The Optimal Depth of Variational Quantum Algorithms Is QCMA-Hard to Approximate

Lennart Bittel
Institute for Theoretical Physics, Heinrich Heine University Düsseldorf, Germany

Sevag Gharibian
Department of Computer Science, and Institute for Photonic Quantum Systems, Universität Paderborn, Germany

Martin Kliesch
Institute for Theoretical Physics, Heinrich Heine University Düsseldorf, Germany
Institute for Quantum-Inspired and Quantum Optimization, Technische Universtiät Hamburg, Germany

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.

2012 ACM Subject Classification Theory of computation → Quantum complexity theory

Keywords and phrases Variational quantum algorithms (VQA), Quantum Approximate Optimization Algorithm (QAOA), circuit depth minimization, Quantum-Classical Merlin-Arthur (QCMA), hardness of approximation, hybrid quantum algorithms

Digital Object Identifier 10.4230/LIPIcs.CCC.2023.34

Related Version *Full Version*: `https://arxiv.org/abs/2211.12519`

Funding *Lennart Bittel and Martin Kliesch*: German Federal Ministry of Education and Research (BMBF), funding program "Quantum Technologies – from Basic Research to Market" via the joint project MANIQU (grant number 13N15578); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the grant number 441423094 within the Emmy Noether Program.
Sevag Gharibian: 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); project "PhoQC" from the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia.

Acknowledgements We thank Ashley Montanaro for helpful discussions.

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. [13, 8] for reviews), with Farhi, Goldstone and Gutmann's Quantum Approximate Optimization Algorithm (QAOA) being a prominent example [14].

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 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 [14] for MAX-CUT being a well-known example. Thus, estimating the optimal depth for a 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 $S_i \subseteq [n]$ 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 $(\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$ and a list $(G_1, \ldots, G_m)$ of Hamiltonians from H (repetitions permitted) such that $|\psi\rangle := e^{i\theta_m G_m} \cdots e^{i\theta_1 G_1}|0\cdots 0\rangle$ 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), $|\psi\rangle := e^{i\theta_{m'} G_{m'}} \cdots e^{i\theta_1 G_1}|0\cdots 0\rangle$ satisfies $\langle\psi|M|\psi\rangle \geq 2/3$.*

[1] In practice, this typically means heuristic optimization.

[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 some Hamiltonian H and evolution time $t \geq 0$, i.e. $U = e^{iHt}$.

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 $e^{i\theta_j G_j}$, 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[3]. (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 5). 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 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 $N^{1-\epsilon}$:

▶ **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 angles* $(\theta_1, \ldots, \theta_m)$ *that minimize expectation* $\langle\psi|M|\psi\rangle$ *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 [14] 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 [14], in which neither H_b nor H_c need be diagonal in the standard

[3] 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 $U_m \cdots U_1$.

basis. (In this sense, our definition is closer to the more general Quantum Alternating Operator Ansatz, also with acronym QAOA [23].) For our hardness results, it will suffice for H_b and H_c to be *k-local* Hamiltonians[4]. 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* $m'/m \geq N^{1-\epsilon}$*, for* N *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 [14]) 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 [12] have shown an interesting negative result – QAOA to any *constant* level/depth cannot outperform the classical Goemans-Williams algorithm for MAX CUT [19]. 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 [15] 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 [9], 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 [9]:

1. [9] 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.

[4] 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.

2. [9] 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.) 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[5]), 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 [14]).

In this direction, two crucial problems can arise in the classical optimization part of the standard VQA setup: (i) barren plateaus [30], which lead to vanishing gradients, and (ii) local minima [9], 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 [30] can be avoided by tailored initialization [40], and there are indications that barren plateaus are a less significant challenge than traps [3]. In general, the optimization can be improved using natural gradients [36], multitask learning type approach [39], optimization based on trigonometric model functions [26], neural network-based optimization methods [31], brick-layer structures of generic unitaries [32], and operator pool-based methods [22, 11]. ADAPT-VQEs [22] 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 [21]. CoVar [11] is based on similar ideas complemented with estimating several properties of the variational state in parallel using classical shadows [24]. The optimization strategies are of a heuristic nature, and analytic results are scarce. Finally, it has been numerically observed [33, 37] and analytically shown [27] 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[6] for both MIN-VQA and MIN-QAOA.

To begin, recall that in a QCMA proof system (Definition 5), 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

[5] Note that for $\log(N)$-size instances of QAOA as in [9], 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.

[6] The prover sends angles θ_j, and the verifier simulates each $e^{i\theta_j H_j}$ via known Hamiltonian simulation algorithms [29].

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 [5, 6] for such results, which remains a notorious open question for both QCMA and QMA[7] [1]. (2) Problem 1 places no restrictions on which Hamiltonians are applied, in which order, and with which rotation angles. 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 [34]. 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 [18] then showed that [34] 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 6) 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 [25] 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[8] "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 [16]. 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 $P_j := X_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}}$ 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

$$Q_j := (V_j)_{AB} \otimes |01\rangle\langle 10|_{C_{|A|+j},C_{|A|+j+1}} + (V_j^\dagger)_{AB} \otimes |10\rangle\langle 01|_{C_{|A|+j},C_{|A|+j+1}}, \tag{1}$$

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

[7] Quantum Merlin-Arthur (QMA) is QCMA but with a quantum proof.

[8] 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.

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 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) $\frac{g'|D|+L}{g|D|+L} \approx \frac{g'}{g}$ for $|D| \in \omega(L)$, 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 definition, 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

$$S \coloneqq \Big\{ V_{s-|A|} \cdots V_1 |y\rangle_A |0 \cdots 0\rangle_B |\widetilde{s}\rangle_C |\widetilde{t}\rangle_D \mid y \in \{0,1\}^{|A|}, s \in \{1,\ldots,|C|\}, t \in \{1,\ldots,|D|\} \Big\},$$

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 [9] – we introduce certain local terms G_j (Equation (27)) 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 (23) of Problem 3). This is in contrast to MIN-VQA, where the initial state $|0\cdots 0\rangle$ 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 the definition of S, 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 20).

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 [14]? 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 [7], Boulebnane and Montanaro [10], and Anshu and Metger [4], 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 [4], 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 [10] 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 [17, 20, 35], 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?

Organization. In Section 2, we show Theorem 1. Section 3 shows Theorem 4. All omitted proofs are in the full version.

2 QCMA-hardness of approximation for VQAs

In this section, we show Theorem 1. We begin in Section 2.1 with relevant definitions and lemmas. Section 2.2 proves Theorem 1.

2.1 Definitions and required facts

Throughout, the relation $:=$ denotes a definition, and $[n] := \{1, 2, \ldots, n\}$. 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.

▶ **Definition 5** (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 $\Pr[V_x \text{ accepts } y] \geq 2/3$ and*

(ii) *if $x \in \Pi_{\mathrm{no}}$ and all strings $y \in \{0,1\}^{p(|x|)}$ it holds that $\Pr[V_x \text{ accepts } y] \leq 1/3$.*

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 $\Pr[V_x \text{ accepts } (x,y)] = \langle y;0|V_x^{(n)\dagger} M^{(B_1)} V_x^{(n)}|y;0\rangle$, 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 [2], 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?") [38] and Ground State Connectivity (GSCON) (i.e. is the ground space of a local Hamiltonian "connected"?) [17]. The latter remains hard (specifically, $\mathsf{QCMA}_{\mathrm{EXP}}$-hard) in the 1D translation-invariant setting [35]. Next, we will introduce a QCMA-complete problem related to monotone sets.

▶ **Definition 6** (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 7** (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 a monotone.*

▶ **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 8** (Gharibian and Kempe [18]). *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.

2.2 QCMA-completeness

▶ **Theorem 1.** MIN-VQA(k,l) *is* QCMA*-complete for $k \geq 4$, $l = 2$, and $m \leq$ 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 $m' \geq m$).

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 [28, 29]. 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'. Suppose V' takes in n'_V input qubits in register A' and m'_V ancilla qubits in register B'. To ease our soundness analysis, we make two assumptions about V' without loss of generality:

▶ **Assumption 9.** *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 10.** *The output qubit of V' is set to $|0\rangle$ until V'_L is applied. For 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 $n_V := n'_V$ input qubits, $m_V := m'_V + n'_V + 1$ ancilla qubits, and consists of $L := L' + n'_V + 1$ gates.

Proof organization. The remainder of the proof is organized as follows. Section 2.2.1 constructs the MIN-VQA instance. Section 2.2.2 proves observations and lemmas required for the completeness and soundness analyses. Sections 2.2.3 and 2.2.4 show completeness and soundness, respectively. Finally, Section 2.2.5 analyzes the hardness ratio achieved. All omitted proofs are in the full version.

2.2.1 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 $0 < \delta < 1$ chosen at the end of the proof in Section 2.2.5.

Our construction will ensure that C (respectively, D) always remains in the span of logical time steps, $\mathcal{T}_C := \{|\widetilde{s}\rangle\}_{s=1}^{|C|}$ (respectively, $\mathcal{T}_D := \{|\widetilde{t}\rangle\}_{t=1}^{|D|}$), defined as:

$$|\widetilde{s}\rangle := |0\rangle^{\otimes s-1}|1\rangle|0\rangle^{\otimes |C|-s} \quad \text{for } 1 \leq s \leq |C| \tag{2}$$

$$|\widetilde{t}\rangle = |0\rangle^{\otimes t-1}|1\rangle|0\rangle^{\otimes |D|-t} \quad \text{for } 1 \leq t \leq |D|\,. \tag{3}$$

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 $|1\rangle^{\otimes t}|0\rangle^{\otimes N-t}$ [25]. 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:

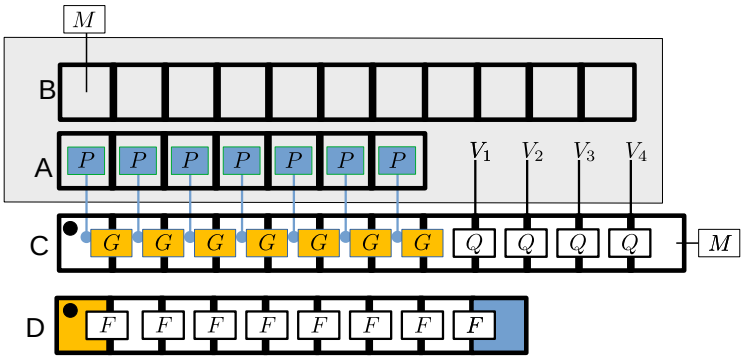

Figure 1 Sketch describing the VQA instance. A colored square (say, blue) at index j of a register means that register's jth qubit must be in $|1\rangle$ for any blue gates to act non-trivially. So, for example, the G gates increment the first clock register C, but only if the D register is in the state $|1\rangle_{D_1}$. 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.

- (F) For propagation of the second clock, D, define 2-local Hamiltonians as

$$F_j \coloneqq |01\rangle\langle 10|_{D_{j,j+1}} + |10\rangle\langle 01|_{D_{j,j+1}} \text{ for all } j \in \{1, \ldots, |D| - 1\}. \tag{4}$$

- (G) For propagation of the first clock, C, define 3-local Hamiltonians as

$$G_j \coloneqq \left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1} \text{ for all } j \in \{1, \ldots, |A|\}. \tag{5}$$

- (P) For each qubit $j \in \{1, \ldots, |A|\}$ of A, define 3-local Hamiltonian as

$$P_j \coloneqq X_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}}. \tag{6}$$

- (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

$$Q_k \coloneqq (V_k)_{R_k} \otimes |01\rangle\langle 10|_{C_{|A|+k,|A|+k+1}} + (V_k^\dagger)_{R_k} \otimes |10\rangle\langle 01|_{C_{|A|+k,|A|+k+1}}. \tag{7}$$

Denote the union of these four sets of Hamiltonians as $S_{FGPQ} \coloneqq F \cup G \cup P \cup Q$. Set a 2-local observable

$$M \coloneqq I - |1\rangle\langle 1|_{B_1} \otimes |1\rangle\langle 1|_{C_{|C|}} \tag{8}$$

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 Table 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

$$|\phi\rangle \coloneqq |0 \cdots 0\rangle_{AB} |10^{|C|-1}\rangle_C |10^{|D|-1}\rangle_D = |0 \cdots 0\rangle_{AB} |\widetilde{1}\rangle_C |\widetilde{1}\rangle_D, \tag{9}$$

Table 1 Terms used in the proof of Theorem 1.

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. 9 and 10	$V = V_L \cdots V_1$
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	$\lvert D\rvert = \lceil L^{1+\delta} \rceil$, see Section 2.2.5 for δ
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.	$m = g \cdot (2\lvert D\rvert - 1) + \lvert A\rvert + L$, $m' = g' \cdot (2\lvert D\rvert - 1) + \lvert A\rvert + L$.

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.

▶ **Observation 11.** *Fix any standard basis state* $|x\rangle_{ABCD} = \overline{X}|0\cdots 0\rangle_{ABCD}$*, for* $\overline{X} \coloneqq X_1^{x_1} \otimes \cdots \otimes X_N^{x_N}$ *with* $N \coloneqq |A| + |B| + |C| + |D|$*. Consider the updated set* $S'_{FGPQ} \coloneqq \left\{\overline{X}H\overline{X} \mid H \in S_{FGPQ}\right\}$*, 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 from* S_{FGPQ}*,*

$$e^{i\theta_m H_m} \cdots e^{i\theta_2 H_2} e^{i\theta_1 H_1} |x\rangle_{ABCD} = e^{i\theta_m H'_m} \cdots e^{i\theta_2 H'_2} e^{i\theta_1 H'_1} |0\cdots 0\rangle_{ABCD}. \tag{10}$$

Moreover, each H *and* H' *have the same locality.*

2.2.2 Helpful observations and lemmas

We next state all observations and technical lemmas for the later correctness analysis of our construction. All omitted proofs are in the full version.

▶ **Observation 12.** *For all $\theta \in \mathbb{R}$, and all $F_j \in F$, $G_j \in G$, $P_j \in P$ and $Q_k \in Q$,*

$$e^{i\theta F_j} = \cos(\theta)(|01\rangle\langle 01| + |10\rangle\langle 10|)_{D_{j,j+1}} + i\sin(\theta)F_j + (I - |01\rangle\langle 01| - |10\rangle\langle 10|)_{D_{j,j+1}} \tag{11}$$

$$e^{i\theta G_j} = \cos(\theta)\left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1} + i\sin(\theta)G_j + \left(I - \left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1}\right) \tag{12}$$

$$e^{i\theta P_j} = (\cos(\theta)I + i\sin(\theta)X)_{A_j} \otimes |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}} + (I - |1\rangle\langle 1|_{C_j} \otimes |1\rangle\langle 1|_{D_{|D|}}) \tag{13}$$

$$e^{i\theta Q_k} = \cos(\theta)I_{AB} \otimes (|01\rangle\langle 01| + |10\rangle\langle 10|)_{C_{|A|+k,|A|+k+1}} + i\sin(\theta)Q_k + I_{AB} \otimes (I - |01\rangle\langle 01| - |10\rangle\langle 10|)_{C_{|A|+k,|A|+k+1}}. \tag{14}$$

Any register not explicitly listed in equations above is assumed to be acted on by identity.

▶ **Definition 13** (Support only on logical time steps)**.** *We say state $|\psi\rangle_{ABCD}$ is* supported only on logical time steps *if it can be written $|\psi\rangle_{ABCD} = \sum_{s=1}^{|C|}\sum_{t=1}^{|D|}\alpha_{st}|\eta_{st}\rangle_{AB}|\widetilde{s}\rangle_C|\widetilde{t}\rangle_D$ 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 (2) and Equation (3), respectively.*

▶ **Observation 14.** *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 $H_j \in S_{FGPQ}$, $e^{i\theta_m H_m}\cdots e^{i\theta_2 H_2}e^{i\theta_1 H_1}|\phi\rangle$ is supported only on logical time steps.*

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 15.** *Define*

$$S := \left\{V_{s-|A|}\cdots V_1|y\rangle_A|0\cdots 0\rangle_B|\widetilde{s}\rangle_C|\widetilde{t}\rangle_D \mid y \in \{0,1\}^{|A|}, s \in \{1,\ldots,|C|\}, t \in \{1,\ldots,|D|\}\right\}, \tag{15}$$

where we adopt the convention that the V gates are present only when $s > |A|$. Then, for any $m \in \mathbb{N}$, $\Pi_{u=1}^m e^{i\theta_u H_u}|\phi\rangle \in \mathrm{Span}(S)$ for any angles $\theta_u \in \mathbb{R}$ and sequence of Hamiltonians $H_u \in S_{FGPQ}$.

Next, we relate the circuit depth of a state generated by our VQA to the Hamming weight of the proof string y.

▶ **Lemma 16.** *Let $(H_u)_{u=1}^m$ be a sequence of Hamiltonians drawn from S_{FGPQ} which maps the initial state (9) to $|\phi_m\rangle := \Pi_{u=1}^m e^{i\theta_u H_u}|\phi\rangle$. 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 $F \cup G \cup P$.*

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 17.** *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*

$$\langle\phi_m|M|\phi_m\rangle = 1. \tag{16}$$

Proof. By Lemma 15, $|\phi_m\rangle \in S$ for S from Equation (15). Next, by Observation 12, 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 10 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. ◀

2.2.3 Completeness

With all observations and lemmas of Section 2.2.2 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 12 and Equation (7) imply that

$$e^{i\frac{\pi}{2}Q_j}|\psi_{j-1}\rangle = \left((V_j)_{R_j} \otimes |\widetilde{|A|+j+1}\rangle\langle\widetilde{|A|+j}|_C\right)|\psi_{j-1}\rangle, \tag{17}$$

  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 $\langle\psi_L|M|\psi_L\rangle \le 1/3$, 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.

2.2.4 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 15 says we may write

$$|\phi_{m'}\rangle = \sum_{y\in\{0,1\}^{|A|}} \sum_{s=1}^{|C|} \sum_{t=1}^{|D|} \alpha_{y,s,t} |\eta_{y,s,t}\rangle \in \mathrm{Span}(S) \tag{18}$$

with $\sum_{y,s,t} |\alpha_{y,s,t}|^2 = 1$. Now, for the observable (8) it follows that

$$\langle\phi_{m'}|M|\phi_{m'}\rangle = 1 - \langle\phi_{m'}| \left(|1\rangle\langle 1|_{B_1} \otimes |1\rangle\langle 1|_{C_{|C|}}\right) |\phi_{m'}\rangle = 1 - \langle\eta||1\rangle\langle 1|_{B_1}|\eta\rangle \text{ for} \tag{19}$$

$$|\eta\rangle \coloneqq \sum_{y\in\{0,1\}^{|A|}} \sum_{t=1}^{|D|} \alpha_{y,|A|+L+1,t} V_L \cdots V_1 |y\rangle_A |0\rangle_B ||A|\widetilde{+L}+1\rangle_C |\widetilde{t}\rangle_D, \tag{20}$$

where we have used Equation (18) 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 17 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 16 hence implies the right hand side of Equation (19) equals $1-\langle\eta_{g'}||1\rangle\langle 1|_{B_1}|\eta_{g'}\rangle$ for[9]

$$|\eta_{g'}\rangle \coloneqq \sum_{y \text{ s.t. } \mathrm{HW}(y)\leq g'} \sum_{t=1}^{|D|} \alpha_{y,|A|+L+1,t} V_L \cdots V_1 |y\rangle_A |0\rangle_B ||A|\widetilde{+L}+1\rangle_C |\widetilde{t}\rangle_D, \tag{21}$$

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.

2.2.5 Hardness ratio

Finally, we show our reduction has the desired approximation ratio. Observe

$$\frac{m'}{m} = \frac{g'(2\,|D|-1)+|A|+L}{g(2\,|D|-1)+|A|+L} = \frac{g'(2\lceil L^{1+\delta}\rceil-1)+|A|+L}{g(2\lceil L^{1+\delta}\rceil-1)+|A|+L}. \tag{22}$$

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 8 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'. By appropriately comparing N' to the encoding size N of our MIN-VQA instance Π, one can in fact show that for any $\epsilon>0$, $g'/g \geq N^{1-\epsilon}$ for large enough V', as desired. The proof is in the full version. ◀

3 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 [14], and closer to that of [23].

[9] Below, $\mathrm{HW}(y)$ denotes the Hamming weight of string y.

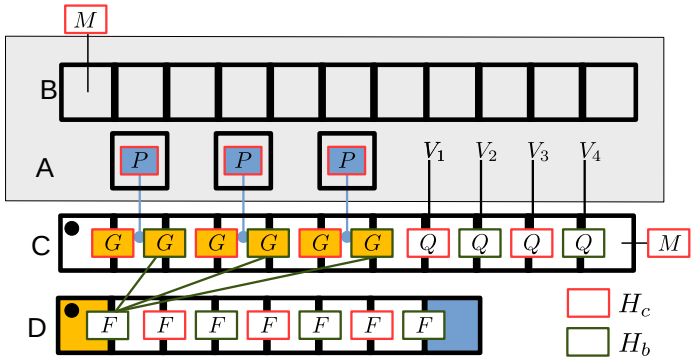

Figure 2 Figure describing the QAOA instance (see Figure 1 for further details). The border color of each gate indicates if the generator belongs to H_b or H_c. Compared to the previous VQA instance, the P now only act at even time steps in C and the even-indexed G_j and the F_1 generator are combined into one generator, denoted by the red and dark green edges.

▶ **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*[10] $(\theta_i)_{i=1}^{m} \in \mathbb{R}^m$, *such that*

$$|\psi\rangle := e^{i\theta_m H_b} e^{i\theta_{m-1} H_c} \cdots e^{i\theta_2 H_b} e^{i\theta_1 H_c} |\mathrm{gs}_b\rangle \tag{23}$$

   satisfies $\langle\psi|H_c|\psi\rangle \leq \frac{1}{3}$.
2. *NO if for all sequences of angles $(\theta_i)_{i=1}^{m'} \in \mathbb{R}^{m'}$*

$$|\psi\rangle := e^{i\theta_{m'} H_b} e^{i\theta_{m'-1} H_c} \cdots e^{i\theta_2 H_b} e^{i\theta_1 H_c} |\mathrm{gs}_b\rangle, \tag{24}$$

   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 $m \leq \mathrm{poly}(n)$. Moreover, for any $\epsilon > 0$, it is* QCMA*-hard to distinguish between the YES and NO cases of* MIN-QAOA *even if $m'/m \geq N^{1-\epsilon}$, for N 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 with $V = V_L \cdots V_1$ being a sequence of L two-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.

[10] Throughout Problem 3, for clarity we assume all angles are specified to $\mathrm{poly}(n)$ bits.

Proof organization. The proof is organized as follows. In Section 3.1 we explain the modifications of the VQA instances to obtain the QAOA instances of our construction. Section 3.1.2 and Section 3.1.3 explain how we recover the desired initial state and cost function. Section 3.1.4 provides notation preliminary technical results needed for the QCMA-completeness proof. Then, completeness is shown in Section 3.1.5 and soundness in Section 3.1.6. Finally, in Section 3.1.7, we analyze the hardness ratio achieved by the reduction. All omitted proofs are in the full version.

3.1 QCMA completeness for QAOAs

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 2. Briefly recapping the proof techniques outline in Section 1.3, we:

- **(i)** 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,
- **(ii)** introduce terms G_j from Equation (27) with 3-cyclic behavior, i.e. allowing three logical actions,
- **(iii)** add new constraints to H_b to ensure its unique ground state encodes the correct start state (see Equation (23) of Problem 3), and
- **(iv)** 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 will show via Lemma 20 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,

$$F_j \coloneqq |01\rangle\langle 10|_{D_{j,j+1}} + |10\rangle\langle 01|_{D_{j,j+1}} - 2|00\rangle\langle 00|_{D_{j,j+1}} \text{ for all } j \in \{2, \dots, |D| - 1\}. \quad (25)$$

- (G) We double the number of qubits G acts on,

$$G_j \coloneqq \left(|01\rangle\langle 10|_{C_{j,j+1}} + |10\rangle\langle 01|_{C_{j,j+1}}\right) \otimes |1\rangle\langle 1|_{D_1} - 2|001\rangle\langle 001|_{C_{j,j+1},D_1} \quad \text{for all } j \in \{1, 3, \dots, 2|A| - 1\}, \quad (26)$$

$$G_j \coloneqq \frac{i}{\sqrt{3}}\big(|0110\rangle\langle 1010| + |1001\rangle\langle 0110| + |1010\rangle\langle 1001| - |1010\rangle\langle 0110| - |0110\rangle\langle 1001| - |1001\rangle\langle 1010|\big)_{C_{j,j+1},D_{1,2}} - 2|0010\rangle\langle 0010|_{C_{j,j+1},D_{1,2}} \quad \text{for all } j \in \{2, 4, \dots, 2|A|\}. \quad (27)$$

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:

$$e^{i\frac{\pi}{3}G_i}|\widetilde{i},\widetilde{1}\rangle_{C,D} = e^{i\frac{\pi}{2}G_i^{(o)}}|\widetilde{i},\widetilde{1}\rangle_{C,D} \propto |\widetilde{i+1},\widetilde{1}\rangle_{C,D}, \quad (28)$$

$$e^{i\frac{2\pi}{3}G_i}|\widetilde{i},\widetilde{1}\rangle_{C,D} = e^{i\frac{\pi}{2}F_1^{(o)}}|\widetilde{i},\widetilde{1}\rangle_{C,D} \propto |\widetilde{i},\widetilde{2}\rangle_{C,D}, \quad (29)$$

where, in this case, "$\propto$" means equality up to a phase.

- (P) For each qubit $j \in \{1, \ldots, |A|\}$ of A, we define the X-operators, but now they only act on even values in the clock register,

$$P_j \coloneqq X_{A_j} \otimes |1\rangle\langle 1|_{C_{2j}} \otimes |1\rangle\langle 1|_{D_{|D|}} - 2|00\rangle\langle 00|_{C_{2j}, D_{|D|}} \text{ for all } j \in \{1, \ldots, |A|\}. \tag{30}$$

- (Q) We shift the C-indices of the Q-gates because reading in the proof takes longer time now,

$$Q_k \coloneqq (V_k)_{R_k} \otimes |01\rangle\langle 10|_{C_{2|A|+k, 2|A|+k+1}} + (V_k^\dagger)_{R_k} \otimes |10\rangle\langle 01|_{C_{2|A|+k, 2|A|+k+1}} \tag{31}$$

$$- 2|00\rangle\langle 00|_{C_{2|A|+k, 2|A|+k+1}} \text{ for all } k \in \{1, \ldots, L\} \tag{32}$$

- (M), (H_0) We add the two operators

$$H_0 = -\left(\sum_{i \in [|A|]} |0\rangle\langle 0|_{A_i} + \sum_{i \in [|B|]} |0\rangle\langle 0|_{B_i} \right) \otimes |1\rangle\langle 1|_{C_1}, \tag{33}$$

$$M = I - |1\rangle\langle 1|_{B_1} \otimes |1\rangle\langle 1|_{C_{|C|}} \tag{34}$$

to the set of generators.

To construct our desired QAOA instance, we define a partition of all gates into two groups:

$$\mathcal{G}_1 = \{G_i\}_{i \in \{2,4,\ldots,2|A|\}} \cup \{F_i\}_{i \in \{3,5,\ldots,|D|-1\}} \cup \{Q_i\}_{i \in \{2,4,\ldots,L\}}, \tag{35}$$

$$\mathcal{G}_2 = \{G_i\}_{i \in \{1,3,\ldots,2|A|-1\}} \cup \{F_i\}_{i \in \{2,4,\ldots,|D|-2\}} \cup \{Q_i\}_{i \in \{1,3,\ldots,L-1\}} \cup \{P_i\}_{i \in [|A|]}. \tag{36}$$

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 9 and Assumption 10 from the VQA section, we shall also use the following.

▶ **Assumption 18.** *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 k times repetition of the circuit V, followed by a majority vote. This increases 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 $k = O(\log(N))$.*

Due to this assumption, our encoding size increases by a multiplicative log factor, which does not affect our final approximation ratio calculation.

3.1.1 The Min-QAOA instance

The QAOA instance we use to prove Theorem 4 takes the generators

$$H_b = \sum_{\Gamma \in \mathcal{G}_1} \Gamma + H_0\,, \tag{37}$$

$$H_c = \kappa \sum_{\Gamma \in \mathcal{G}_2} \Gamma + M \tag{38}$$

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 18, 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 first characterize the initial state and cost function as defined in Problem 3.

3.1.2 Initial state

Recall that in Problem 3 the initial state has to be a ground state of H_b (given as input via a preparation circuit U_b). We want this initial state to be

$$|\mathrm{gs}_b\rangle = |0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD}, \tag{39}$$

which can trivially be prepared by a constant-sized circuit U_b. To see that we indeed obtain this ground state, note below 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 turns out to be unique because for each qubit, the state is uniquely determined by one of the generators, which implies that the entire state is unique. Specifically, we have

$$\begin{aligned}
F_i|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -2|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} \quad \forall i \in \{3,5,\dots,|D|-1\}, \quad &\|F_i\|_\infty &= 2,\\
G_i|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -2|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} \quad \forall i \in \{2,4\dots,2\,|A|\}, &\|G_i\|_\infty &= 2,\\
Q_i|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -2|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} \quad \forall i \in \{2,4\dots,L\}, &\|Q_i\|_\infty &= 2,\\
H_0|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD} &= -(|A|+|B|)|0,0,\widetilde{1},\widetilde{1}\rangle_{ABCD}, &\|H\|_\infty &= |A|+|B|\,.
\end{aligned}$$

Indeed, since the state (39) 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.

3.1.3 Cost function

In the QAOA setup, the measured observable is H_c. For our construction we wish to use the observable M. Fortunately, we can find an upper bound for the difference of these operators. Namely, for every $|\Psi\rangle \in \mathcal{H}$

$$|\langle\Psi|(H_c - M)|\Psi\rangle| = \kappa|\langle\Psi|\textstyle\sum_{\Gamma\in\mathcal{G}_2}\Gamma|\Psi\rangle| \leq 2\kappa|\mathcal{G}_2| \leq \frac{1}{12} \tag{40}$$

where we used (1) that $\|g\|_\infty \leq 2$ for all $\Gamma \in \mathcal{G}_2$, and (2) the definition of κ.

3.1.4 Preliminaries for the completeness proof

We first define the set of states comprising our logical computation space,

$$S \coloneqq \{V_{t-2|A|-1}\cdots V_1|y\rangle_A|0\cdots 0\rangle_B|\widetilde{t}\rangle_C|\widetilde{s}\rangle_D \,|\, \forall (y,t,s) \in I_S\} \tag{41}$$

with

$$I_S = \left\{(y,t,s)\middle| y\in\{0,1\}^{|A|}, t\in\{1,\dots,|C|\}, s\in\begin{cases}\{1,\dots,|D|\} & \text{if } t\in\{2,4,\cdots 2\,|A|\}\\ \{1\} & \text{otherwise}\end{cases}\right\}$$

being the allowed index set. Here, the notation means that V_1 is only applied if $t > 2\,|A|+1$. Below, we often write a state $|\Psi_S\rangle \in \mathrm{span}(S)$ as

$$|\Psi_S\rangle = \sum_{(y,t,s)\in I_S} a_{y,t,s} V_{t-2|A|-1}\cdots V_1|y\rangle_A|0\cdots 0\rangle_B|\widetilde{t}\rangle_C|\widetilde{s}\rangle_D =: \sum_{(y,t,s)\in I_S} a_{y,t,s}|\Psi_{y,t,s}\rangle\,.$$

We also define the function W, which is intended to capture a lower bound on the number of Hamiltonian evolutions required to prepare a given logical state $|\Psi_{y,t,s}\rangle$:

$$W(y,t,s) \coloneqq (2\,|D|-4)\mathrm{HW}(y) + t + (-1)^{\delta_{y_{\lceil t/2 \rceil},1}}(s+\delta_{s,1}-2), \tag{42}$$

where $\mathrm{HW}(y)$ denotes the Hamming weight of y.

We next show a helpful lemma regarding the action of each Hamiltonian on our logical computation space, S.

▶ **Lemma 19.** *The following two statements hold:*

- *For every $|\Psi_{y,t,s}\rangle \in S$ and $H_i \in \{H_b, H_c\}$, $e^{\mathrm{i}H_i\theta}|\Psi_{y,t,s}\rangle = e^{\mathrm{i}\alpha^{(i)}_{y,t,s}\theta}e^{\mathrm{i}\Gamma^{(i)}_{y,t,s}\theta}|\Psi_{y,t,s}\rangle$ 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 $\Gamma \in \mathcal{G}_1 \cup \mathcal{G}_2 \cup \{H_0\}$, $\exists$ amplitudes $\{a_{y,t,s}\}$ such that*

$$e^{\mathrm{i}\Gamma\theta}|\Psi_{y,t,s}\rangle = \sum_{\substack{(y',t',s')\in I_S \\ W(y',t',s')\leq W(y,t,s)+1}} a_{y,t,s}|\Psi_{y',t',s'}\rangle \tag{43}$$

In words, the application of Γ can only increase value of the W-function by at most 1.

3.1.5 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 the verifier circuit V. We use shorthand notation $(y)_j = (y_1, \ldots y_{j-1}, 0, \ldots, 0)$ 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 19, claim 1). The honest prover proceeds as follows:

- (Prepare classical proof) Prepare state (up to global phase) $|\psi_0\rangle \coloneqq |y\rangle_A|0\rangle_B|\widetilde{2\,|A|+1}\rangle_C|\widetilde{1}\rangle_D$ 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
  $$|(y)_{j-1}, 0, \widetilde{2j-1}, \widetilde{1}\rangle_{ABCD} \quad\mapsto\quad |(y)_{j-1}, 0, \widetilde{2j}, \widetilde{1}\rangle_{ABCD}. \tag{44}$$
  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.
     $$|(y)_{j-1}, 0, \widetilde{2j}, \widetilde{1}\rangle_{ABCD} \quad\mapsto\quad |(y)_{j-1}, 0, \widetilde{2j}, \widetilde{2}\rangle_{ABCD}. \tag{45}$$
     - 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.
     $$|(y)_{j-1}, 0, \widetilde{2j}, \widetilde{2}\rangle_{ABCD} \quad\mapsto\quad |(y)_{j-1}, 0, \widetilde{2j}, |\widetilde{D}|\rangle_{ABCD}. \tag{46}$$
     - 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.
     $$|(y)_{j-1}, 0, \widetilde{2j}, |\widetilde{D}|\rangle_{ABCD} \quad\to\quad |(y)_{j}, 0, \widetilde{2j}, |\widetilde{D}|\rangle_{ABCD}. \tag{47}$$
     - 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.
     $$|(y)_{j}, 0, \widetilde{2j}, |\widetilde{D}|\rangle_{ABCD} \quad\mapsto\quad |(y)_{j}, 0, \widetilde{2j}, \widetilde{2}\rangle_{ABCD}. \tag{48}$$

- 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.

$$|(y)_j, 0, \widetilde{2j}, \widetilde{2}\rangle_{ABCD} \quad \mapsto \quad |(y)_j, 0, \widetilde{2j+1}, \widetilde{1}\rangle_{ABCD} \tag{49}$$

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.

$$|(y)_{j-1}, 0, \widetilde{2j}, \widetilde{1}\rangle_{ABCD} \quad \mapsto \quad |(y)_j, 0, \widetilde{2j+1}, \widetilde{1}\rangle_{ABCD}. \tag{50}$$

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, $\exp(\mathrm{i}\frac{\pi}{2\kappa}H_c) \sim \exp(\mathrm{i}\frac{\pi}{2}Q_1)$, $\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}Q_2)$, $\ldots$, $\exp(\mathrm{i}\frac{\pi}{2}H_b) \sim \exp(\mathrm{i}\frac{\pi}{2}Q_L)$ for a total L gates. This implements the verifier, i.e.

$$|y, 0, 2\widetilde{|A|+1}, \widetilde{1}\rangle_{ABCD} \quad \to \quad |\Psi_L\rangle \coloneqq V_L \cdots V_1 |y, 0, \widetilde{|C|}, \widetilde{1}\rangle_{ABCD}. \tag{51}$$

Since V accepts proof y of the QMSA instance with probability at least $1 - \epsilon_Q$, we conclude using Equation (40) that

$$\langle \Psi_L | H_c | \psi_L \rangle \leq \langle \Psi_L | M | \Psi_L \rangle + \frac{1}{12} \leq 1 - (1 - \epsilon_Q) + \frac{1}{12} \leq \frac{1}{3} \tag{52}$$

as desired. The number of Hamiltonians applied in this case is $m = g(2|D| - 2) + 2|A| + L = g(2|D| - 2) + |C| - 1$, also as desired.

3.1.6 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, for our MIN-QAOA construction, the M operator (embedded in H_c) does *not* necessarily preserve the space $\mathrm{span}(S)$ (see Claim 2 of Lemma 19). 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 20** (Rounding lemma). *In the NO case, after $\zeta \geq 1$ applications of H_c and H_b, $|\Psi_\zeta\rangle \in \Gamma_\zeta \coloneqq \left\{ \prod_{i=1}^{\zeta} e^{iH_i\theta_i} |\mathrm{gs}_b\rangle \mid H_i \in \{H_b, H_c\}, \theta \in \mathbb{R}^\zeta \right\}$ will be $\epsilon \leq 4\zeta\sqrt{\epsilon_Q}$ close to the span of S i.e.*

$$\forall |\Psi_\zeta\rangle \in \Gamma_\zeta,\ \exists |\Psi'_\zeta\rangle \in \mathrm{span}(S) : \left\| |\Psi_\zeta\rangle\langle\Psi_\zeta| - |\Psi'_\zeta\rangle\langle\Psi'_\zeta| \right\|_{\mathrm{tr}} \leq 4\zeta\sqrt{\epsilon_Q} \tag{53}$$

and it additionally holds that $|\Psi'_\zeta\rangle = \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\leq\zeta+1}} a_{y,t,s} |\Psi_{y,t,s}\rangle$.

The proof is in the full version. This lemma is needed because the time evolution of the observable M (in H_c) may leave the sub-space $\mathrm{Span}(S)$. 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 18), meaning the evolution in M only adds to a global phase.

We are finally ready to prove soundness. For this, we need to show that in the NO case, all sequences of $\zeta \leq m' = g'(2\,|D| - 4) + |C| - 1$ gates produce cost function value $\langle \Psi_\zeta | H_c | \Psi_\zeta \rangle \geq \frac{2}{3}$. This follows since for all $\zeta \leq m'$,

$$\langle \Psi_\zeta | H_c | \Psi_\zeta \rangle \geq \langle \Psi_\zeta | M | \Psi_\zeta \rangle - \frac{1}{12} \tag{54}$$

$$\geq \langle \Psi'_\zeta | M | \Psi'_\zeta \rangle - |\mathrm{Tr}[M(|\Psi_\zeta\rangle\langle\Psi_\zeta| - |\Psi'_\zeta\rangle\langle\Psi'_\zeta|)]| - \frac{1}{12} \tag{55}$$

$$\geq \langle \Psi'_\zeta | M | \Psi'_\zeta \rangle - \|M\|_\infty \left\| |\Psi_\zeta\rangle\langle\Psi_\zeta| - |\Psi'_\zeta\rangle\langle\Psi'_\zeta| \right\|_{\mathrm{tr}} - \frac{1}{12} \tag{56}$$

$$\geq \langle \Psi'_\zeta | M | \Psi'_\zeta \rangle - 4m' \sqrt{\epsilon_Q} - \frac{1}{12} \tag{57}$$

$$\geq \langle \Psi'_\zeta | M | \Psi'_\zeta \rangle - \frac{1}{6} \tag{58}$$

where the first statement follows from Equation (40), the third by Hölder's inequality, the fourth by Lemma 20, and the last since $\sqrt{\epsilon_Q} \leq \frac{1}{48m'}$. By Lemma 19, we can expand $|\Psi'_\zeta\rangle$ in the basis $|\Psi'_\zeta\rangle = \sum_{\substack{(y,t,s)\in I_S \\ W(y,t,s)\leq m'+1}} a_{y,t,s} |\Psi_{y,t,s}\rangle$ which gives

$$\langle \Psi'_\zeta | M | \Psi'_\zeta \rangle = 1 - \sum_{y\in\{0,1\}^{|A|} | \mathrm{HW}(y)\leq g'} |a_{y,|C|,1}|^2 \langle \Psi_{y,|C|,1} | |1\rangle\langle 1|_{B_1} | \Psi_{y,|C|,1} \rangle \geq 1 - \epsilon_Q \tag{59}$$

as M only acts non-trivial on $t = |C|$ and $W(y, |C|, 1) \leq m' + 1$ reduces to $\mathrm{HW}(y) \leq g'$, and in the NO case QMSA accepts such a y with at most ϵ_Q probability. Combining the two results we get

$$\langle \Psi_\zeta | H_c | \Psi_\zeta \rangle \geq 1 - \epsilon_Q - \frac{1}{6} > \frac{2}{3} \tag{60}$$

which shows soundness for all gates-sequences of length $\zeta \leq m'$.

3.1.7 Hardness ratio

The analysis is essentially identical to that for MIN-VQA, and is in the full version. ◀

References

1 Dorit Aharonov, Itai Arad, and Thomas Vidick. Guest column: The quantum PCP conjecture. *SIGACT News*, 44(2):47–79, June 2013. doi:10.1145/2491533.2491549.

2 Dorit Aharonov and Tomer Naveh. Quantum NP - a survey, 2002. arXiv:quant-ph/0210077.

3 Eric R. Anschuetz and Bobak T. Kiani. Beyond barren plateaus: Quantum variational algorithms are swamped with traps, 2022. arXiv:2205.05786.

4 Anurag Anshu and Tony Metger. Concentration bounds for quantum states and limitations on the QAOA from polynomial approximations, 2022. arXiv:2209.02715.

5 S. Arora and S. Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the ACM*, 45(1):70–122, 1998. Prelim. version FOCS '92. doi:10.1145/273865.273901.

6 Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998. Prelim. version FOCS '92. doi:10.1145/278298.278306.

7 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, 2022. doi:10.48550/arXiv.2204.10306.

8 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. *Rev. Mod. Phys.*, 94(1):015004, January 2022. doi:10.1103/RevModPhys.94.015004.

9 Lennart Bittel and Martin Kliesch. Training variational quantum algorithms is NP-hard. *Phys. Rev. Lett.*, 127:120502, September 2021. doi:10.1103/PhysRevLett.127.120502.

10 Sami Boulebnane and Ashley Montanaro. Solving boolean satisfiability problems with the quantum approximate optimization algorithm, 2022. doi:10.48550/arXiv.2208.06909.

11 Gregory Boyd and Bálint Koczor. Training variational quantum circuits with covar: covariance root finding with classical shadows, 2022. arXiv:2204.08494.

12 Sergey Bravyi, Alexander Kliesch, Robert Koenig, and Eugene Tang. Obstacles to variational quantum optimization from symmetry protection. *Phys. Rev. Lett.*, 125:260505, December 2020. doi:10.1103/PhysRevLett.125.260505.

13 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. *Nat. Rev. Phys.*, 3:625–644, 2021. doi:10.1038/s42254-021-00348-9.

14 Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm, 2014. arXiv:1411.4028.

15 Edward Farhi and Aram W Harrow. Quantum supremacy through the quantum approximate optimization algorithm, 2016. arXiv:1602.07674.

16 Richard P Feynman. Quantum mechanical computers. *Found. Phys.*, 16(6):507–531, 1986. URL: http://www.cs.princeton.edu/courses/archive/fall05/frs119/papers/feynman85_optics_letters.pdf.

17 S. Gharibian and J. Sikora. Ground state connectivity of local Hamiltonians. In *42nd International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 617–628, 2015. doi:10.1007/978-3-662-47672-7_50.

18 Sevag Gharibian and Julia Kempe. Hardness of approximation for quantum problems. In *39th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 387–398, 2012. doi:10.1007/978-3-642-31594-7_33.

19 M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995. doi:10.1145/227683.227684.

20 David Gosset, Jenish C. Mehta, and Thomas Vidick. QCMA hardness of ground space connectivity for commuting Hamiltonians. *Quantum*, 1:16, July 2017. doi:10.22331/q-2017-07-14-16.

21 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, 2022. arXiv:2204.07179.

22 Harper R. Grimsley, Sophia E. Economou, Edwin Barnes, and Nicholas J. Mayhall. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nat. Commun.*, 10:3007, July 2019. doi:10.1038/s41467-019-10988-2.

23 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. *Algorithms*, 12(2):34, 2019. doi:https://www.mdpi.com/1999-4893/12/2/34.

24 Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16:1050–1057, June 2020. doi:10.1038/s41567-020-0932-7.

25 Alexei Yu Kitaev, Alexander Shen, and Mikhail N Vyalyi. *Classical and quantum computation*, volume 47. American Mathematical Society, 2002. URL: https://bookstore.ams.org/gsm-47.

26 Bálint Koczor and Simon C. Benjamin. Quantum analytic descent. *Phys. Rev. Research*, 4(2):023017, April 2022. doi:10.1103/PhysRevResearch.4.023017.

27 Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. Theory of overparametrization in quantum neural networks, 2021. arXiv:2109.11676.

28 Seth Lloyd. Universal quantum simulators. *Science*, 273(5278):1073–1078, 1996. doi:10.1126/science.273.5278.1073.

29 Guang Hao Low and Isaac L. Chuang. Optimal Hamiltonian simulation by quantum signal processing. *Phys. Rev. Lett.*, 118:010501, January 2017. doi:10.1103/PhysRevLett.118.010501.

30 Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.*, 9:4812, November 2018. doi:10.1038/s41467-018-07090-4.

31 Javier Rivera-Dean, Patrick Huembeli, Antonio Acín, and Joseph Bowles. Avoiding local minima in variational quantum algorithms with neural networks, 2021. arXiv:2104.02955.

32 Lucas Slattery, Benjamin Villalonga, and Bryan K. Clark. Unitary block optimization for variational quantum algorithms. *Phys. Rev. Research*, 4(2):023072, April 2022. doi:10.1103/PhysRevResearch.4.023072.

33 Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity. Learning unitaries by gradient descent, 2020. arXiv:2001.11897.

34 C. Umans. Hardness of approximating Σ_2^p minimization problems. In *40th Symposium on Foundations of Computer Science*, pages 465–474, 1999.

35 J. D. Watson, J. Bausch, and S. Gharibian. The Complexity of Translationally Invariant Problems beyond Ground State Energies. In *40th Symposium on Theoretical Aspects of Computer Science (STACS 2023)*, 2023.

36 David Wierichs, Christian Gogolin, and Michael Kastoryano. Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer. *Phys. Rev. Research*, 2(4):043246, November 2020. doi:10.1103/PhysRevResearch.2.043246.

37 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. *PRX Quantum*, 1:020319, 2020. doi:10.1103/PRXQuantum.1.020319.

38 P. Wocjan, D. Janzing, and T. Beth. Two QCMA-complete problems. *Quantum Information & Computation*, 3(6):635–643, 2003. doi:10.5555/2011556.2011563.

39 Dan-Bo Zhang and Tao Yin. Collective optimization for variational quantum eigensolvers. *Phys. Rev. A*, 101(3):032311, March 2020. doi:10.1103/PhysRevA.101.032311.

40 Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D. Lukin. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Phys. Rev. X*, 10:021067, 2020. doi:10.1103/PhysRevX.10.021067.