MIN-QAOA *even if* $m'/m \geq N^{1-\epsilon}$, *where* N *is the number of strictly* k*-local terms comprising* H_b *and* H_c.

Proof. Containment in QCMA is again straightforward and thus omitted. For QCMA-hardness of approximation, we again use a reduction from an instance $\Pi = (V, g, g')$ of QMSA, for $V = V_L \cdots V_1$ a sequence of L 2-qubit gates taking in n_V input bits and m_V ancilla qubits. All those terms are defined as in the proof of Theorem 1.

Proof organization. The proof is organized as follows. In Section 4.1 we explain the modifications of the VAQ instances to obtain the QAOA instances of our construction. Section 4.2 provides notation preliminary technical results needed for the QCMA-completeness proof. Then, completeness is shown in Section 4.3 and soundness in Section 4.4. Finally, in Section 4.5, we analyze the hardness ratio achieved by the reduction.

4.1 Modifications of the VQA instances to obtain the QAOA instances

To specify our QAOA instance, we modify the set S_{FGPQ} from the proof of Theorem 1 to suit our reduction here as follows. The structural changes are illustrated in Figure 3. Briefly recapping the proof techniques outline in Section1.3, we: (1) implement the reduction with only two generators by alternating even and odd steps of the honest prover's actions, so that H_b simulates the even steps, and H_c the odd ones, (2) introduce terms G_j (Equation (61)) with 3-cyclic behavior, i.e. allowing three logical actions, (3) add new constraints to H_b to ensure its unique ground state encodes the correct start state (see Equation (57) of Problem 3), and (4) add the observable O to H_c (scaled larger than other terms in H_c) to obtain the correct cost function. An undesired side effect of this is that evolution by H_c allows one to leave the desired logical computation space, S. We show via Lemma 5 that the states obtained are still close to the set, which suffices for our soundess analysis.

To begin, we use registers composed of $|A| = n_V$, $|B| = m_V$, $|C| = L + 2n_V + 1$, and $|D| = \lceil L^{1+\delta} \rceil$ qubits, respectively, where $0 < \delta < 1$ is fixed by specified later. Without loss of generality, we assume $|D|$ and L to be even. Additionally to the changes we outline, we also add diagonal terms additional diagonal terms. This will be relevant for defining the initial state later on.

- (F) We remove F_1,

$$F_j := |01\rangle\langle 10|_{D_{j,j+1}} + |10\rangle\langle 01|_{D_{j,j+1}} - 2|00\rangle\langle 00|_{D_{j,j+1}} \text{ for all } j \in \{2, \ldots, |D| - 1\}. \tag{59}$$

- (G) We double the number of qubits G acts on,

$$G_j \coloneqq \Big(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\Big) \otimes |1\rangle\langle 1|_{D_1} - 2|001\rangle\langle 001|_{C_{j,j+1},D_1} \quad \text{for all } j \in \{1,3,\ldots,2\,|A|-1\}, \tag{60}$$

$$G_j \coloneqq \frac{i}{\sqrt{3}}\big(|0110\rangle\langle 1010| + |1001\rangle\langle 0110| + |1010\rangle\langle 1001| - |1010\rangle\langle 0110| - |0110\rangle\langle 1001| - |1001\rangle\langle 1010|\big)_{C_{j,j+1},D_{1,2}} - 2|0010\rangle\langle 0010|_{C_{j,j+1},D_{1,2}} \quad \text{for all } j \in \{2,4,\ldots,2\,|A|\}. \tag{61}$$

  While odd numbered gates can only change the clock, even numbered ones can still increment C, but also have the option of moving $|\widetilde{1}\rangle_D \to |\widetilde{2}\rangle_D$, which is the operation performed by $F_1^{(o)}$ in the proof of Theorem 1 on MIN-VQA. The superscript (o) refers to the gates of the previous VQA proof. The following relations hold:

$$e^{i\frac{\pi}{3}G_i}|\widetilde{i},\widetilde{1}\rangle_{C,D} = e^{i\frac{\pi}{2}G_i^{(o)}}|\widetilde{i},\widetilde{1}\rangle_{C,D} \propto |\widetilde{i+1},\widetilde{1}\rangle_{C,D}, \tag{62}$$

$$e^{i\frac{2\pi}{3}G_i}|\widetilde{i},\widetilde{1}\rangle_{C,D} = e^{i\frac{\pi}{2}F_1^{(o)}}|\widetilde{i},\widetilde{1}\rangle_{C,D} \propto |\widetilde{i},\widetilde{2}\rangle_{C,D}, \tag{63}$$

  where the last step means equality up to a phase.

- (P) For each qubit $j \in \{1,\ldots,|A|\}$ of A, we define the flip operator, but now it only acts on even values in the clock register.

$$P_j \coloneqq X_{A_j} \otimes |1\rangle\langle 1|_{C_{2j}} \otimes |1\rangle\langle 1|_{D_{|D|}} - 2|00\rangle\langle 00|_{C_{2j},D_{|D|}} \text{ for all } j \in \{1,\ldots,|A|\}. \tag{64}$$

- (Q) We shift the C-indices of the Q-gates because reading in the proof takes longer time now,

$$Q_k \coloneqq (V_k)_{R_k} \otimes |01\rangle\langle 10|_{C_{2|A|+k,2|A|+k+1}} + (V_k^\dagger)_{R_k} \otimes |10\rangle\langle 01|_{C_{2|A|+k,2|A|+k+1}} \tag{65}$$

$$- 2|00\rangle\langle 00|_{C_{2|A|+k,2|A|+k+1}}. \tag{66}$$

- (M), (H_0) We add these two operators

$$H_0 = -\left(\sum_{i\in[|A|]} |0\rangle\langle 0|_{A_i} + \sum_{i\in[|B|]} |0\rangle\langle 0|_{B_i}\right) \otimes |1\rangle\langle 1|_{C_1} \tag{67}$$

$$M = I - |1\rangle\langle 1|_{B_1} \otimes |1\rangle\langle 1|_{C_{|C|}} \tag{68}$$

  to the set of generators.

To construct our desired QAOA instance, we define a partition of all gates into two groups:

$$\mathcal{G}_1 = \{G_i\}_{i\in\{2,4,\ldots,2|A|\}} \cup \{F_i\}_{i\in\{3,5,\ldots,|D|-1\}} \cup \{Q_i\}_{i\in\{2,4,\ldots,L\}}, \tag{69}$$

$$\mathcal{G}_2 = \{G_i\}_{i\in\{1,3,\ldots,2|A|-1\}} \cup \{F_i\}_{i\in\{2,4,\ldots,|D|-2\}} \cup \{Q_i\}_{i\in\{1,3,\ldots,L-1\}} \cup \{P_i\}_{i\in[|A|]}. \tag{70}$$

Intuitively, $\mathcal{G}_1$ (respectively, $\mathcal{G}_2$) will be part of our Hamiltonian H_b (respectively, H_c). Note also that all operators in $\mathcal{G}_1$ (respectively, $\mathcal{G}_2$) pairwise commute, a fact we will use in our analysis. Finally, in addition to Assumption 6 and Assumption 7 from VQA, we shall use the following.

Assumption 11. *The acceptance probability of V in the YES (respectively, NO) case is at least $1-\epsilon_Q$ (respectively, at most ϵ_Q), where $\epsilon_Q = O(N^{-1})$. This is achieved via standard parallel repetition of V k times, followed by a majority vote. This will increase the encoding size of V — for k repetitions, the new gate sequence length scales with $L' = k(L+O(1))$, and yields $\epsilon'_Q = \epsilon_Q^{O(k)}$. For the precision we require, it suffices to set $k = O(\log(N))$. Thus, our encoding size increases by a multiplicative log factor, which does not affect our final approximation ratio calculation.*

4.1.1 The QAOA instance

The QAOA instance we use to prove Theorem 4 takes the generators

$$H_b = \sum_{\Gamma\in\mathcal{G}_1} \Gamma + H_0\,, \tag{71}$$

$$H_c = \kappa \sum_{\Gamma\in\mathcal{G}_2} \Gamma + M \tag{72}$$

with $m = g(2\,|D|-2)+|C|-1$ and $m' = g'(2\,|D|-4)+|C|-1$. Crucially, the generators/operators comprising H_b (respectively, H_c) pairwise commute. The Q gates are taken from a QMSA circuit where using Assumption 11, we set the acceptance threshold of the circuit to $\sqrt{\epsilon_Q} = \frac{1}{48m'}$. Also, we set $\kappa = \frac{1}{24|\mathcal{G}|}$.

We proceed by first characterizing the initial state and cost function as defined in Problem 3.

4.1.2 Preliminaries

Initial state Recall in Problem 3, the initial state should be a ground state of H_b (given as input via a preparation circuit U_b). We want this initial state to be

$$|\mathrm{gs}_b\rangle = |0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD}, \tag{73}$$

which can trivially be prepared by a constant-sized circuit U_b. To see that we indeed obtain this $|\mathrm{gs}_b\rangle$, note that for all generators except G_1, M, which are not included in H_b, $|\mathrm{gs}_b\rangle$ is a ground state of the generator. Moreover, the groundstate is unique because for each qubit, the state is uniquely determined by one of the generators, which implies that the entire state is unique.

$$\begin{aligned}
F_i|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -2|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} \quad \forall i\in\{3,5,\ldots,|D|-1\}, && \|F_i\|_\infty = 2,\\
G_i|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -2|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} \quad \forall i\in\{2,4\ldots,2\,|A|\}, && \|G_i\|_\infty = 2,\\
Q_i|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -2|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} \quad \forall i\in\{2,4\ldots,L\}, && \|Q_i\|_\infty = 2,\\
H_0|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -(|A|+|B|)|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD}, && \|H\|_\infty = |A|+|B|\,.
\end{aligned}$$

Since the state (73) is a ground state of all the generators of H_b and the terms of H_b mutually commute, it is also a ground state of H_b. Moreover, since every qubit is non-trivially supported by at least one generator of H_b, it is also the unique ground state for the entire Hilbert space, i.e., $|\mathrm{gs}_b\rangle$ represents the unique one-dimensional subspaces where each gate acts non-trivially.

Cost function In the QAOA setup, the measured observable is H_c. For our construction we want to use the observable M. We can find an upper bound for the difference of these operators. Namely, $\forall|\Psi\rangle\in\mathcal{H}$

$$|\langle\Psi|(H_c-M)|\Psi\rangle| = \kappa|\langle\Psi|\textstyle\sum_{\Gamma\in\mathcal{G}_2}\Gamma|\Psi\rangle| \le 2\kappa|\mathcal{G}_2| \le \frac{1}{12} \tag{74}$$

where we used (1) that $\|g\|_\infty \le 2$ for all $\Gamma\in\mathcal{G}_2$, and (2) the definition of κ.

4.2 Preliminaries for the completeness proof

We first define the set of states comprising our logical computation space,

$$S \coloneqq \{V_{t-2|A|-1}\cdots V_1|y\rangle_A|0\cdots 0\rangle_B|\widetilde{t}\rangle_C|\widetilde{s}\rangle_D \,|\, \forall (y,t,s)\in I_S\} \tag{75}$$

with

$$I_S = \left\{(y,t,s)\middle| y\in\{0,1\}^{|A|}, t\in\{1,\ldots,|C|\}, s\in\begin{cases}\{1,\ldots,|D|\} & \text{if } t\in\{2,4,\cdots 2|A|\}\\ \{1\} & \text{otherwise}\end{cases}\right\} \tag{76}$$

is the allowed index set. Here, the notation means that V_1 is only applied if $t > 2|A|+1$. Below, will often write a state $|\Psi_S\rangle \in \mathrm{span}(S)$ as

$$|\Psi_S\rangle = \sum_{(y,t,s)\in I_S} a_{y,t,s} V_{t-2|A|-1}\cdots V_1|y\rangle_A|0\cdots 0\rangle_B|\widetilde{t}\rangle_C|\widetilde{s}\rangle_D \tag{77}$$

$$=: \sum_{(y,t,s)\in I_S} a_{y,t,s}|\Psi_{y,t,s}\rangle \tag{78}$$

We also define the function W, or observable, which is roughly intended to capture a lower bound on the number of Hamiltonian evolutions required to prepare a given logical state $|\Psi_{y,t,s}\rangle$:

$$W(y,t,s) \coloneqq (2|D|-4)\mathrm{HW}(y) + t + (-1)^{\delta_{y_{\lceil t/2\rceil},1}}(s+\delta_{s,1}-2), \tag{79}$$

for $\mathrm{HW}(y)$ the Hamming weight of y.

We next show a helpful lemma regarding the action of each Hamiltonian on our logical computation space, S.

Lemma 4. *The following two statements hold:*

- *For every $|\Psi_{y,t,s}\rangle \in S$ and $H_i \in \{H_b, H_c\}$*

$$e^{iH_i\theta}|\Psi_{y,t,s}\rangle = e^{i\alpha^{(i)}_{y,t,s}\theta}e^{i\Gamma^{(i)}_{y,t,s}\theta}|\Psi_{y,t,s}\rangle \tag{80}$$

  for some phase $\alpha\in\mathbb{R}$. In words, applying H_i simulates application of a single gate $\Gamma^{(b)}_{y,t,s}\in\mathcal{G}_1$, $\Gamma^{(c)}_{y,t,s}\in\kappa\mathcal{G}_2\cup\{M\}$ up to global phase $\alpha_{y,t,s}$, where $\kappa\mathcal{G}_2 = \{\kappa\Gamma \mid \Gamma\in\mathcal{G}_2\}$.

- *For every $|\Psi_{y,t,s}\rangle\in S$ and every gate $\Gamma\in\mathcal{G}_1\cup\mathcal{G}_2\cup\{H_0\}$, there exist amplitudes $\{a_{y,t,s}\}$ such that*

$$e^{i\Gamma\theta}|\Psi_{y,t,s}\rangle = \sum_{\substack{(y',t',s')\in I_S\\ W(y',t',s')\le W(y,t,s)+1}} a_{y,t,s}|\Psi_{y',t',s'}\rangle \tag{81}$$

  In words, the application of g can only increase value of the W-function by at most 1.

Proof. We proceed by case analysis. The first claim will follow if for every $|\Psi_{y,t,s}\rangle$, $|\Psi_{y,t,s}\rangle$ is an eigenvector of all but (at most one) generator $\Gamma^{(i)}_{y,t,s}$ comprising H_i. To see why, define $H^{(i)}_{y,t,s} \coloneqq H_i - \Gamma^{(i)}_{y,t,s}$. Then:

$$e^{i\theta H_i}|\Psi_{y,t,s}\rangle = e^{i\theta\left(\Gamma^{(i)}_{y,t,s}+H^{(i)}_{y,t,s}\right)}|\Psi_{y,t,s}\rangle = e^{i\theta\Gamma^{(i)}_{y,t,s}}e^{i\theta H^{(i)}_{y,t,s}}|\Psi_{y,t,s}\rangle = e^{i\theta\Gamma^{(i)}_{y,t,s}}e^{i\theta\alpha^{(i)}_{y,t,s}}|\Psi_{y,t,s}\rangle, \tag{82}$$

Where $\alpha^{(i)}_{y,t,s}$ is the corresponding eigenvalue of $H^{(i)}_{y,t,s} \coloneqq H_i - \Gamma^{(i)}_{y,t,s}$. The second step uses that all generators in H_i commute with each other restricted to states where C and D are in

valid logical time states, which one can verify by direct calculation. As for the second claim of the lemma, it will follow if every generator maps only between states $|\Psi_{y,t,s}\rangle \leftrightarrow |\Psi_{y',t',s'}\rangle$ with $|W(y,t,s) - W(y',t',s')| \leq 1$. To obtain these claims, we first list all non-trivial generator transitions of H_b, where one can transition between various states $|\Psi_{y,t,s}\rangle$ listed below. For example, in Equation (83) and Equation (84), F_i can map $|y\rangle_A|0\rangle_B|\widetilde{2j}\rangle_C|\widetilde{i}\rangle_D$ to $|y\rangle_A|0\rangle_B|\widetilde{2j}\rangle_C|\widetilde{i+1}\rangle_D$ and vice versa. To the right of each of these states, we list the value of the W-function for that state (which, recall, is only a function of (y,t,s)).

- $(F_i,\ i \in \{3,5\ldots,|D|-1\}),\ \forall j \in [|A|],\ y \in \{0,1\}^{|A|}$

$$|y\rangle_A|0\rangle_B|\widetilde{2j}\rangle_C|\widetilde{i}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2j + (-1)^{\delta_{y_{\lceil t/2\rceil},1}}(i-2) \tag{83}$$

$$|y\rangle_A|0\rangle_B|\widetilde{2j}\rangle_C|\widetilde{i+1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2j + (-1)^{\delta_{y_{\lceil t/2\rceil},1}}(i-1) \tag{84}$$

- $(G_i,\ i \in \{2,4\ldots,2|A|\}),\ \forall y \in \{0,1\}^{|A|}$

$$|y\rangle_A|0\rangle_B|\widetilde{i}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + i \tag{85}$$

$$|y\rangle_A|0\rangle_B|\widetilde{i+1}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + i + 1 \tag{86}$$

$$|y\rangle_A|0\rangle_B|\widetilde{i}\rangle_C|\widetilde{2}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + i \tag{87}$$

- $(Q_i,\ i \in \{2,4,\ldots,L\}),\ \forall y \in \{0,1\}^{|A|}$

$$V_{i-1}\cdots V_1|y\rangle_A|0\rangle_B|2|\widetilde{A|+1+i}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2|A| + 1 + i \tag{88}$$

$$V_{i}\cdots V_1|y\rangle_A|0\rangle_B|2|\widetilde{A|+2+i}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2|A| + 2 + i \tag{89}$$

  We note that, by Assumption 6, since V_i can only be controlled via register A (as opposed to acting on A as a target register), it cannot change the string y in A. Thus, W is only affected via the change on the C register.

- $(H_0),\ \forall y \in \{0,1\}^{|A|}$

$$|y\rangle_A|0\rangle_B|\widetilde{1}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 1 \tag{90}$$

In all cases above, the change in W is at most 1, every logical state $|\Psi_{y,t,s}\rangle$ in S is appears in precisely one set, and H_0 always acts trivially, proving the lemma for H_b. Repeating this approach for H_c yields a similar conclusion:

- $(F_i,\ i \in \{2,4\ldots,|D|-2\}),\ \forall j \in [|A|],\ y \in \{0,1\}^{|A|}$

$$|y\rangle_A|0\rangle_B|\widetilde{2j}\rangle_C|\widetilde{i}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2j + (-1)^{\delta_{y_{\lceil t/2\rceil},1}}(i-2) \tag{91}$$

$$|y\rangle_A|0\rangle_B|\widetilde{2j}\rangle_C|\widetilde{i+1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2j + (-1)^{\delta_{y_{\lceil t/2\rceil},1}}(i-1) \tag{92}$$

- $(G_i,\ i \in \{1,3\ldots,2|A|-1\}),\ \forall y \in \{0,1\}^{|A|}$

$$|y\rangle_A|0\rangle_B|\widetilde{i}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + i \tag{93}$$

$$|y\rangle_A|0\rangle_B|\widetilde{i+1}\rangle_C|\widetilde{1}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + i + 1 \tag{94}$$

- $(P_i,\ i \in \{1,\ldots,|A|\}),\ \forall y \in \{\{0,1\}^{|A|} | y_i = 0\}$

$$|y\rangle_A|0\rangle_B|\widetilde{2i}\rangle_C|\widetilde{|D|}\rangle_D: \quad W = (2|D|-4)\mathrm{HW}(y) + 2i + (|D|-2) \tag{95}$$

$$|y\oplus e_i\rangle_A|0\rangle_B|\widetilde{2i}\rangle_C|\widetilde{|D|}\rangle_D: \quad W = (2|D|-4)(\mathrm{HW}(y)+1) + 2i - (|D|-2) \tag{96}$$

- $(Q_i,\, i \in \{1,3,\ldots,L-1\})$, $\forall y \in \{0,1\}^{|A|}$

$$V_{i-1}\cdots V_1|y\rangle_A|0\rangle_B|2\,|\widetilde{A|+1}+i\rangle_C|\widetilde{1}\rangle_D: \quad W=(2\,|D|-4)\mathrm{HW}(y)+2\,|A|+1+i \tag{97}$$

$$V_{i}\cdots V_1|y\rangle_A|0\rangle_B|2\,|\widetilde{A|+2}+i\rangle_C|\widetilde{1}\rangle_D: \quad W=(2\,|D|-4)\mathrm{HW}(y)+2\,|A|+2+i \tag{98}$$

- (M), $\forall y \in \{0,1\}^{|A|}$

$$V_L\cdots V_1|y\rangle_A|0\rangle_B||\widetilde{C}|\rangle_C|\widetilde{1}\rangle_D: \quad W=(2\,|D|-4)\mathrm{HW}(y)+|C| \tag{99}$$

$$Z_{B_1}V_L\cdots V_1|y\rangle_A|0\rangle_B||\widetilde{C}|\rangle_C|\widetilde{1}\rangle_D: \quad W=(2\,|D|-4)\mathrm{HW}(y)+|C| \tag{100}$$

Note that the first claim of the lemma indeed holds for M, but since $Z_{B_1}V_L\cdots V_1|y\rangle_A|0\rangle_B||\widetilde{C}|\rangle_C|\widetilde{1}\rangle_D \notin \mathrm{span}(S)$, the second claim does not (and thus why we write $\Gamma \in \mathcal{G}_1 \cup \mathcal{G}_2 \cup \{H_0\}$ in the second claim). □

With this we can proceed to show completeness and soundness.

4.3 Completeness

In the YES case, there exists a sequence of gates with proof $y \in \{0,1\}^{|A|}$ of Hamming weight at most g accepted with probability at least $1-\epsilon_Q$ by V. We use shorthand $(y)_j = (y_1,\ldots y_{j-1},0,\ldots,0)$ to indicate the partially written proof string. Also, $\exp(\mathrm{i}\theta H_i) \sim \exp(\mathrm{i}\theta\Gamma)$ indicates which generator (Γ) in H_i performs the non-trivial operation (as per Lemma 4, claim 1). The honest prover proceeds as follows:

- (Prepare classical proof) Prepare state (up to global phase) $|\psi_0\rangle \coloneqq |y\rangle_A|0\rangle_B|2\,|\widetilde{A|+1}\rangle_C|\widetilde{1}\rangle_D$ as follows. Starting with $|\mathrm{gs}_b\rangle = |(y)_0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD}$:
  1. Set $j=1$.
  2. Apply $\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}G_{2j-1})$ to map $|\widetilde{2j-1}\rangle_C \to |\widetilde{2j}\rangle_C$. This maps
  $$|(y)_{j-1},0,\widetilde{2j-1},\widetilde{1}\rangle_{ABCD} \quad\mapsto\quad |(y)_{j-1},0,\widetilde{2j},\widetilde{1}\rangle_{ABCD}. \tag{101}$$
  3. If $y_j=1$ then
     - Apply $\exp(\mathrm{i}\frac{2\pi}{3}H_b) \sim \exp(\mathrm{i}\frac{2\pi}{3}G_{2j})$, to map $|\widetilde{1}\rangle_D \to |\widetilde{2}\rangle_D$, i.e.
     $$|(y)_{j-1},0,\widetilde{2j},\widetilde{1}\rangle_{ABCD} \quad\mapsto\quad |(y)_{j-1},0,\widetilde{2j},\widetilde{2}\rangle_{ABCD}. \tag{102}$$
     - Apply, in order, $\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}F_2)$, $\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}F_3),\ldots,\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}F_{|D|-1})$, in total $|D|-2$ operations . This maps $|\widetilde{2}\rangle_D \to ||\widetilde{D}|\rangle_D$, i.e.
     $$|(y)_{j-1},0,\widetilde{2j},\widetilde{2}\rangle_{ABCD} \quad\mapsto\quad |(y)_{j-1},0,\widetilde{2j},|\widetilde{D}|\rangle_{ABCD}. \tag{103}$$
     - Apply $\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}P_j)$, to map $|0\rangle_{A_j} \to |1\rangle_{A_j}$, i.e.
     $$|(y)_{j-1},0,\widetilde{2j},|\widetilde{D}|\rangle_{ABCD} \quad\to\quad |(y)_{j},0,\widetilde{2j},|\widetilde{D}|\rangle_{ABCD}. \tag{104}$$
     - Apply, in order, $\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}F_{|D|-1})$,$\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}F_{|D|-2})$ $\ldots$, $\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}F_2)$, in total $|D|-2$ operations. This maps $||\widetilde{D}|\rangle_D \to |\widetilde{2}\rangle_D$, i.e.
     $$|(y)_{j},0,\widetilde{2j},|\widetilde{D}|\rangle_{ABCD} \quad\mapsto\quad |(y)_{j},0,\widetilde{2j},\widetilde{2}\rangle_{ABCD}. \tag{105}$$

– Apply $\exp(\mathrm{i}\frac{2\pi}{3}H_b) \sim \exp(\mathrm{i}\frac{2\pi}{3}G_{2j})$, to map $|\widetilde{2}\rangle_D \to |\widetilde{1}\rangle_D$ and $|\widetilde{2j}\rangle_C \to |\widetilde{2j+1}\rangle_C$, i.e.

$$|(y)_j, 0, \widetilde{2j}, \widetilde{2}\rangle_{ABCD} \quad \mapsto \quad |(y)_j, 0, \widetilde{2j+1}, \widetilde{1}\rangle_{ABCD} \tag{106}$$

4. else
   – Apply $\exp(\mathrm{i}\frac{\pi}{3}H_b) \sim \exp(\mathrm{i}\frac{\pi}{3}G_{2j})$, to map $|\widetilde{2j}\rangle_C \mapsto |\widetilde{2j+1}\rangle_C$, i.e.

$$|(y)_{j-1}, 0, \widetilde{2j}, \widetilde{1}\rangle_{ABCD} \quad \mapsto \quad |(y)_j, 0, \widetilde{2j+1}, \widetilde{1}\rangle_{ABCD}. \tag{107}$$

5. Set $j = j + 1$.
6. If $j < |A|$, return to line 2 above.

This process applies $2g(|D| - 1) + 2|A|$ gates.

- (Simulate verifier) Apply in order, $\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}Q_1)$, $\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}Q_2)$, ...,$\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}Q_L)$ for a total L gates. This implements the verification circuit, i.e.

$$|y, 0, \widetilde{2|A|+1}, \widetilde{1}\rangle_{ABCD} \quad \to \quad |\Psi_L\rangle := V_L \cdots V_1 |y, 0, |\widetilde{C}|, \widetilde{1}\rangle_{ABCD}. \tag{108}$$

Since V accepts proof y of the QMSA instance with probability at least $1 - \epsilon_Q$, we conclude

$$\langle\Psi_L|H_c|\psi_L\rangle \leq \langle\Psi_L|M|\Psi_L\rangle + \frac{1}{12} \leq 1 - (1 - \epsilon_Q) + \frac{1}{12} \leq \frac{1}{3}, \tag{109}$$

as desired, first the first inequality follows from Equation (74). The number of Hamiltonians applied in this case is $m = g(2|D| - 2) + 2|A| + L = g(2|D| - 2) + |C| - 1$, as desired.

4.4 Soundness

In the proof of Theorem 1 for MIN-VQA, we showed that all Hamiltonian evolutions keep us in our desired logical computation space, S. In contrast, here for MIN-QAOA, the M operator (embedded in H_c) does *not* necessarily preserve the space span(S) (see Claim 2 of Lemma 4). We thus first require the following lemma, which allows us to "round" our intermediate state back to one in S for our analysis and also establishes $W(y, t, s)$ as a proper lower bound for the number of gate applications required to reach the states in S.

Lemma 5 (Rounding lemma). *In the NO case, after $\zeta \geq 1$ applications of H_c and H_b, the state*

$$|\Psi_\zeta\rangle \in \Gamma_\zeta := \left\{ \prod_{i=1}^{\zeta} e^{iH_i\theta_i} |\mathrm{gs}_b\rangle \mid H_i \in \{H_b, H_c\}, \theta \in \mathbb{R}^\zeta \right\} \tag{110}$$

will be $\epsilon \leq 4\zeta\sqrt{\epsilon_Q}$ close to the span of S i.e.

$$\forall |\Psi_\zeta\rangle \in \Gamma_\zeta, \ \exists |\Psi'_\zeta\rangle \in \mathrm{span}(S) : \left\| |\Psi_\zeta\rangle\langle\Psi_\zeta| - |\Psi'_\zeta\rangle\langle\Psi'_\zeta| \right\|_{\mathrm{tr}} \leq 4\zeta\sqrt{\epsilon_Q} \tag{111}$$

and it additionally holds that

$$|\Psi'_\zeta\rangle = \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\leq\zeta+1}} a_{y,t,s} |\Psi_{y,t,s}\rangle. \tag{112}$$

This lemma is needed because the time evolution of the observable M (in H_c) may leave the sub-space Span(S). The rounding step is possible, because in the NO case, the state in the B_1 register, after applying the circuit V ($\tilde{s} = |D|$), is always close to $|0\rangle_{B_1}$ (using Assumption 11), meaning the evolution in M only adds to a global phase.

Proof. For our construction we use

$$|\Psi'_{\zeta+1}\rangle = e^{\mathrm{i}\theta_{\zeta+1}(H_{\zeta+1} - M\delta_{H_{\zeta+1},H_c})}|\Psi'_\zeta\rangle \tag{113}$$

i.e. the same VQA but without the M generator in H_c. We show the proof by induction. The lemma statement holds trivially for the base case $\zeta = 0$, since

$$|\Psi_{\zeta=0}\rangle = |\Psi'_{\zeta=0}\rangle = |0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} = \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\leq 1}} a_{y,t,s}|\Psi_{y,t,s}\rangle \tag{114}$$

with $W(0,1,1) = 1$.

Induction step: For the norm inequality (Equation (111)), only M maps states outside of S, meaning we only have to consider the action of H_c. Then,

$$\begin{aligned}
\left\||\Psi_{\zeta+1}\rangle\langle\Psi_{\zeta+1}| - |\Psi'_{\zeta+1}\rangle\langle\Psi'_{\zeta+1}|\right\|_{\mathrm{tr}} &= \left\|e^{\mathrm{i}H_c\theta}|\Psi_\zeta\rangle\langle\Psi_\zeta|e^{-\mathrm{i}H_c\theta} - e^{\mathrm{i}(H_c-M)\theta}|\Psi'_\zeta\rangle\langle\Psi'_\zeta|e^{-\mathrm{i}(H_c-M)\theta}\right\|_{\mathrm{tr}} \\
&= \left\||\Psi_\zeta\rangle\langle\Psi_\zeta| - e^{-\mathrm{i}M\theta}|\Psi'_\zeta\rangle\langle\Psi'_\zeta|e^{\mathrm{i}M\theta}\right\|_{\mathrm{tr}} \\
&\leq \left\||\Psi'_\zeta\rangle\langle\Psi'_\zeta| - e^{-\mathrm{i}M\theta}|\Psi'_\zeta\rangle\langle\Psi'_\zeta|e^{\mathrm{i}M\theta}\right\|_{\mathrm{tr}} + \left\||\Psi'_\zeta\rangle\langle\Psi'_\zeta| - |\Psi_\zeta\rangle\langle\Psi_\zeta|\right\|_{\mathrm{tr}} \\
&\leq \left\||\Psi'_\zeta\rangle\langle\Psi'_\zeta| - e^{-\mathrm{i}M\theta}|\Psi'_\zeta\rangle\langle\Psi'_\zeta|e^{\mathrm{i}M\theta}\right\|_{\mathrm{tr}} + 4\zeta\epsilon_Q,
\end{aligned}$$

where the second statement holds since $[H_c, M] = 0$ and by the unitary invariance of the trace norm, the third by the triangle inequality, and the fourth by the induction hypothesis. Now, for M we have the following non-trivial action:

$$\begin{aligned}
e^{-i\theta}e^{iM\theta}|\Psi'_\zeta\rangle - |\Psi'_\zeta\rangle &= \sum_{(y,t,s)\in I_S} a_{y,t,s}(e^{\mathrm{i}(M-1)\theta} - 1)|\Psi_{y,t,s}\rangle &(115)\\
&= \sum_{y\in\{0,1\}^{|A|}} a_{y,|C|,1}(e^{-\mathrm{i}\theta|1\rangle\langle 1|_{B_1}} - 1)|\Psi_{y,|C|,1}\rangle &(116)\\
&= (e^{-i\theta} - 1)\sum_{y\in\{0,1\}^{|A|}} a_{y,|C|,1}|1\rangle\langle 1|_{B_1}|\Psi_{y,|C|,1}\rangle, &(117)
\end{aligned}$$

where we used that M only acts non-trivially for $t = |C|$. This means

$$\begin{aligned}
\left\||\Psi'_\zeta\rangle\langle\Psi'_\zeta| - e^{-\mathrm{i}M\theta}|\Psi'_\zeta\rangle\langle\Psi'_\zeta|e^{\mathrm{i}M\theta}\right\|_{\mathrm{tr}} &\leq 2\left\|e^{-\mathrm{i}\theta}e^{iM\theta}|\Psi'_\zeta\rangle - |\Psi'_\zeta\rangle\right\|_2 &(118)\\
&\leq 4\sqrt{\sum_{y\in\{0,1\}^{|A|}} |a_{y,|C|,1}|^2\langle\Psi_{y,|C|,1}||1\rangle\langle 1|_{B_1}|\Psi_{y,|C|,1}\rangle} &(119)\\
&\leq 4\sqrt{\sum_{y\in\{0,1\}^{|A|}} |a_{y,|C|,1}|^2\epsilon_Q} \leq 4\sqrt{\epsilon_Q}, &(120)
\end{aligned}$$

where the first line is a known norm inequality[20] and we used that $\langle\Psi_{y,|C|,1}||1\rangle\langle 1|_{B_1}|\Psi_{y,|C|,1}\rangle \leq \epsilon_Q$ is the acceptance probability of a QMSA NO instance. This shows the first claim of the lemma. For the second claim, a similar induction setup, coupled with Lemma 4, yields

[20] due to $\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_{\mathrm{tr}} = 2\sqrt{1 - |\langle\psi|\phi\rangle|^2}$ and $\||\psi\rangle - |\phi\rangle\|_2 = \sqrt{2 - 2\,\mathrm{Re}(\langle\psi|\phi\rangle)}$ for all $|\psi\rangle, |\phi\rangle \in \mathcal{H}$ and $\sqrt{1-x^2} \leq \sqrt{2-2x} \quad \forall x \in [0,1]$.

$$|\Psi'_{\zeta+1}\rangle = e^{i\theta_{\zeta+1}(H_{\zeta+1}-M\delta_{H_{\zeta+1},H_c})}|\Psi'_\zeta\rangle \tag{121}$$

$$= \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\le\zeta+1}} a_{y,t,s} e^{i\theta_{\zeta+1}(\Gamma^{(\zeta+1)}_{y,t,s}+\alpha^{(\zeta+1)}_{y,t,s})}|\Psi_{y,t,s}\rangle \tag{122}$$

$$= \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\le\zeta+1}} a_{y,t,s} \sum_{\substack{(y',t',s')\in I_S \\ W(y',t',s')\le W(y,t,s)+1}} b^{(y,t,s)}_{y',t',s'}|\Psi_{y',t',s'}\rangle \tag{123}$$

$$= \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\le\zeta+2}} a'_{y,t,s}|\Psi_{y,t,s}\rangle, \tag{124}$$

where the second and third statements follow from the first and second claims of Lemma 4, respectively, and the last statement just recombines the a and b indices into new indices a'. □

We are finally ready to prove soundness. For this, we need to show that in the NO case, all sequences of $\zeta \le m' = g'(2|D|-4)+|C|-1$ gates produce cost function value $\langle\Psi_\zeta|H_c|\Psi_\zeta\rangle \ge \frac{2}{3}$. This follows since for all $\zeta \le m'$,

$$\langle\Psi_\zeta|H_c|\Psi_\zeta\rangle \ge \langle\Psi_\zeta|M|\Psi_\zeta\rangle - \frac{1}{12} \tag{125}$$

$$\ge \langle\Psi'_\zeta|M|\Psi'_\zeta\rangle - |\mathrm{Tr}[M(|\Psi_\zeta\rangle\langle\Psi_\zeta| - |\Psi'_\zeta\rangle\langle\Psi'_\zeta|)]| - \frac{1}{12} \tag{126}$$

$$\ge \langle\Psi'_\zeta|M|\Psi'_\zeta\rangle - \|M\|_\infty \left\||\Psi_\zeta\rangle\langle\Psi_\zeta| - |\Psi'_\zeta\rangle\langle\Psi'_\zeta|\right\|_{\mathrm{tr}} - \frac{1}{12} \tag{127}$$

$$\ge \langle\Psi'_\zeta|M|\Psi'_\zeta\rangle - 4m'\sqrt{\epsilon_Q} - \frac{1}{12} \tag{128}$$

$$\ge \langle\Psi'_\zeta|M|\Psi'_\zeta\rangle - \frac{1}{6} \tag{129}$$

where the first statement follows from Equation (74), the third by Hölder's inequality, the fourth by Lemma 5, and the last since $\sqrt{\epsilon_Q} \le \frac{1}{48m'}$. By Lemma 4, we can expand $|\Psi'_\zeta\rangle$ in the basis

$$|\Psi'_\zeta\rangle = \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\le m'+1}} a_{y,t,s}|\Psi_{y,t,s}\rangle \tag{130}$$

which gives

$$\langle\Psi'_\zeta|M|\Psi'_\zeta\rangle = 1 - \sum_{y\in\{0,1\}^{|A|}|\mathrm{HW}(y)\le g'} |a_{y,|C|,1}|^2\langle\Psi_{y,|C|,1}||1\rangle\langle 1|_{B_1}|\Psi_{y,|C|,1}\rangle \ge 1-\epsilon_Q \tag{131}$$

as M only acts non-trivial on $t=|C|$ and $W(y,|C|,1) \le m'+1$ reduces to $\mathrm{HW}(y) \le g'$, and in the NO case QMSA accepts such a y with at most ϵ_Q probability. Combining the two results we get

$$\langle\Psi_\zeta|H_c|\Psi_\zeta\rangle \ge 1-\epsilon_Q-\frac{1}{6} > \frac{2}{3} \tag{132}$$

which shows soundness for all gates-sequences of length $\zeta \le m'$.

4.5 Hardness ratio

The analysis is essentially identical to that for MIN-VQA, so we sketch it briefly. Since we set $|D| = \lceil L^{1+\delta} \rceil$, we have that in ratio

$$\frac{m'}{m} = \frac{g'(2|D| - 4) + |C| - 1}{g(2|D| - 2) + |C| - 1}, \tag{133}$$

the dominant term is again $|D|$. Thus, $m'/m \approx g'/g \geq O((N')^{1-\epsilon'})$, for N' the encoding size of the QMSA instance, and for any desired $\epsilon' > 0$. Since the encoding size of H_b and H_c can also be seen to scale as $O((L')^{1+\delta})$ (recall L' the number of gates in the original QMSA circuit V'), we can apply Equation (55) and the surrounding approximation ratio analysis from MIN-VQA to argue again that $N \approx O((N')^{1+\epsilon'})$, for N the encoding size of our MIN-QAOA instance with logarithmic overhead to satisfy Assumption 11 and only $O(L)$ overhead for the changes performed to the gate set. Thus, for any desired $\epsilon > 0$, we may choose $\epsilon' > 0$ and $\delta > 0$ so that $g'/g \geq (N')^{1-\epsilon'} \geq N^{1-\epsilon}$, as desired.

□

Acknowledgements

LB and MK are supported by the German Federal Ministry of Education and Research (BMBF) within the funding program "Quantum Technologies – from Basic Research to Market" via the joint project MANIQU (grant number 13N15578) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the grant number 441423094 within the Emmy Noether Program. SG was supported by the DFG under grant numbers 450041824 and 432788384, the BMBF within the funding program "Quantum Technologies – from Basic Research to Market" via project PhoQuant (grant number 13N16103), and the project "PhoQC" from the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors.

References

[AAV13] Dorit Aharonov, Itai Arad, and Thomas Vidick. "Guest column: the quantum PCP conjecture". In: *SIGACT News* 44.2 (June 2013), pp. 47–79. arXiv: 1309.7495 [quant-ph].

[AK22] Eric R. Anschuetz and Bobak T. Kiani. "Beyond barren plateaus: quantum variational algorithms are swamped with traps". arXiv: 2205.05786 [quant-ph]. 2022.

[AM22] Anurag Anshu and Tony Metger. "Concentration bounds for quantum states and limitations on the QAOA from polynomial approximations". arXiv: 2209.02715 [quant-ph]. 2022.

[AN02] Dorit Aharonov and Tomer Naveh. "Quantum NP - a survey". arXiv: quant-ph/0210077 [quant-ph]. 2002.

[Aro+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. "Proof verification and the hardness of approximation problems". In: *J. ACM* 45.3 (1998). Prelim. version FOCS '92, pp. 501–555.

[AS98] Sanjeev Arora and Shmuel Safra. "Probabilistic checking of proofs: a new characterization of NP". In: *J. ACM* 45.1 (1998). Prelim. version FOCS '92, pp. 70–122.

[AT03] Dorit Aharonov and Amnon Ta-Shma. "Adiabatic quantum state generation and statistical zero knowledge". In: *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*. STOC '03. San Diego, CA, USA: Association for Computing Machinery, 2003, pp. 20–29. arXiv: quant-ph/0301023 [quant-ph].

[Bas+22] Joao Basso, David Gamarnik, Song Mei, and Leo Zhou. "Performance and limitations of the QAOA at constant levels on large sparse hypergraphs and spin glass models". arXiv: 2204.10306 [quant-ph]. 2022.

[Bha+22] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. "Noisy intermediate-scale quantum (NISQ) algorithms". In: *Rev. Mod. Phys.* 94.1, 015004 (Jan. 2022), p. 015004. arXiv: 2101.08448 [quant-ph].

[BK21] Lennart Bittel and Martin Kliesch. "Training variational quantum algorithms is NP-hard". In: *Phys. Rev. Lett.* 127 (12 Sept. 2021), p. 120502. arXiv: 2101.07267 [quant-ph].

[BK22] Gregory Boyd and Bálint Koczor. "Training variational quantum circuits with covar: covariance root finding with classical shadows". arXiv: 2204.08494 [quant-ph]. 2022.

[BM22] Sami Boulebnane and Ashley Montanaro. "Solving boolean satisfiability problems with the quantum approximate optimization algorithm". arXiv: 2208.06909 [quant-ph]. 2022.

[Bra+20] Sergey Bravyi, Alexander Kliesch, Robert Koenig, and Eugene Tang. "Obstacles to variational quantum optimization from symmetry protection". In: *Phys. Rev. Lett.* 125 (26 Dec. 2020), p. 260505. arXiv: 1910.08980 [quant-ph].

[Cer+21] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. "Variational quantum algorithms". In: *Nat. Rev. Phys.* 3 (2021), pp. 625–644. arXiv: 2012.09265 [quant-ph].

[Fey86] Richard P Feynman. "Quantum mechanical computers". In: *Found. Phys.* 16.6 (1986), pp. 507–531.

[FGG14] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm". arXiv: 1411.4028 [quant-ph]. 2014.

[FH16] Edward Farhi and Aram W Harrow. "Quantum supremacy through the quantum approximate optimization algorithm". arXiv: 1602.07674 [quant-ph]. 2016.

[GK12] Sevag Gharibian and Julia Kempe. "Hardness of approximation for quantum problems". In: *39th International Colloquium on Automata, Languages and Programming (ICALP)*. 2012, pp. 387–398. arXiv: 1209.1055 [quant-ph].

[GMV17] David Gosset, Jenish C. Mehta, and Thomas Vidick. "QCMA hardness of ground space connectivity for commuting Hamiltonians". In: *Quantum* 1 (July 2017), p. 16. arXiv: 1610.03582 [cs.CC].

[Gri+19] Harper R. Grimsley, Sophia E. Economou, Edwin Barnes, and Nicholas J. Mayhall. "An adaptive variational algorithm for exact molecular simulations on a quantum computer". In: *Nat. Commun.* 10, 3007 (July 2019), p. 3007. arXiv: 1812.11173 [quant-ph].

[Gri+22] Harper R. Grimsley, George S. Barron, Edwin Barnes, Sophia E. Economou, and Nicholas J. Mayhall. "ADAPT-VQE is insensitive to rough parameter landscapes and barren plateaus". arXiv: 2204.07179 [quant-ph]. 2022.

[GS15] Sevag Gharibian and Jamie Sikora. "Ground state connectivity of local Hamiltonians". In: *42nd International Colloquium on Automata, Languages, and Programming (ICALP)*. 2015, pp. 617–628. arXiv: 1409.3182 [quant-ph].

[GW95] Michel X Goemans and David P Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *J. ACM* 42 (1995), pp. 1115–1145.

[Had+19] Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G. Rieffel, Davide Venturelli, and Rupak Biswas. "From the quantum approximate optimization algorithm to a quantum alternating operator ansatz". In: *Algorithms* 12.2 (2019), p. 34. arXiv: 1709.03489 [quant-ph].

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". In: *Nature Physics* 16 (June 2020), pp. 1050–1057. arXiv: 2002.08953 [quant-ph].

[KB22] Bálint Koczor and Simon C. Benjamin. "Quantum analytic descent". In: *Phys. Rev. Research* 4.2, 023017 (Apr. 2022), p. 023017. arXiv: 2008.13774 [quant-ph].

[KSV02] Alexei Yu Kitaev, Alexander Shen, and Mikhail N Vyalyi. *Classical and quantum computation*. Vol. 47. American Mathematical Society, 2002.

[Lar+21] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. "Theory of overparametrization in quantum neural networks". arXiv: 2109.11676 [quant-ph]. 2021.

[LC17] Guang Hao Low and Isaac L. Chuang. "Optimal Hamiltonian simulation by quantum signal processing". In: *Phys. Rev. Lett.* 118 (1 Jan. 2017), p. 010501. arXiv: 1606.02685 [quant-ph].

[Llo96] Seth Lloyd. "Universal quantum simulators". In: *Science* 273.5278 (1996), pp. 1073–1078.

[McC+18] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Barren plateaus in quantum neural network training landscapes". In: *Nat. Commun.* 9, 4812 (Nov. 2018), p. 4812. arXiv: 1803.11173 [quant-ph].

[Riv+21] Javier Rivera-Dean, Patrick Huembeli, Antonio Acín, and Joseph Bowles. "Avoiding local minima in variational quantum algorithms with neural networks". arXiv: 2104.02955 [quant-ph]. 2021.

[SVC22] Lucas Slattery, Benjamin Villalonga, and Bryan K. Clark. "Unitary block optimization for variational quantum algorithms". In: *Phys. Rev. Research* 4.2, 023072 (Apr. 2022), p. 023072. arXiv: 2102.08403 [quant-ph].

[TLM20] Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity. "Learning unitaries by gradient descent". arXiv: 2001.11897 [quant-ph]. 2020.

[Uma99] C. Umans. "Hardness of approximating Σ_2^p minimization problems". In: *40th Annual Symposium on Foundations of Computer Science*. 1999, pp. 465–474.

[WBG20] James D. Watson, Johannes Bausch, and Sevag Gharibian. "The complexity of translationally invariant problems beyond ground state energies". arXiv: 2012.12717 [quant-ph]. 2020.

[WGK20] David Wierichs, Christian Gogolin, and Michael Kastoryano. "Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer". In: *Phys. Rev. Research* 2.4, 043246 (Nov. 2020), p. 043246. arXiv: 2004.14666 [quant-ph].

[Wie+20] Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. "Exploring entanglement and optimization within the Hamiltonian variational ansatz". In: *PRX Quantum* 1 (2020), p. 020319. arXiv: 2008.02941 [quant-ph].

[WJB03] Pawel Wocjan, Dominik Janzing, and Thomas Beth. "Two QCMA-complete problems". In: *Quantum Information & Computation* 3.6 (2003), pp. 635–643. arXiv: quant-ph/0305090 [quant-ph].

[Zho+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". In: *Phys. Rev. X* 10 (2020), p. 021067. arXiv: 1812.01041 [quant-ph].

[ZY20] Dan-Bo Zhang and Tao Yin. "Collective optimization for variational quantum eigensolvers". In: *Phys. Rev. A* 101.3, 032311 (Mar. 2020), p. 032311. arXiv: 1910.14030 [quant-ph].

A Additional proofs

Proof of Corollary 2. Suppose there exists an algorithm A for computing estimate $m_{\text{est}} \in [m_{\text{opt}}, N^{1-\epsilon} m_{\text{opt}}]$. We show how to use A to decide MIN-VQA, yielding QCMA-hardness. Specifically, given an instance Π of MIN-VQA, run A. If A's output is less than or equal to m', accept. Otherwise, reject.

To see that this is correct, observe that in the YES case, $m_{\text{opt}} \leq m$. Since $m'/m \geq N^{1-\epsilon}$, A outputs estimate $m_{\text{est}} \leq m'$, from which we conclude Π is cannot be a NO instance, and thus must be a YES instance (due to the promise that Π is either a YES or NO instance). Conversely, in the NO case, $m_{\text{est}} \geq m_{\text{opt}} > m'$, from which we conclude Π is a NO instance. □

E

Scalable approach to many-body localization via quantum data

Title:	Scalable approach to many-body localization via quantum data
Authors:	Alexander Gresch, Lennart Bittel, Martin Kliesch
Journal:	Machine Learning: Science and Technology
Date of submission:	20 March 2022
Publication status:	In peer-review

This publication corresponds to the article [89]. The summary of the results is presented in section 4.4.1.

Contribution: The publication resulted from the master thesis of AG, who I supervised at the time. The majority of the work was done by AG in close discussions with MK and me.

Scalable approach to many-body localization via quantum data

Alexander Gresch,* Lennart Bittel, and Martin Kliesch
Quantum Technology Research Group, Heinrich Heine University Düsseldorf, Germany

We are interested in how quantum data can allow for practical solutions to otherwise difficult computational problems. A notoriously difficult phenomenon from quantum many-body physics is the emergence of many-body localization (MBL). So far, is has evaded a comprehensive analysis. In particular, numerical studies are challenged by the exponential growth of the Hilbert space dimension. As many of these studies rely on exact diagonalization of the system's Hamiltonian, only small system sizes are accessible.

In this work, we propose a highly flexible neural network based learning approach that, once given training data, circumvents any computationally expensive step. In this way, we can efficiently estimate common indicators of MBL such as the adjacent gap ratio or entropic quantities. Our estimator can be trained on data from various system sizes at once which grants the ability to extrapolate from smaller to larger ones. Moreover, using transfer learning we show that already a two-dimensional feature vector is sufficient to obtain several different indicators at various energy densities at once. We hope that our approach can be applied to large-scale quantum experiments to provide new insights into quantum many-body physics.

I. INTRODUCTION

The goal of quantum computing is to efficiently solve practically relevant problems that are intractable on classical computers. Many those problems require a fault-tolerant, universal quantum computer. This requirement, in turn, comes in conjunction with the need for quantum error correction which yields a daunting overhead in the qubit numbers. Both requirements exceed the current available quantum hardware substantially. Hence, in the meantime, the potential of hybrid quantum algorithms is explored. They aim to optimally use the few dozens of available qubits with no or little error mitigation schemes. Most of their pragmatic approaches are centered around variational quantum algorithms (VQAs) [1, 2]. These algorithms provide heuristics for problems such as finding the ground-state energy in the field of quantum chemistry [3] or solving combinatorial problems [4]. Even though the encountered practical constraints impose a tall hurdle, those efforts appear promising for near-future applications. Such hopes are furthermore fueled by the achievements in the field of deep learning, especially during the last decade. Despite the absence of rigorous performance guarantees, there has been a tremendous success of deep learning methods in diverse fields ranging from computer vision, natural language processing to finance and beyond [5].

Over the last year, rigorous performance guarantees for machine-learning-based approaches to quantum many-body physics have been found [6–8]. These findings suggest that machine learning algorithms are well suitable to generalize efficiently on *quantum data* that is obtained by quantum experiments or a quantum simulation. In particular, with the recent development in hybrid quantum algorithms such as the variational quantum eigensolver (VQE) [3, 9], variational methods become interesting, viable experimental alternatives. Alterations to the originally proposed scheme allow for the study of a few eigenvalues and -states around a target energy [10] which does not need to be the ground state [11]. The VQE's setting suits the study of MBL quite well [12].

To demonstrate the importance of the quantum data, difficult problems from quantum physics are needed. These problems are rendered as such because of their evasive behavior under analytical or numerical analyses. One of such notoriously difficult problems is the phenomenon of localization in interacting quantum many-body systems, known as MBL [13–15], see e.g. Refs. [16–18] for reviews. It originates from the well-known Anderson model of non-interacting fermions in a disordered potential where localization occurs above a certain disorder threshold [19]. The seminal works [13, 20] proved the sur-

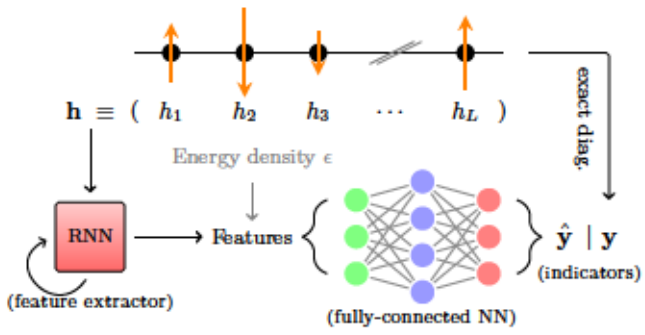

FIG. 1. Workflow for training our model architecture to predict indicator values $\hat{\mathbf{y}}$ from the system's disorder vector $\mathbf{h}$. We pass the latter into a recurrent neural network as in Fig. 2 which extracts general features of $\mathbf{h}$ in a scalable fashion. These features can be augmented by the respective energy density ϵ we are considering. Together, they are fed into a fully-connected neural network that maps them to $\hat{\mathbf{y}}$. They are compared to the results $\mathbf{y}$ obtained from exactly diagonalizing the system's Hamiltonian in the corresponding energy density ϵ.

* alexander.gresch@hhu.de

vival of the localization under the introduction of a weak interaction in terms of a perturbation. This localization can be pinpointed to the emergence of macroscopically many conserved quantities [16, 21–24] that suppress the flow of correlations through the system. In the regime of strong interactions (or conversely, a negligible disordered potential), MBL does not occur which indicates a phase transition between the MBL phase and the delocalized one. The latter can be explored deploying e.g. classically motivated ergodic arguments [25]. However, little is known about the transition region between the two phases and its underlying mechanism. The emergence of MBL connects to the fundamental question of thermalization in quantum mechanics [26–28], possibly bridged by the eigenstate thermalization hypothesis (ETH) [16, 29]. Numerical studies of the transition either apply exact diagonalization methods [14, 30, 31] or approximate methods such as renormalization group techniques [32]. Around the presumed transition region between the two phases, the numerical methods suffer from the curse of dimensionality because the Hilbert space dimension grows exponentially with the chain length L. Moreover, a numerical extrapolation to the thermodynamic limit at which the transition is expected to be chararacterized by a single value for the critical disorder parameter h_c is hampered by finite-size effects [33].

A. Related works

The idea of applying neural networks (NNs) to physical problems and, in particular, phase classification, arises as a consequence of its success with feature extraction e.g. for conventional image classification, where the classifiers could achieve a higher prediction accuracy than human test groups [34]. It has led to a surge of explorations in applying similar methods to difficult problems in (quantum) many-body physics [35–38]. The phenomenon of MBL, in particular, has attracted many numerical approaches using machine learning [39–41] or deep learning [42–45]. The previous attempts typically utilized NNs for the phase classification in order to extract a phase diagram of the transition in an energy-density- and disorder-parameter-resolved way. Employing a recurrent neural network (RNN) to study the behavior of MBL was – to the best of our knowledge – first accomplished by Ref. [43] who trace the temporal evolution of an observable as a phase classification task. In variation to those approaches, we propose to employ an RNN to characterize a given instance of the Hamiltonian's components in terms of quantum data. For the characterization, there has been an explorative work done by Nieuwenburg, Baum, and Refael [45] in the same direction. They show the learnability of the adjacent gap ratio by means of convolutional NNs from the disorder vector joined with the corresponding disorder parameter, i.e. from $\mathbf{h} \oplus h$ [45, Appendix]. Their efforts, however, resort to a proof-of-principle demonstration and use it for data augmentation. Moreover, their architecture is not scalable in the system size L because the output size of the convolutional layers grows linearly with L. Such convolutional layers can be made scalable with the input size as demonstrated by Saraceni, Cantori, and Pilati [46]. They propose an architecture where the number of extracted features does not grow with the input size and can thus be mapped to a fixed output size. Apart from this last instance, all the previous methods are restricted to a given, fixed chain length and therefore not applicable to data from a larger system. Another bottleneck is the fact that the typical input for these approaches consists of heavily preprocessed data such as the entanglement spectrum [42] or even a whole eigenvector of the Hamiltonian [44]. Both are obtained by exact diagonalization and thus lack a feasible source of training data from the transition regime for system sizes $L \gtrapprox 20$.

B. Our contribution

In this work, we propose an NN-architecture that is both applicable to data from different system sizes and not necessitating any computationally costly preprocessing of the input data. We accomplish this by directly presenting the local disorder values $\mathbf{h} = (h_1, \ldots, h_L)$ to an RNN. This step lifts the system size constraint by treating $\mathbf{h}$ as a sequence of inputs such that the sequence length corresponds to the system size. The output of the RNN serves as the extracted feature vector from the disorder sequence. Typically, such features do not yet resemble the indicators. Rather, they are global properties of the input which are not tied to a specific regression task. This view is adapted from results in computer vision where the first layers of image classifying networks merely detect edges and corners, independent of the underlying classification problem [47]. Hence, we use a final fully-connected NN as sketched in Fig. 1 that maps the extracted features to the indicator estimates. With this choice for our architecture, we can investigate in the features further by means of *transfer learning* [48]. To this end, we show that a set of features extracted from some indicators can be generalized to other previously unseen indicators. Moreover, we show that we can achieve this goal with only two features of the input without a significant drop in performance. Finally, we demonstrate the efficiency of our architecture to enhance the resolution of the phase diagram of the test data set. We achieve this because our trained network is capable of predicting the indicator values for various choices of the energy density ϵ and disorder parameter h at once.

We emphasize that this NN-based approach to the phenomenon of MBL differs from previous attempts drastically. Previously, NNs have been used for the classification task of preprocessed inputs [42–45]. Such an ansatz depends completely on the availability of the preprocessed input. We take a step further and demonstrate that distinctive signatures of MBL, encoded in the indicator values, are directly learnable from a given disorder

realization in a spin chain. That is, we only enter the defining values of the Hamiltonian and regard the processed indicators as targets, not as inputs to our NN. We obtain these estimates for each disorder realization and for various energy densities at once, i.e. we do not require any averages beforehand.

C. Outline

In the next Section II A, we introduce artificial NNs and in particular our model architecture that is based on a recurrent variant. We proceed by introducing the quantum many-body system of interest for the study of MBL in Section II B. As a test bed for our set-up, this will be the disordered Heisenberg spin chain. To this end, we present prominent indicators of MBL and their behavior in each of the two phases. In Section III A, we demonstrate the scalability of our architecture to predict data for system sizes beyond the training set. This includes a quantitative benchmark of the quality of the network's output. As the next step, we emphasize in Section III B by the means of transfer learning that the relevant global features of the input are recognized. Moreover, this hints towards a compatibility between the various indicators which is understood in the study of Anderson localization but remains unclear for MBL. Lastly, we show the numerical efficiency of our method in Section III C to obtain a high-resolution phase diagram of the MBL-transition. We complement our work with a summary and an outlook for future directions in Section IV.

II. PRELIMINARIES

In the following, we start with providing the required background of RNNs, accompanied by a physical model featuring MBL, the Heisenberg spin chain.

A. Recurrent artificial neural networks

We use artificial NNs and in particular their recurrent variant (RNN). NNs are loosely inspired by their biological counterpart in the human brain. Effectively, they serve as a black-box approach to a universal function approximator. They are modularily built by so-called parameterized layers, usually of the form $\mathbf{y}_l = \sigma(W_l \mathbf{y}_{l-1} + b_l)$ where the parameters of the l-th layer (W_l, b_l) are called weights and biases, respectively. The linearity is broken by a so-called activation function σ which is a non-linear function, usually applied element-wise to its argument. This way, a predefined type of input $\mathbf{x} =: \mathbf{y}_0$ is processed layer by layer. This is referred to as the feed-forward pass of the NN. As a consequence, we can consider the NN as a parameterized black-box function $f_\theta(\mathbf{x}) = \hat{\mathbf{y}}$ with parameters θ given by the weights and biases. In the *supervised learning* setting, the input $\mathbf{x}$ is tied to a target value $\mathbf{y}$ of which $\hat{\mathbf{y}}$ is an estimation. The quality of the estimation is quantifiable by the so-called *loss function*. Its gradient with respect to the network's parameters θ can be computed efficiently by the method of *backpropagation*. It is used in an update rule, such as gradient descent, for the parameters to iteratively find a set of parameters that minimizes the loss [47].

The key limitation of the plain-vanilla NN is the restriction in the fixed input shape. RNNs have a special architecture that allows e.g. for an arbitrary input and output length. This feature is heavily utilized in the field of natural language processing. The recurrent behavior of a layer is achieved by the introduction of a *hidden state* $\mathcal{H}$. To this end, we regard the input $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ as a sequence of T individual inputs. The hidden state can be repeatedly updated according to the network's parameters θ and the current input, i.e. $\mathcal{H}_t = \mathcal{H}_t(\theta, \mathcal{H}_{t-1})$ with $t = 1, \ldots, T$. Importantly, the same parameters θ are used for every update of the hidden state. The final hidden state $\mathcal{H}_T$ serves as the output of the recurrent layer. A schematic is shown in Fig. 2.

B. The model for MBL

A common model often consulted on for the study of MBL is the one-dimensional Heisenberg spin chain of length L whose Hamiltonian reads as

$$H = J \sum_{i=1}^{L} \sum_{\alpha \in \{x,y,z\}} \sigma_\alpha^{(i)} \sigma_\alpha^{(i+1)} + \sum_{i=1}^{L} h_i \sigma_z^{(i)}, \qquad (1)$$

where $\sigma_{x/y/z}^{(i)}$ denotes the respective Pauli matrix acting on the i-th site. We work with periodic boundary conditions, i.e. $\sigma_{x/y/z}^{(L+1)} \equiv \sigma_{x/y/z}^{(1)}$. The *parameters* $\mathbf{h} = (h_1, \ldots, h_L)$ are the local disorder strengths which are sampled independently from a uniform distribution over the interval

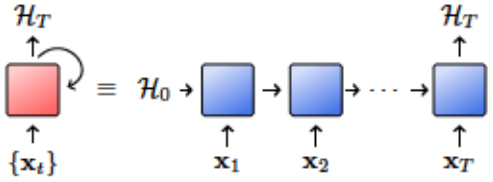

FIG. 2. Scheme of an RNN cell as used in Fig. 1. On the left, the cell is shown as a black-box that iterates over an input sequence $\{\mathbf{x}_t\}$ and produces an output state $\mathcal{H}_T$. Unfolding the cell results in the scheme on the right. An initial hidden state $\mathcal{H}_0$ is evolved over T time steps during which the sequence elements are fed into the network one after another. The final evolved hidden state is released as the network's output. Each box on the right corresponds to the same cell architecture, i.e. having the same weights and biases for each time step. The recurrent cell can process inputs of arbitrary sequence lengths T.

$h_i \in [-h, h]$ for each site i. The variable h is called the *disorder parameter*. The nearest-neighbor interaction strength J can be set to unity as we are only considering its relation to the value of h, i.e. we report values for h in units of J.

We note that the total magnetization $S_z^{\text{tot}} := \sum_{i=1}^{L} \sigma_z^{(i)}$ commutes with the Hamiltonian (1), and we restrict our considerations to the $S_z^{\text{tot}} = 0$ sector and even chain lengths $L \in 2\mathbb{N}$. The dimensionality of this sector is $\binom{L}{L/2}$. This model displays delocalized eigenstates for $h \to 0$ because the Hamiltonian becomes rotationally invariant in this limit. On the other hand, i.e. for $h \to \infty$ the interaction term is negligible, and we recover the localization behavior of the Anderson model. In between these limits, a phase transition from the delocalized phase to the many-body localized one is therefore assumed. Numerical studies report an estimation of the critical disorder parameter h_c of $h_c \approx 6$[1], which has an additional slight dependence on the considered energy density $\epsilon(E) := (E - E_{\min})/(E_{\max} - E_{\min})$ [15]. This numerically observed so-called *mobility edge* is debated from theoretical grounds and attributed to finite-size effects [24].

There are several properties of the two phases which are shared with the Anderson metal-insulator transition. Such properties like the system's entanglement or its spectral statistics are typically aimed to be summarized by a single real number. Since it varies in its numerical value from one phase to the other, it is referred to as an *indicator* for many-body localization. This is not an order parameter as there exists no mean-field theory for MBL [18]. Indicators can be divided into three groups of origin: (i) spectral indicators (function of the eigenvalues), (ii) functions of the eigenvectors (e.g. entanglement entropies), and (iii) time-averaged observables after a quench. As one example for a spectral indicator, it is known that the distribution of the spectral gaps of the Hamiltonian varies between the two phases. In particular, for $h \to 0$ the gaps are distributed according to the Wigner-Dyson distribution whereas the distribution is Poissonian in the MBL phase [13]. These two limiting cases are incorporated by the *adjacent gap ratio* $\langle r \rangle$. This ratio can be computed for the i-th spectral gap $\delta_i = E_{i+1} - E_i \geq 0$ as

$$r_i := \frac{\min\{\delta_{i+1}, \delta_i\}}{\max\{\delta_{i+1}, \delta_i\}} \,. \tag{2}$$

Averaging over all eigenvalues close to a target energy density and over different disorder realizations yields $\langle r \rangle_{\text{deloc}} \approx 0.53$ in the delocalized limit and $\langle r \rangle_{\text{MBL}} = 2\ln(2) - 1 \approx 0.39$ in the MBL phase when $h \to \infty$.

Localization is not only traceable by spectral statistics. Another prominent measure is the *half-chain entanglement entropy* [14]. To this end, we split the chain in half and calculate the reduced density matrix of the first half $\rho_A := \text{Tr}_B[\rho_{AB}]$ by tracing out the second half of the joint density matrix ρ_{AB}. The density operator is constructed for each eigenstate $|n\rangle$ of the Hamiltonian, i.e. $\rho_{AB} = |n\rangle\langle n|$. The entanglement entropy $\langle S_A \rangle$ is given by computing

$$S_A := \text{Tr}[\rho_A \ln(\rho_A)] \tag{3}$$

and averaging again over eigenstates and disorder realizations. We normalize this quantity with the expected maximal half-chain entropy which is the Page entropy [49]. In this way, the indicator varies from 1 in the delocalized regime to approaching 0 in the MBL phase as entanglement is suppressed by the local disorder. Moreover, we note a volume-law scaling of the entanglement entropy with respect to the system size in the delocalized phase but only an area-law scaling in the localized regime [50].

In addition, the eigenstates carry information about the transport behavior of the spin which is a global conserved quantity. The *dynamical spin fraction* $\langle \mathcal{F} \rangle$ quantifies the degree of relaxation of an initial inhomogeneous spin density [14]. It is given as

$$\begin{aligned} \mathcal{F} &:= 1 - \frac{\langle M^\dagger M \rangle}{\langle M^\dagger \rangle \langle M \rangle} \\ \text{with } M &= \sum_{j=1}^{L} \sigma_z^{(j)} \exp\left(2\pi \mathrm{i}\, \frac{j-1}{L}\right) \end{aligned} \tag{4}$$

where the expectation value is taken for all eigenstates close to a target energy. Again, we average $\mathcal{F}$ over many disorder realizations. The persistent spin inhomogeneity in the MBL phase means that $\langle \mathcal{F} \rangle \to 0$ whereas in the delocalized regime $\langle \mathcal{F} \rangle \to 1$.

III. RESULTS

In this work, we report on a highly flexible deep learning architecture whose workflow we depict in Fig. 1 that learns the quantum data obtained from an experiment or a numerical study. In this way, predictions can be made for single instances at various energy levels at once, and we do not need any averages over input configurations. Moreover, the set-up lifts the restriction of a fixed system size for the available quantum data and only requires the relevant parameters of the underlying Hamiltonian. We demonstrate that the set-up extracts global, i.e. task-independent features from the input which makes it applicable to predicting a broad class of quantum data. Thus, our approximation scheme serves as a computationally cheap alternative to demanding numerical methods such as exact diagonalization. We emphasize that, in a broader sense, our method is not limited to the study of MBL but applicable to many more problems in quantum many-body physics.

[1] Due to our definition of Eq. (1) via Pauli matrices, the critical value is twice as large as typically reported in the literature.

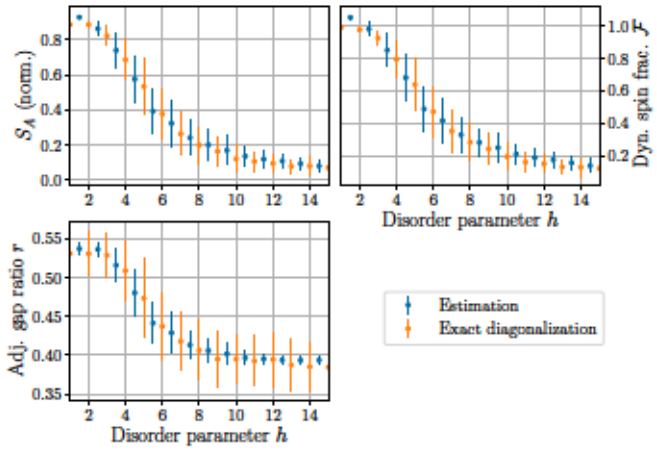

FIG. 3. Estimation of the indicator statistics by the RNN as a function of the disorder parameter h for the $L = 14$ chain at an energy density $\epsilon = 0.5$. We also provide the respective standard deviations around the means which are reproduced by the NN for the first two indicators as well.

A. Scalable indicator approximation

Over the last two decades of approaching Anderson localization analytically and subsequently MBL mostly numerically, several properties of the phenomenon have been demonstrated to be summarized by the aid of the aforementioned indicators. We demonstrate that they can be approximated efficiently by an NN. Intuitively, this comes as no surprise for the indicator values are functions of the Hamiltonian's parameters which are taken as the input of the NN. The defining parameters of the Hamiltonian (1) are the local disorder values $\mathbf{h} = (h_1, \ldots, h_L)$ as we consider isotropic nearest-neighbor interactions of relative unit strength. As we explain later in Section III C, our architecture is capable of estimating the indicators for various values of the energy density ϵ at once. For now, however, we restrict ourselves to the infinite temperature regime, i.e. with $\epsilon = 0.5$ fixed. In order to accommodate disorder vectors of different lengths, we use an RNN architecture that treats the disorder vector as a sequence of the local disorder values. RNNs have specifically been designed to handle variable sequence lengths by virtue of their recursive design, see Fig. 2 and further details in Appendix A. As loss function, we choose the mean-squared-error (MSE) between the obtained estimations of the RNN and the actual values obtained by exact diagonalization of the Hamiltonian. As a framework for setting up the NNs and its training, we rely on PyTorch [51]. We publish our data and the code for performing the training of the NNs and for creating all here presented plots online [52].

Figure 3 shows a plot of the learned indicator statistics for $L = 14$ where the network has been trained on data from chain lengths $L = 10, 12$. We interleave the plotting of the underlying target data with the corresponding output from the NN. For various values of the disorder parameter h, we sampled disorder vectors that make up different Hamiltonians. For each of these, we obtained the vector of indicator values $\mathbf{y}$ from Section II B via exact diagonalization. Each of the disorder vectors was fed into our NN to output an estimation $\hat{\mathbf{y}}$ of $\mathbf{y}$. In the plot, we show the mean and the standard deviation (that results from different realizations of the disorder vector sampled with the same disorder parameter h) of $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively. Especially the entanglement entropy S_A (3) and the dynamical spin fraction $\mathcal{F}$ (4) show a good agreement up to the second moment of the data distribution. For the adjacent gap ratio r (2), only the mean is well-approximated which indicates that the dependence of r on the level of the particular disorder realization may be harder to learn. Importantly, we demonstrate that our NN-architecture can be queried on data belonging to an arbitrary chain length L. Here, we have trained on smaller system sizes and find a qualitative agreement for the larger system size, $L = 14$, in the plot.

Additionally, we can quantitatively benchmark the performance of our network using the *coefficient of determination* R^2. It is used as a benchmarking tool in linear regression and is defined as

$$R^2 := 1 - \frac{\sum_i (f(x_i) - y_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\mathrm{MSE}[f(X), Y]}{\mathrm{Var}[Y]} \quad (5)$$

where the sum runs over all data point pairs $\{(x_i, y_i)\}$ in the test set, the mean over the targets y_i is denoted by $\bar{y}$, f represents the NN and $\mathrm{Var}[Y]$ denotes the variance of Y. So, it essentially compares the MSE of the network outputs with the variance in the data. For a non-linear function f the second term on the right-hand-side is unbounded from above and the corresponding R^2 value will lie in the interval $(-\infty, 1]$ which is unwanted for a squared expression. The coefficient of determination (5) can be transformed to a non-negative number by introducing $R^2_{\mathrm{norm.}} := 1/(2 - R^2) \in [0, 1]$ [53]. Here, $R^2_{\mathrm{norm.}} = 1$ means an approximation being exact and $1/2$ constitutes a baseline value, which is attained for f being the constant function that outputs the target mean. We calculate the normalized coefficient indicator-wise for each value of the disorder parameter h.

The result for the same energy density as in Fig. 3 is presented in Fig. 4. We emphasize that the network has not encountered any training data from the largest system size, $L = 14$. Yet, it is qualitatively able to estimate values beyond its training set system sizes. This quantitative observation corroborates our first qualitative one in Fig. 3. Since the entanglement entropy and the dynamical spin fraction have been well-matched, we see a large value of $R^2_{\mathrm{norm.}}$ for values $h \gtrapprox 3$ accordingly. The breakdown for disorder parameter values below that can be attributed to the vanishing variance in the test set for $h \to 0$ due to the vanishing disorder in the Hamiltonian. As a consequence, it does not pose a threat to our set-up as it could easily be circumvented by weighting the corresponding training

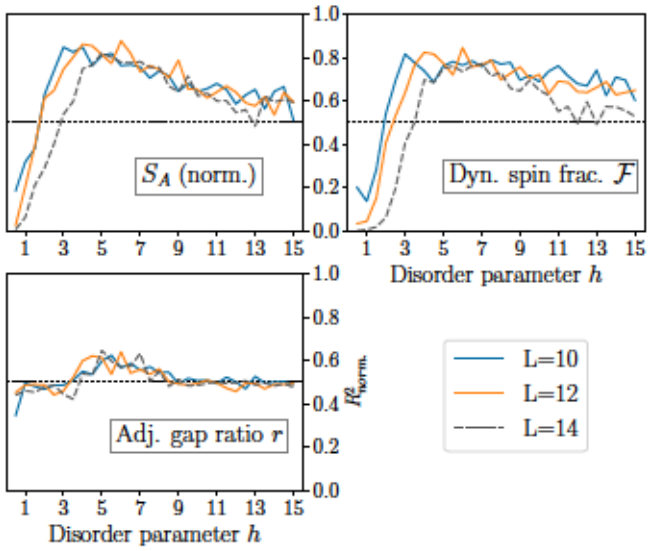

FIG. 4. Normalized coefficient of determination $R^2_{\text{norm.}}$ for each MBL indicator as a function of the disorder parameter h at an energy density $\epsilon = 0.5$. We average all results over five independently trained models. The network has not encountered any data from the $L = 14$ chain (dashed line), yet is capable of capturing the significant part of the indicator statistics. The dotted line is plotted at $R^2_{\text{norm.}} = 1/2$ to serve as a baseline. The breakdown of the quality for small disorder parameter values is due to a vanishing variance in the test set which is a consequence of the vanishing disorder in the system, see main text.

data accordingly. As we have seen already, the adjacent gap ratio can only estimate the mean of the data distribution faithfully. Hence, the corresponding normalized coefficient of determination barely exceeds the baseline value. We attribute this to the unsteadiness in the definition of the adjacent gap ratio caused by the division. Here, similar Hamiltonians in terms of their respective disorder vectors $\mathbf{h}$ can have very different spectra and, in consequence, a very different spectral indicator value. Moreover, it differs in the limit of vanishing disorder as the spectral indicator can be sufficiently described by the Wigner-Dyson distribution from random matrix theory. We therefore do not observe a vanishing variance in our numerics which explains the difference in the limit $h \to 0$ compared to the other indicators.

Lastly, we experimented with the number of required number of samples in the training set. This is a crucial figure of merit since obtaining the training data always poses a bottleneck in deep-learning approaches to quantum many-body physics. Since each disorder realization of a given disorder parameter value h is sampled from a uniform distribution over the interval $[-h, h]$, the corresponding variance for a single local disorder strength h_i increases quadratically with h. However, we found no qualitative difference in the approximation quality when considering a training set with a massively increased proportion of data from the MBL side. As the bottleneck of benchmarking our approach is the generation of the training set (due to the cost intensity of the exact diagonalization), we are interested in how the network copes with a shrunken training data set. We refer to Appendix B for the analysis and plots. In essence, we find that we can shrink the training data set if we allow for more training epochs in return. This way, we can reduce the training data set down to a number close to the number of trainable parameters in the network. These observations are crucial for obtaining a data set from an actual experiment in the future where determining indicator values for even a single realization might be expensive.

B. Transfer learning

The common notion in deep learning is that there exists a hierarchy of abstraction in what the different layers of an NN are capable of identifying. This view has been corroborated by inspecting the first layers of state-of-the-art image classifiers which correspond to edge and corner detection [47]. Since such tasks are detached from the actual classification task, the first layers are said to detect task-unspecific, general *features* of the input and thus regarded as feature extractors. Only the last layers of a (deep) NN map these extracted features to the specific problem at hand.

In this section, we inspect whether such a behavior is exhibited by our proposed model. We approach this question with the aid of transfer learning [48]. The idea is, assuming that the RNN actually extracts general features of the disorder vector $\mathbf{h}$, to keep the RNN fixed after we have trained it on a set of MBL indicators. We can now switch the targets in the training set, i.e. exchange the target indicators with some new indicators which the network has not encountered before. As the RNN-output is detached from the choice of the target indicators, we only retrain the NN that maps the features to the newly chosen indicators. If the output of the RNN corresponds to features of the input that are task-independent, the prediction quality should be comparable to the case where we retrain the full model from scratch on the new data.

We select the dynamical spin fraction $\mathcal{F}$ (4) as the transfer target indicator. To this end, we train our model on the adjacent gap ratio r (2) and on the entanglement entropy S_A (3) for system sizes $L = 10, 12$. Thus, we exclude $\mathcal{F}$ explicitly from the training set. Once the training succeeds, we keep the RNN's parameters fixed and only retrain the subsequent NN to predict the spin fraction given the output of the RNN. We benchmark the prediction quality with a model of the same architecture that is trained to predict only $\mathcal{F}$ from scratch. Furthermore, we compare both predictions with the previous model from Fig. 4 that has been trained on all three indicators at once and which we call the multitask network. A quantitative comparison using the normalized coefficient of determination (5) is given in Fig. 5. The transferred features lead

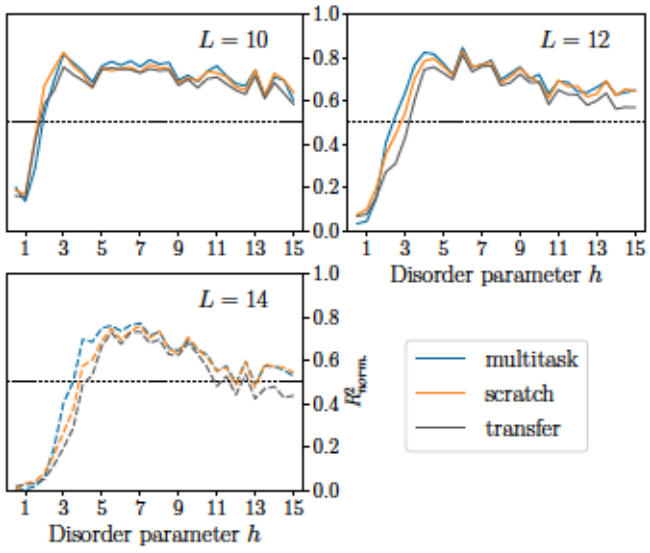

FIG. 5. Plot of the normalized coefficient of determination (5) for the dynamical spin fraction $\mathcal{F}$. We compare the model trained via transfer learning (gray line) with an uninitialized model that learns from scratch (orange line) and the previously trained model on all indicators at once (blue line). Once again, we have excluded data for $L = 14$ from the training set which is indicated by the dashed lines in the lower left panel. The dotted line is plotted at $R^2_{\text{norm.}} = 1/2$ to serve as a baseline. We averaged the outcome over five independent training procedures.

to a comparable performance as a model that is retrained from scratch and thus tailored to the specific indicator. Additionally, the performance of the these two networks is very similar to the multitask network. The differences between any two curves is due to statistical errors. We find a similar situation when selecting the adjacent gap ratio or the entanglement entropy as the transfer target indicator, respectively (data not shown). We can attribute the congruence of all three different types of training to the following two reasons. First, there appears no qualitative difference in the learnability of each of the indicators. Moreover, they seem to be compatible with each other in the sense that they can all be obtained from the same features. In our case, we are able to apply the transfer learning scheme using only two features. We provide more details in Appendix A. This indicates that the extracted features are general enough to allow for the estimation of a variety of indicators which, in turn, do not rely on a specific set of features produced during a specific training procedure.

C. Energy dependency

Lastly, we demonstrate that predictions from our trained estimator recover the results from previous numerical studies of MBL in the limit of averaging over many disorder realizations. Namely, we recover the phase diagram of the transition for various chain lengths L that show the indicator values in dependence of the considered disorder parameter h and energy density ϵ. To this end, we can generate predictions of unseen trial disorder realizations, i.e. random instances of disorder vectors for a given chain length and disorder parameter. These instances are fed into our NN to accumulate a trial data set for various energy densities ϵ at once. The latter is straight-forwardly incorporated by augmenting the output of the RNN by the corresponding value for ϵ. Since we solely focus on the network's prediction, we do not need to perform the exact diagonalization procedure for these new instances. Therefore, generating this large data set is efficient in the system size. The resulting phase diagram for the dynamical spin fraction $\mathcal{F}$ is presented in Fig. 6.

Most importantly, we are now able to generate images of the phase diagram to an arbitrary resolution with numerical efficiacy. Moreover, we are not limited by the initial resolution in the training data. This is because we only require forward passes through the NN which scales both linearly in the number of queried values for both the disorder parameter and the energy density. We provide further insights in Appendix C.

IV. CONCLUSION AND OUTLOOK

We have constructed a RNN architecture that approximates values for certain indicators for MBL directly from the variable part of the Hamiltonian, i.e. the local disorder strengths. The recurrent set-up ensures that the network can process data for an arbitrary system size L and produce a good estimation output provided the trial system size is not too far off the training set. Moreover, our approach does not require any further computationally expensive preprocessing of the input data. In this way, we are able to characterize single disorder realizations by pro-

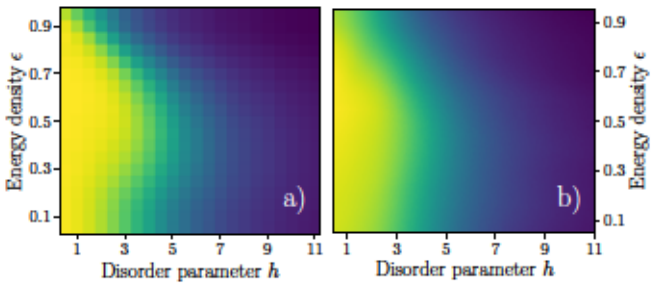

FIG. 6. Phase diagram of the dynamical spin fraction $\mathcal{F}$ for various energy densities ϵ, disorder parameter values h and for a chain length of $L = 14$. In (a), we show the qualitative diagram of the transition obtained by averaging over many disorder realizations. It is faithfully reproduced by the averaged predictions of the NN (b). Moreover, as the NN allows to estimate data for arbitrary values of ϵ and h we can efficiently increase the resolution of the diagram.

viding the corresponding indicator values. By inspecting the intermediate features of the RNN by means of transfer learning, we observe that all considered indicators can be derived from two features alone. Furthermore, they serve as an archetype for various indicators at arbitrary energy densities at once. This enables us to study the transition region by means of phase diagrams that can be rendered to an arbitrary resolution.

Outlook

With a training set that consists of indicator sets from different system sizes, we envision an interplay between an actual experiment and our architecture. The experiment can address systems consisting of dozens of spins or qubits. Thus, it delivers the training set for the architecture beyond what is reachable by exact diagonalization studies. As we demonstrated, our architecture is not inclined to a specific data type. Thus, the experiment is not restricted to a certain indicator but can provide the most amenable one (such as the growth of the entanglement entropy [54], the imbalance after a quench [54–57] or even characteristics of the energy spectrum [58]) for the training set. Motivated by our findings in Section III A, we conjecture that only a few realizations per disorder parameter are sufficient as to merely guide the extrapolation. In addition, the indicators are expected to become more and more pronounced in their respective shape. Therefore, we do not expect large deviations from the case of smaller system sizes up to finite-size effects. The whole premise of transfer learning relies on the assumption that the additional data for a larger system size only serves as a guidance for the overall learned structure on the training set. This boosts training the NN significantly [48, 59]. Given experimental training input, the network can in turn provide estimates for data outside of or in between gaps in the training set which can be benchmarked by the experiment in return [60, 61]. Other possibilities of enriching the training set is to resort to numerical approximations, for example by tensor networks methods which are well-suited deep within the MBL phase [62] or yet another NN architecture to even speed up those methods [63]. With the data at hand, a more detailed examination of the compability of different indicators allows to shed some light on their yet unknown coaction towards MBL. Diving deeper into the interpretation of the archetypical feature and the compatibility of various indicators is an interesting research direction for future works.

Our proposed scheme aims to bring together the often independent advances in experiments and numerics, and we see possible research directions in the now scalable phase classification task and a better understanding of the learning process of the recurrent feature extractor. Furthermore, the connection of our method with a VQA is of broader interest ranging from applications in condensed matter and statistical physics to the field of (hybrid) quantum computation or quantum machine learning. Compared to the existing traditional numerical methods, the interplay of a quantum experiment or its simulation with our method may constitute a new type of quantum advantage in the sense that we can obtain an efficient classical method only via accessing a quantum data set. Such a pairing provides a potentially powerful computational tool that is yet to be augmented with experimental data in the future.

APPENDIX

In this appendix, we provide more details on our network architecture and the training procedure. Starting with Appendix A, we describe the generation of the data sets and detail the architecture of our approach. In Appendix B, we examine the network's performance under a shrinking data set size. Finally, we give some more comments on the obtained phase diagram in Section III C and its analysis in Appendix C.

A. Details on the network architecture and the training procedure

We briefly describe how we set up the training and the test set as well as the network architecture used for the results in the main text. We set up a grid for the disorder parameter h, i.e. we chose 30 values $h = 0.5, 1, 1.5, \ldots, 15$ which lie well around the assumed critical disorder parameter value of $h_c \approx 6$. For each chain length $L = 10, 12, 14$ we have sampled disorder vectors $\mathbf{h}$ with entries h_i independently and identically distributed from the uniform distribution, such that $h_i \in [-h, h]$ for a given disorder parameter value h. For each h and L, this was done $N_{\text{train}} = 1000$ and $N_{\text{test}} = 100$ times for the two data sets, respectively. Each of these disorder vectors yields a realization of the Hamiltonian (1). Its eigenvalues and -vectors were found via exact diagonalization. We have chosen a grid of $N_\epsilon = 19$ energy densities $\epsilon = 0.05, 0.1, 0.15, \ldots, 0.95$ and have kept the 100 next closest eigenvalues and their corresponding eigenvectors for calculating the three indicators from Section II B.

The architecture of our proposed network scheme is summarized in Fig. 7 and we explain its choice in the following. The first part consists of an RNN-cell that serves as a feature extractor of the input. The RNN is presented each disorder parameter h_i successively and updates its hidden state according to its parameters and the value of h_i. The hidden state was initialized as zero. After having fed in h_L, the final updated hidden state is released as the output of the RNN. We treat this output as the feature vector of the disorder vector. Due to this recursive procedure, RNNs can be unstable during training because of exploding or vanishing gradients in the optimization procedure. In order to circumvent this problem, the long short-term memory (LSTM) cell [64] and the gated recurrent unit (GRU) [65] have been proposed with competing performance-efficiency trade-offs [66]. We find the latter to be slightly better in performance during training. Concerning the number of output features of the RNN, we find qualitative good results when choosing a feature dimension of 2. A larger dimensionality does increase the performance of the indicator approximation, however, we observe a severly decreased performance when applying the transfer learning scheme from Fig. 5. We have only used a single RNN cell of depth one. Lastly, we have performed a computationally inexpensive preprocessing of the disorder vector. We regroup the elements of the disorder vector in pairs of two, i.e. transform according to $[h_1, h_2, h_3, \ldots, h_L] \mapsto [(h_1, h_2), (h_2, h_3), \ldots, (h_{L-1}, h_L), (h_L, h_1)]$. Regroupings into even larger tuples are also possible. The pairing in two, however, fits in well with the nearest-neighbor interactions and the periodic boundary condition and, furthermore, leads to the best performance. Afterwards, the feature vector is augmented by the value for the energy density ϵ under consideration. Together, we map them to the three indicator values by a fully-connected NN of hidden size 10. As the loss we choose the mean-squared-error (MSE) and train the model for $N_{\text{epochs}} = 15$ on the training data. We use the Adam optimizer with default values [67], a batch-size of 128 and a learning rate $\eta = 10^{-3}$.

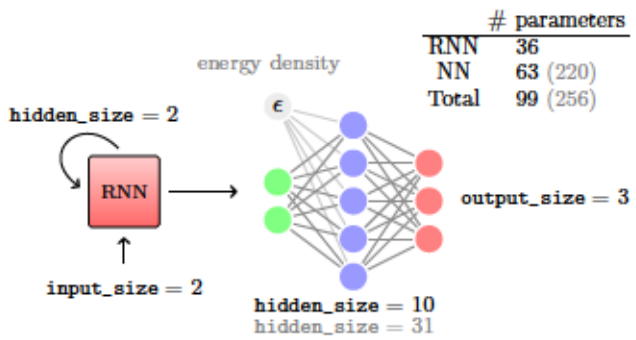

FIG. 7. Details of our model architecture of Fig. 1. The recurrent neural network as in Fig. 2 takes in the preprocessed input iteratively. Afterwards, the final hidden state is fed into the fully-connected NN. It can be augmented by the respective energy density ϵ as done for Section III C. The corresponding alterations in the network architecture are emphasized by the gray font. The total number of trainable parameters (including biases) are given in the table.

For the transfer learning scheme of Section III B and for creating the model that is capable of dealing with an arbitrary energy density ϵ in Section III C, the training consists of two stages: we first proceed as outlined above This pretraining is necessary to facilitate an easier focussed training of the RNN to extract meaningful features which we show in Fig. 8. Then, we fix the parameters of the RNN and thus the intermediate features, and train the subsequent fully-connected NN on the full training data for 30 more epochs with a decreased learning rate of 10^{-4} following the Adam optimizer routine. This fine-tuning of the NN yields a greater performance compared to training the two components of the model jointly. The choice for the hyperparameters (architecture of the two individual components, feature size, number of hidden neurons and the optimizer parameters) above has been determined on a held-out validation data set.

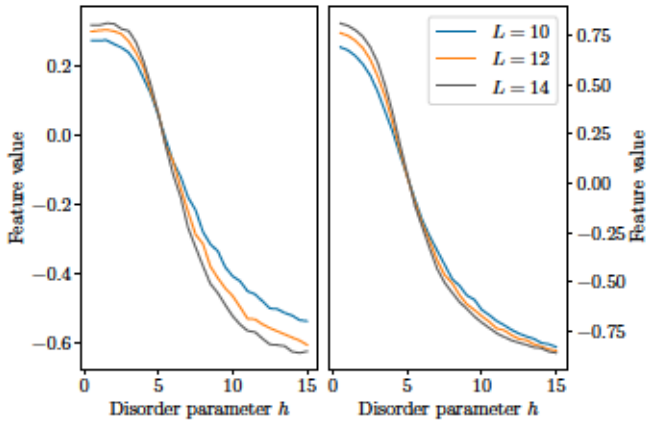

FIG. 8. Typical features produced after the training averaged over the training set. Error on the means are on the scale of the line thickness. The crossing points vary from training to training which makes retraining and averaging a necessity.

B. Examination of the data set size

In this section, we provide details on the results of Section III A. In particular, we investigate the performance dependence on the size of the training data set. We can test this quantitatively by decreasing the number of samples per disorder parameter N_{train}. In this setting, half a value in N_{train} corresponds to a two-fold reduction in the training set size. If we were to train now for a fixed number of epochs N_{epochs}, that is, until the network encountered each data point N_{epochs} times during training, we expect a better performance with a larger N_{train}. In this case, the network receives more update iterations to minimize the MSE objective, hence the performance gain. For a fairer comparison, we track both the training and the test loss during training after each update step. Hence, the total number of iteration steps is to be made a constant, i.e. on a training set of twice the size we allow the network to train for half the epochs. In this setting, each training run allows the NN the same total amount of update steps.

In particular, this has resulted in very long training loops for a small N_{train} as we have trained for several hundreds of epochs. Due to the mini-batching during training, we track the actual number of received update steps during training for various values of N_{train} and exclude the system size of $L = 14$ from the training set. We set a value of $N_{\text{epochs}} = 30$ for training on the largest data set size with $N_{\text{train}} = 100$ and adjusted that value accordingly for smaller sizes. In all considered cases, this leads to a convergence of the models and we extract the remaining average MSE on both the training and the test set after convergence. For each value of N_{train}, we reinitialize and train the model ten times. In all cases, when we decreased the training set, we have done so by always picking a random subset of the full training data set for each training reinitialization. We show the two averaged losses in Fig. 9. This reveals that shrinking the training set down to $N_{\text{train}} \approx 3$ (this corresponds to a total number of training points of around 180) yields no qualitative increase of neither of the two losses after training. This threshold is of the order or trainable parameters of the model (cf. Fig. 7). Below it, we observe a decreased training loss while the test loss is increased. In this limit of scarce data, the model begins to overfit the training data at the expense of a larger loss on the test set. This small number is encouraging for the model application to data that stems from an actual experiment as we have to repeat the same experiment only a handful of times for each point in the phase diagram we are interested in. This highlights the feasability of our approach to actual data stemming from a quantum experiment.

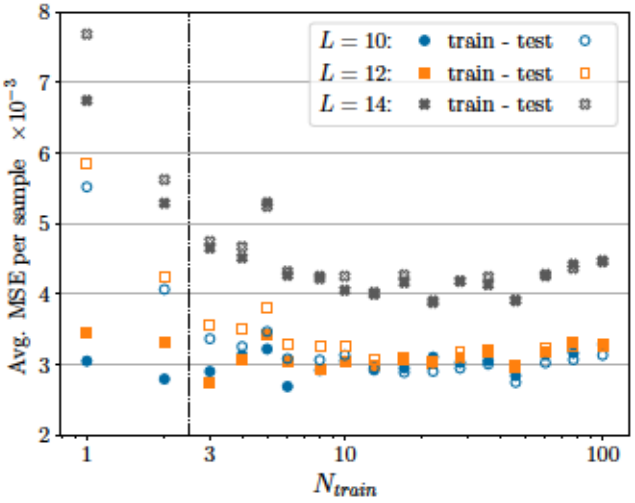

FIG. 9. Dependence of the training and the test loss on the size of the corresponding training set. N_{train} denotes how many realizations for each disorder parameter h and chain length L have been included in the training set. Both losses are reported after convergence (around 1.350 update steps). We distinguish losses for different system size by color and the train from the test loss by different symbols, respectively. We have averaged over ten independent training procedures. There is no qualitative improvement for a training set with $N_{\text{train}} \geq 3$ (vertical, dotted line). Below this threshold, the network tends to overfit the available data, indicated by an increasing test error despite a decreased train error. We have excluded data for $L = 14$ from the training set, hence the increased losses for this system size.

C. Further details on the phase diagrams

In Section III C, we highlight that our model is capable of dealing with various values for the energy density ϵ. Due to the choice of our architecture, ϵ is taken as an input feature for the subsequent fully-connected NN. We have experimented with various ways in presenting different

values for ϵ to our model. One initial alternative consists of various fully-connected NNs that are individually trained to predict the indicator values at a single ϵ each. While this, at first, has appeared beneficial with respect to the validation loss, there are a few drawbacks of this approach. The first one is the increased model complexity opposed to our scheme now. Here, we only require one single NN whereas the naive approach would require an NN for every ϵ of interest. Secondly, this approach limits the resolution of the prediction when in comes to obtaining the phase diagram in Fig. 6 as we require a data set for every ϵ of interest. Our approach circumvents both issues by the introduction of ϵ as an intermediate feature. This way, we can set up a much tighter grid for both ϵ as well as the disorder parameter h and make predictions for each possible combination. To this end, we sample $N = 100$ new samples of disorder vectors $\mathbf{h}$ for each h and obtain the feature value by feeding it to the RNN. Then, we augment this value with every value of ϵ of interest and parse everything to the NN. Lastly, we average over N and show this mean in dependence of ϵ and h in the phase diagram. Since we only require forward passes through our model, this procedure is highly efficient: the run time is proportional to the chain length L and to the number of queried values for both h and ϵ and in that sense optimal.

We have also experimented with analysing the model's predictions with a more quantitative measure such as the finite-size scaling analysis (FSSA) [68]. This method is aimed at mitigating the finite-size effects in the data and to obtain quantitative estimates of the critical disorder parameter h_c and the critical exponent of the transition ν. To this end, data from various chain lenghts is given to the FSSA and fitted around the assumed value for h_c. We have tried to query our model at chain lengths beyond those in the training set, i.e. $L > 14$ but failed to reproduce previous approaches [15] as we have not observed signs of the ϵ-dependent mobility edge in the transition. We attribute this observation to two different origins. First, we observe that the approximation is of higher quality around the transition region (cf. Fig. 4) and significantly so in the middle of the spectrum (at $\epsilon \approx 0.5$). The latter might leave a bias in the data at either side of the spectrum which is observed in the phase diagram. The second reason is due to our choice of the RNN architecture as feature extractor. In Fig. 8, we have shown the typical feature vector produced by the RNN after training. One important aspect is that there exists a cross-over point that is independent of the chain length L of the input data but whose position depends on the initialization of the network parameters. This introduces a bias in the indicators since this cross-over is not apparent in the training data. We have tried to average the output over multiple retrainings (and therefore feature vectors) and by increasing the number of features but failed to lift this bias. However, we conjecture that with a more careful design of the RNN architecture, this is possible. In any case, the investigation of finding the right feature architecture is both interesting from a numerical and a theoretical perspective as it helps to shine some light on the nature of the MBL transition.

ACRONYMS

FSSA	finite-size scaling analysis	11
GRU	gated recurrent unit	9
LSTM	long short-term memory	9
MBL	many-body localization	1
MSE	mean-squared-error	5
NN	neural network	2
RNN	recurrent neural network	2
VQA	variational quantum algorithm	1
VQE	variational quantum eigensolver	1

[1] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, *Variational quantum algorithms*, Nature Reviews Physics **3**, 625 (2021), arXiv:2012.09265.

[2] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, *Noisy intermediate-scale quantum algorithms*, Rev. Mod. Phys. **94**, 015004 (2022), arXiv:2101.08448 [quant-ph].

[3] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, *A variational eigenvalue solver on a photonic quantum processor*, Nat. Commun. **5**, 4213 (2014), arXiv:1304.3061.

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

[5] M. I. Jordan and T. M. Mitchell, *Machine learning: Trends, perspectives, and prospects*, Science **349**, 255 (2015).

[6] H. Y. Huang, R. Kueng, and J. Preskill, *Information-Theoretic Bounds on Quantum Advantage in Machine Learning*, Phys. Rev. Lett. **126**, 190505 (2021), arXiv:2101.02464.

[7] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, *Power of data in quantum machine learning*, Nat. Commun. 2021 12:1 **12**, 1 (2021), arXiv:2011.01938 [quant-ph].

[8] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, *Provably efficient machine learning for quantum many-body problems*, arXiv:2106.12627.

[9] P. J. J. O'Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tran-

ter, N. Ding, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Jeffrey, E. Lucero, A. Megrant, J. Y. Mutus, M. Neeley, C. Neill, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, P. V. Coveney, P. J. Love, H. Neven, A. Aspuru-Guzik, and J. M. Martinis, *Scalable quantum simulation of molecular energies*, Phys. Rev. X **6**, 031007 (2016), arXiv:1512.06860.

[10] K. M. Nakanishi, K. Mitarai, and K. Fujii, *Subspace-search variational quantum eigensolver for excited states*, Phys. Rev. Research **1**, 033062 (2019), arXiv:1810.09434.

[11] O. Higgott, D. Wang, and S. Brierley, *Variational quantum computation of excited states*, Quantum **3**, 156 (2019), arXiv:1805.08138.

[12] S. Liu, S.-X. Zhang, C.-Y. Hsieh, S. Zhang, and H. Yao, *Probing many-body localization by excited-state VQE*, arXiv:2111.13719.

[13] V. Oganesyan and D. A. Huse, *Localization of interacting fermions at high temperature*, Phys. Rev. B **75**, 155111 (2007), arXiv:cond-mat/0610854 [cond-mat.str-el].

[14] A. Pal and D. A. Huse, *Many-body localization phase transition*, Phys. Rev. B **82**, 174411 (2010), arXiv:1010.1992 [cond-mat.dis-nn].

[15] D. J. Luitz, N. Laflorencie, and F. Alet, *Many-body localization edge in the random-field Heisenberg chain*, Phys. Rev. B **91**, 081103 (2015), arXiv:1411.0660 [cond-mat.dis-nn].

[16] R. Nandkishore and D. A. Huse, *Many-body localization and thermalization in quantum statistical mechanics*, Annu. Rev. Condens. Matter Phys. **6**, 15 (2015), arXiv:1404.0686 [cond-mat.stat-mech].

[17] J. Eisert, M. Friesdorf, and C. Gogolin, *Quantum many-body systems out of equilibrium*, Nat. Phys. **11**, 124 (2015), arXiv:1408.5148 [quant-ph].

[18] F. Alet and N. Laflorencie, *Many-body localization: An introduction and selected topics*, Comptes Rendus Physique **19**, 498 (2018), arXiv:1711.03145 [cond-mat.str-el].

[19] P. W. Anderson, *Absence of diffusion in certain random lattices*, Phys. Rev. **109**, 1492 (1958).

[20] D. Basko, I. Aleiner, and B. Altshuler, *Metal–insulator transition in a weakly interacting many-electron system with localized single-particle states*, Annals of Physics **321**, 1126 (2006), arXiv:cond-mat/0506617 [cond-mat.mes-hall].

[21] A. Chandran, I. H. Kim, G. Vidal, and D. A. Abanin, *Constructing local integrals of motion in the many-body localized phase*, Phys. Rev. B **91**, 085425 (2015), arXiv:1407.8480 [cond-mat.dis-nn].

[22] I. H. Kim, A. Chandran, and D. A. Abanin, *Local integrals of motion and the logarithmic lightcone in many-body localized systems*, arXiv:1412.3073 [cond-mat.dis-nn].

[23] L. Rademaker, M. Ortuño, and A. M. Somoza, *Many-body localization from the perspective of integrals of motion*, Ann. Phys. **529**, 1600322 (2017), arXiv:1610.06238 [cond-mat.str-el].

[24] J. Z. Imbrie, V. Ros, and A. Scardicchio, *Local integrals of motion in many-body localized systems*, Ann. Phys. **529**, 1600278 (2017), arXiv:1609.08076 [cond-mat.dis-nn].

[25] D. J. Luitz and Y. B. Lev, *The ergodic side of the many-body localization transition*, Ann. Phys. **529**, 1600350 (2017), arXiv:1610.08993 [cond-mat.dis-nn].

[26] J. M. Deutsch, *Quantum statistical mechanics in a closed system*, Phys. Rev. A **43**, 2046 (1991).

[27] M. Srednicki, *Chaos and quantum thermalization*, Phys. Rev. E **50**, 888 (1994), arXiv:cond-mat/9403051 [cond-mat].

[28] M. Rigol, V. Dunjko, and M. Olshanii, *Thermalization and its mechanism for generic isolated quantum systems*, Nature **452**, 854 (2008), arXiv:0708.1324 [cond-mat.stat-mech].

[29] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, *From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics*, Adv. Phys. **65**, 239 (2016), arXiv:1509.06411.

[30] F. Pietracaprina, N. Macé, D. J. Luitz, and F. Alet, *Shift-invert diagonalization of large many-body localizing spin chains*, SciPost Phys. **5**, 45 (2018), arXiv:1803.05395 [cond-mat.dis-nn].

[31] P. Sierant, M. Lewenstein, and J. Zakrzewski, *Polynomially filtered exact diagonalization approach to many-body localization*, Phys. Rev. Lett. **125**, 156601 (2020), arXiv:2005.09534 [cond-mat.dis-nn].

[32] S. P. Lim and D. N. Sheng, *Many-body localization and transition by density matrix renormalization group and exact diagonalization studies*, Phys. Rev. B **94**, 045111 (2016), arXiv:1510.08145 [cond-mat.str-el].

[33] V. Khemani, S. P. Lim, D. N. Sheng, and D. A. Huse, *Critical properties of the many-body localization transition*, Phys. Rev. X **7**, 021013 (2017), arXiv:1607.05756 [cond-mat.dis-nn].

[34] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification*, arXiv:1502.01852.

[35] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, *Learning phase transitions by confusion*, Nat. Phys. **13**, 435 (2017), arXiv:1610.02048 [cond-mat.dis-nn].

[36] J. Carrasquilla and R. G. Melko, *Machine learning phases of matter*, Nat. Phys. **13**, 431 EP (2017), arXiv:1605.01735 [cond-mat.str-el].

[37] Y.-H. Liu and E. P. L. van Nieuwenburg, *Discriminative cooperative networks for detecting phase transitions*, Phys. Rev. Lett. **120**, 176401 (2018), arXiv:1706.08111 [cond-mat.str-el].

[38] R. G. Melko, G. Carleo, J. Carrasquilla, and J. I. Cirac, *Restricted Boltzmann machines in quantum physics*, Nat. Phys. **15**, 887 (2019).

[39] Y.-T. Hsu, X. Li, D.-L. Deng, and S. Das Sarma, *Machine learning many-body localization: Search for the elusive nonergodic metal*, Phys. Rev. Lett. **121**, 245701 (2018), arXiv:1805.12138.

[40] J. Venderley, V. Khemani, and E.-A. Kim, *Machine learning out-of-equilibrium phases of matter*, Phys. Rev. Lett. **120**, 257204 (2018), arXiv:1711.00020 [cond-mat.dis-nn].

[41] W. Zhang, L. Wang, and Z. Wang, *Interpretable machine learning study of the many-body localization transition in disordered quantum Ising spin chains*, Phys. Rev. B **99**, 054208 (2019), arXiv:1807.02954.

[42] F. Schindler, N. Regnault, and T. Neupert, *Probing many-body localization with neural networks*, Phys. Rev. B **95**, 245134 (2017), arXiv:1704.01578 [cond-mat.dis-nn].

[43] E. van Nieuwenburg, E. Bairey, and G. Refael, *Learning phase transitions from dynamics*, Phys. Rev. B **98**, 060301 (2018), arXiv:1712.00450 [cond-mat.dis-nn].

[44] P. Huembeli, A. Dauphin, P. Wittek, and C. Gogolin, *Automated discovery of characteristic features of phase transitions in many-body localization*, Phys. Rev. B **99**, 104106 (2019), arXiv:1806.00419 [quant-ph].

[45] E. van Nieuwenburg, Y. Baum, and G. Refael, *From*

Bloch oscillations to many-body localization in clean interacting systems, Proc. Nat. Acad. Sci. **116**, 9269 (2019), arXiv:1808.00471 [cond-mat.dis-nn].

[46] N. Saraceni, S. Cantori, and S. Pilati, *Scalable neural networks for the efficient learning of disordered quantum systems*, Phys. Rev. E **102**, 033301 (2020), arXiv:2005.14290.

[47] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016) http://www.deeplearningbook.org.

[48] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, *How transferable are features in deep neural networks?* Adv. Neur. Inf. Processing Sys. **4**, 3320 (2014), arXiv:1411.1792.

[49] D. N. Page, *Average entropy of a subsystem*, Phys. Rev. Lett. **71**, 1291 (1993), arXiv:gr-qc/9305007 [gr-qc].

[50] J. Eisert, M. Cramer, and M. B. Plenio, *Colloquium: Area laws for the entanglement entropy*, Rev. Mod. Phys. **82**, 277 (2010), arXiv:0808.3773 [quant-ph].

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, in *Adv. in Neur. Inf. Processing Sys. 32* (Curran Associates, Inc., 2019) pp. 8024–8035, arXiv:1912.01703 [cs.LG].

[52] *Supplementary code*, (2021).

[53] J. Nossent and W. Bauwens, *Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol' sensitivity analysis of a hydrological model*, in *EGU Gen. Ass. Conf. Abstracts* (2012) p. 237.

[54] K. Xu, J.-J. Chen, Y. Zeng, Y.-R. Zhang, C. Song, W. Liu, Q. Guo, P. Zhang, D. Xu, H. Deng, K. Huang, H. Wang, X. Zhu, D. Zheng, and H. Fan, *Emulating many-body localization with a superconducting quantum processor*, Phys. Rev. Lett. **120**, 050507 (2018), arXiv:1709.07734 [quant-ph].

[55] M. Schreiber, S. S. Hodgman, P. Bordia, H. P. Lüschen, M. H. Fischer, R. Vosk, E. Altman, U. Schneider, and I. Bloch, *Observation of many-body localization of interacting fermions in a quasirandom optical lattice*, Science **349**, 842 (2015), arXiv:1501.05661 [cond-mat.quant-gas].

[56] P. Bordia, H. P. Lüschen, S. S. Hodgman, M. Schreiber, I. Bloch, and U. Schneider, *Coupling identical one-dimensional many-body localized systems*, Phys. Rev. Lett. **116**, 140401 (2016), arXiv:1509.00478 [cond-mat.quant-gas].

[57] T. Kohlert, S. Scherg, X. Li, H. P. Lüschen, S. Das Sarma, I. Bloch, and M. Aidelsburger, *Observation of many-body localization in a one-dimensional system with a single-particle mobility edge*, Phys. Rev. Lett. **122**, 170403 (2019), arXiv:1809.04055 [cond-mat.quant-gas].

[58] L. K. Joshi, A. Elben, A. Vikram, B. Vermersch, V. Galitski, and P. Zoller, *Probing many-body quantum chaos with quantum simulators*, Phys. Rev. X **12**, 011018 (2022), arXiv:2106.15530 [quant-ph].

[59] C. Beetar, J. Murugan, and D. Rosa, *Neural networks as universal probes of many-body localization in quantum graphs*, arXiv:2108.05737 [cond-mat.dis-nn].

[60] N. Mohseni, C. Navarrete-Benlloch, T. Byrnes, and F. Marquardt, *Deep recurrent networks predicting the gap evolution in adiabatic quantum computing*, arXiv:2109.08492 [quant-ph].

[61] C. Miles, R. Samajdar, S. Ebadi, T. T. Wang, H. Pichler, S. Sachdev, M. D. Lukin, M. Greiner, K. Q. Weinberger, and E.-A. Kim, *Machine learning discovery of new phases in programmable quantum simulator snapshots*, arXiv:2112.10789 [quant-ph].

[62] M. Friesdorf, A. H. Werner, W. Brown, V. B. Scholz, and J. Eisert, *Many-body localization implies that eigenvectors are matrix-product states*, Phys. Rev. Lett. **114**, 170505 (2015), arXiv:1409.1252 [quant-ph].

[63] C. Guo, Z. Jie, W. Lu, and D. Poletti, *Matrix product operators for sequence-to-sequence learning*, Phys. Rev. E **98**, 1 (2018), arXiv:1803.10908 [cond-mat.stat-mech].

[64] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neur. Comput. **9**, 1735 (1997).

[65] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, *On the properties of neural machine translation: Encoder–decoder approaches*, in *Proc. of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Stat. Translation* (Association for Computational Linguistics, Doha, Qatar, 2014) pp. 103–111, arXiv:1409.1259 [cs.CL].

[66] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, *Empirical evaluation of gated recurrent neural networks on sequence modeling*, in *NIPS Workshop on Deep Learning* (2014) arXiv:1412.3555 [cs.NE].

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

[68] J. T. Chayes, L. Chayes, D. S. Fisher, and T. Spencer, *Finite-size scaling and correlation lengths for disordered systems*, Phys. Rev. Lett. **57**, 2999 (1986).