

Quantum Space, Ground Space Traversal, and How to Embed Multi-Prover Interactive Proofs into Unentanglement

Sevag Gharibian ✉ 

Universität Paderborn, Germany

Dorian Rudolph¹ ✉ 

Universität Paderborn, Germany

Abstract

A celebrated result in classical complexity theory is Savitch’s theorem, which states that non-deterministic polynomial-space computations (NPSPACE) can be simulated by deterministic poly-space computations (PSPACE). In this work, we initiate the study of a quantum analogue of NPSPACE, denoted Streaming-QCMASPACE (SQCMASPACE), in which an exponentially long classical proof is streamed to a poly-space quantum verifier. We first show that a quantum analogue of Savitch’s theorem is unlikely to hold, in that $SQCMASPACE = NEXP$. For completeness, we also introduce the companion class Streaming-QMASPACE (SQMASPACE) with an exponentially long streamed *quantum* proof, and show $SQMASPACE = QMA_{EXP}$ (the quantum analogue of NEXP). Our primary focus, however, is on the study of exponentially long streaming *classical* proofs, where we next show the following two main results.

The first result shows that, in strong contrast to the classical setting, the solution space of a quantum constraint satisfaction problem (i.e. a local Hamiltonian) is *always* connected when exponentially long proofs are permitted. For this, we show how to simulate any Lipschitz continuous path on the unit hypersphere via a sequence of *local* unitary gates, at the expense of blowing up the circuit size. This shows that quantum error-correcting codes can be unable to detect one codeword erroneously evolving to another if the evolution happens sufficiently slowly, and answers an open question of [Gharibian, Sikora, ICALP 2015] regarding the Ground State Connectivity problem.

Our second main result is that any SQCMASPACE computation can be embedded into “unentanglement”, i.e. into a quantum constraint satisfaction problem with unentangled provers. Formally, we show how to embed SQCMASPACE into the Sparse Separable Hamiltonian problem of [Chailloux, Sattath, CCC 2012] (QMA(2)-complete for $1/\text{poly}$ promise gap), at the expense of scaling the promise gap with the streamed proof size. As a corollary, we obtain the first systematic construction for obtaining QMA(2)-type upper bounds on arbitrary multi-prover interactive proof systems, where the QMA(2) promise gap scales exponentially with the number of bits of communication in the interactive proof. Our construction uses a new technique for exploiting unentanglement to simulate quadratic Boolean functions, which in some sense allows *history* states to encode the *future*.

2012 ACM Subject Classification Theory of computation → Complexity classes; Theory of computation → Quantum complexity theory

Keywords and phrases quantum complexity theory, Quantum Merlin Arthur (QMA), QMA(2), ground state connectivity (GSCON), quantum error correction

Digital Object Identifier 10.4230/LIPIcs.ITCS.2023.53

Related Version *Full Version:* <https://arxiv.org/abs/2206.05243> [15]

Funding *Sevag Gharibian:* Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) projects 450041824 and 432788384.

¹ corresponding author



Dorian Rudolph: Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project 432788384.

Acknowledgements We thank Rolando Somma for pointing us to [9] and for interesting discussions, and Chinmay Nirke for feedback on this manuscript.

1 Introduction

Computational complexity theory studies the resources required to solve a given computational problem. The resources of time and space, in particular, are very well-studied, revealing certain interesting discrepancies. For example, while the question of whether non-deterministic poly-time (NP) equals deterministic poly-time (P) remains a central open problem in the field, in the context of *space*, the answer is well-known: In 1970, Savitch [32] gave his celebrated result that non-deterministic poly-space computations (NPSPACE) could be simulated by deterministic poly-space computations (PSPACE), yielding $\text{PSPACE} = \text{NPSPACE}$.

Motivated by the prospect of a quantum analogue of Savitch’s theorem, in this work, we initiate the study of a “non-deterministic” quantum analogue of PSPACE, which we call SQCMASPACE. To define the latter, recall that NPSPACE may be viewed as a PSPACE machine which receives an *exponential* length proof $y \in \{0, 1\}^{2^n}$. Of course, a PSPACE verifier cannot even write down y given its limited memory, so a natural way to formalize this idea is to allow y to be *streamed*, bit by bit. This is the approach we take² in defining SQCMASPACE.

► **Definition 1** (SQCMASPACE (informal; see Definition 20)). *A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in SQCMASPACE(p, q, r) for polynomially-bounded functions p, q, r , if there exist thresholds $\alpha(n), \beta(n)$ satisfying $\alpha(n) - \beta(n) \geq 2^{-r(n)}$, and a polynomial-space uniform family of quantum circuits $\{Q_n\}$ such that, for any input $x \in \Sigma^n$:*

- *If $x \in A_{\text{yes}}$, \exists streaming proof $y \in \{0, 1\}^{2^{p(n)}}$ s.t. Q_n accepts (x, y) with probability $\geq \alpha$.*
- *If $x \in A_{\text{no}}$, \forall streaming proofs $y \in \{0, 1\}^{2^{p(n)}}$, Q_n accepts (x, y) with probability $\leq \beta$.*

To avoid cluttering the introduction, we leave our formal definition of *streaming proof* to Section 2 (Definition 19 therein), and instead make do with the following intuitive definition: To “stream” the next bit y_i to the verifier, we imagine the prover applies either “proof gate” I (if $y_i = 0$) or X (if $y_i = 1$), for I and X the single-qubit identity and Pauli X (i.e. NOT) gates, respectively, to a designated qubit k in the verifier’s memory, which is initialized to $|0\rangle_k$. The verifier then copies³ this bit into its main memory via Controlled-NOT (CNOT), and the prover subsequently uncomputes bit y_i by re-applying I or X to k , respectively. In other words, there is no separate proof – we view the entire computation as a sequence of gates on the verifier’s memory, some gates of which (the “proof” gates) are *a priori* unknown. For clarity, this is similar to how communication is modelled in quantum interactive proofs, where prover and verifier take turns acting on a shared “message register” (see e.g. [25]).

In Section 1.3, we survey previous works studying quantum notions of PSPACE. Most relevant to our discussion at this point, however, is the work of Fefferman and Remschrin [12], which defines a quantum variant of NPSPACE denoted QMASPACE, and which differs from

² One can in principle consider alternative definitions of SQCMASPACE. For example, Definition 1 allows only one streaming pass of the proof, but one could consider multiple passes. An even stronger access model might allow the ability to query arbitrary single bits of the proof. For our results here, however, a single-pass streaming model suffices, e.g. this definition already captures NEXP, as we show in Theorem 5.

³ The verifier can also simulate the choice *not* to copy the bit into memory, if desired. See the discussion after Definition 19.

SQCMASPACE in three respects: The first two differences are that QMASPACE has a *poly*-length proof which is *quantum*, whereas SQCMASPACE has an *exponential* length streamed proof, which is *classical*. The third difference is that whereas QMASPACE = PSPACE [12], here we show SQCMASPACE = NEXP (Theorem 5, stated shortly). To the best of our knowledge, the current work is the first to formalize and study a quantum analogue of NPSPACE which allows an *exponentially* long classical proof.

Broader theme. Beyond initiating the study of SQCMASPACE itself, the broader theme of this work asks: “*What can one say about exponentially long proofs verified by poly-space quantum verifiers?*” For example, can allowing an exp-length proof “trivialize” a problem which is provably hard for poly-length proofs? Can exponential length proofs be encoded into *poly*-size history state⁴ constructions? Here, we give positive answers to both of these questions, for which we now set up the background.

Theme 1: Exp- versus poly-length proofs, and the solution space of constraint satisfaction problems (CSPs). In 2006, Gopalan, Kolaitis, Maneva and Papadimitriou [18] initiated the study of *reconfiguration problems* for SAT, which ask: Given two solutions x and y to a SAT formula ϕ , is there a path in the hypercube from x to y on which all intermediate vertices z are also solutions? Alternatively, in graph theoretic terms, is the solution space of ϕ *connected*? Reference [18] showed that this decision problem is PSPACE-complete, which in particular implies the problem is not *trivial* – the solution space can be either connected or disconnected, and deciding between the two is hard.

In the quantum setting, one can ask analogous questions about the “solution space” of quantum CSPs, and this has implications for the study of quantum error-correcting codes. To begin, the quantum generalization of a MAX- k -SAT instance ϕ is a *k-local Hamiltonian* H . H is a $2^n \times 2^n$ Hermitian operator acting on n qubits, but specified *succinctly* via a sum of “local clauses” H_i acting on k qubits (analogous to how ϕ is specified locally via an AND of k -local disjunctions), i.e. $H = \sum_i H_i$. The smallest eigenvalue of H , $\lambda_{\min}(H)$, is the *ground state energy* of H (for ϕ , this encodes the maximum number of simultaneously satisfiable clauses), and the corresponding space of eigenvectors the *ground space* (for ϕ , this encodes the space of optimal assignments). In 2002, Kitaev [24] gave his now celebrated “quantum Cook-Levin theorem”, which showed estimating $\lambda_{\min}(H)$, known as the k -local Hamiltonian problem (k -LH), is complete⁵ for Quantum-Merlin Arthur (QMA).

With the definition of local Hamiltonians (i.e. quantum CSPs) in hand, we can now state the quantum analogue of reconfiguration, defined as follows.

► **Definition 2** (Ground State Connectivity (GSCON) [17] (informal); see Definition 25). *Given a k -local Hamiltonian H with ground states $|\psi\rangle$ and $|\phi\rangle$ (represented succinctly via quantum circuits), and parameters m, l , does there exist a sequence of l -local unitaries U_1, \dots, U_m s.t.:*

1. ($|\psi\rangle$ mapped to $|\phi\rangle$) $U_m \cdots U_1 |\phi\rangle \approx |\psi\rangle$, and
2. (*intermediate states have low energy*) $\forall i \in [m], U_i \cdots U_1 |\psi\rangle$ has low energy relative to H ?

In words, GSCON asks whether there exists a sequence of m l -local unitaries that map $|\psi\rangle$ to $|\phi\rangle$ such that intermediate states have low energy (i.e. are also approximate “solutions”) with respect to H . Here, the use of *local* unitaries U_i is crucial, and generalizes the notion of following a path on the hypercube for SAT (which would involve flipping one bit of an assignment per step, or in quantum terms, applying a local X gate). Thus, GSCON asks: Is the ground space of H “connected”?

⁴ A history state [24] is the quantum analogue of a tableau in the Cook-Levin theorem [10, 27]).

⁵ QMA is a quantum analogue of Merlin-Arthur (MA), except with a quantum proof and quantum verifier.

Recall now that in the classical setting, the solution space of a SAT formula can be either connected or disconnected [18]. In this work, we ask the analogous fundamental question about the structure of ground spaces of local Hamiltonians:

► **Question 3.** *Can ground spaces of local Hamiltonians also be “disconnected”?*

It is known that if only *poly*-length sequences of local gates are allowed, the answer to this question is YES – namely, GSCON with a polynomial sequence of 2-local unitaries ($m = \text{poly}(n), l = 2$) and inverse polynomial spectral gap is⁶ QCMA-complete [17]. However, even in the classical case, in the worst case a connecting path in the hypercube might be *exponentially* long! (Indeed, this is what makes the PSPACE-completeness result of [18] possible.) Thus, to answer Question 3, we must allow sequences of *exponentially* many local gates, i.e. GSCON with $m = \text{exp}(n)$, denoted $\text{GSCON}_{\text{exp}}$.

In addition to this fundamental structural motivation, there are two additional reasons why Question 3 is interesting:

- First, from a complexity theory perspective, an instance of $\text{GSCON}_{\text{exp}}$ is straightforwardly in SQCMASPACE – roughly, in step i , the prover streams gate U_i to the verifier, who applies it to map its current state from $U_{i-1} \cdots U_1 |\psi\rangle$ to $U_i \cdots U_1 |\psi\rangle$. Once the proof is fully received, the verifier randomly chooses to check one of the two conditions in the GSCON definition, and accepts if the condition is met. Thus, *if* a “quantum Savitch’s” theorem were to hold, i.e. $\text{SQCMASPACE} = \text{PSPACE}$, then we would immediately obtain $\text{GSCON}_{\text{exp}} \in \text{SQCMASPACE} = \text{PSPACE}$, resolving an open question of [17].
- Second, and perhaps most interesting, is the connection to quantum error-correcting codes. For example, in a stabilizer code [20], the set of valid codewords is the ground space of a local Hamiltonian H . In this case, one desires the ground space of H to be “disconnected” in the following sense. Let $|\psi\rangle$ be a codeword of H . Then, any sufficiently short sequence of local gates (think of these as local errors “corrupting” $|\psi\rangle$) should ideally take one *out* of the ground space, so that measuring the Hamiltonian catches the corrupting process with non-negligible probability. Indeed, this is precisely what quantum codes typically achieve. What is much less obvious, however, is what happens with *exponential length* corrupting processes – by allowing an exponential-length sequences of local gates, can we stealthily map from $|\psi\rangle$ to some other codeword $|\phi\rangle$ while remaining *exponentially close* to the ground space? If so, then a single measurement of the Hamiltonian during this corrupting process is highly unlikely to detect that we are no longer in state $|\psi\rangle$!

Theme 2: Exp-length proofs, poly-size history states, and QMA(2). Our next question asks: *Can exponential length proofs be encoded into poly-size history state/circuit-to-Hamiltonian constructions?* Here, a circuit-to-Hamiltonian construction is the quantum analogue of the Cook-Levin construction [10, 27], i.e. a map from quantum circuits V to local Hamiltonians H , such that the ground space of H encodes the action of V . The basic premise is captured by Kitaev’s 5-local construction, which maps a QMA verification circuit $V = V_m \cdots V_1$ (for 1- and 2-qubit gates V_i) to a local Hamiltonian $H = H_{\text{in}} + H_{\text{prop}} + H_{\text{out}} + H_{\text{stab}}$. Intuitively, each of H_{in} , H_{prop} , and H_{out} plays a role analogous to its classical cousin in the Cook-Levin construction – H_{in} ensures V ’s computation is initialized correctly, H_{prop} that in time step t the gate V_t is applied, and H_{out} that rejecting computations are penalized. Then, the “ideal” quantum assignment perfectly satisfying H_{in} and H_{prop} is the *history state*

⁶ Quantum-Classical Merlin-Arthur (QCMA) is a quantum analogue of MA, except with a classical proof and quantum verifier.

$$|\psi_{\text{hist}}\rangle = \frac{1}{\sqrt{m+1}} \sum_{t=0}^m V_t \cdots V_1 |\psi_{\text{proof}}\rangle_A |0 \cdots 0\rangle_B |t\rangle_C \quad (1)$$

(the quantum analogue of a “tableau”), where in the context of QMA, register A starts with the quantum proof $|\psi_{\text{proof}}\rangle$, B is the ancilla space, and C is the clock keeping track of time.

Returning to the question at hand, the naive approach to encoding an exponentially long proof (given explicitly) into history state $|\psi_{\text{hist}}\rangle$ would result in an *exponential* size proof register A , which is too large for our purposes. However, in our definition of SQCMASPACE, the proof is not given explicitly, but *streamed* via application of local gates. While this may seem *a priori* more difficult to work with, it has a distinct benefit – since all gates V_t encoding streamed proof bits (i.e. “proof gates”) are “part of” the verification circuit itself, we can directly encode them into the history state’s *superposition/sum* over time steps (requiring only poly-space), thus obviating the need for a separate proof register, A ! Of course, now we are out of the frying pan into the fire, for there remains a serious problem – the propagation term $H_{\text{prop}} = \sum_{t=1}^m H_t$, which explicitly encodes each gate V_t into its corresponding local propagation term, H_t , needs to be fully specified in *advance*. However, by definition of streaming proof, the gates V_t which are proof gates are *not* known in advance. Can correct propagation still somehow be enforced? To put it more “dramatically”, can a *history* state be used to encode the *future*?

In and of itself, this seems paradoxical. Yet, there *is* a setting in which special cases of classical proofs can be “compressed” into an exponentially smaller number of qubits – QMA(2) (Definition 23). Informally, QMA(2) is defined as QMA, except where the verifier is promised to get a proof in *tensor product* across some prespecified partition L versus R of the qubits, i.e. an “unentangled” proof of form $|\psi_1\rangle_L \otimes |\psi_2\rangle_R$. In this setting, Blier and Tapp [4] first showed that the NP-complete problem 3-SAT could be verified using just *log-size* “unentangled” proofs, log-space quantum verification, and $1/\text{poly}$ promise gap, i.e. in $\text{PQMA}_{\log}(2)$. Next, Pereszlényi [31] showed a similar result for verifying the NEXP-complete language SUCCINCT-3-COLORING via poly-size unentangled proofs and $1/\text{exp}$ promise gap, i.e. in $\text{PreciseQMA}(2)$ (thus obtaining $\text{PreciseQMA}(2) = \text{NEXP}$). (Further related works in Section 1.3.) However, these constructions are expressly tailored to the problems being reduced from, and *a priori* have nothing to do with streaming. Moreover, to-date, no systematic constructions were known for embedding “long” classical proofs into “small” unentangled quantum systems. We thus ask:

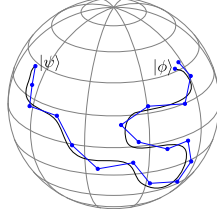
► **Question 4.** *Can unentanglement be exploited to compress streaming proofs into exponentially smaller⁷ history state constructions?*

1.1 Our Results

We divide our results into three parts: SQCMASPACE, ground space traversal, and embedding streaming proofs into unentanglement.

1. The complexity of SQCMASPACE. We first show that a quantum analogue of Savitch’s theorem for SQCMASPACE is unlikely to hold, even in the setting of *constant* promise gap.

⁷ For clarity, “smaller” refers to the number of qubits in the history state. Thus, if the proof has length $f(n)$, then the history state should be a $O(\log(f(n)))$ -qubit state.



■ **Figure 1** (Color online) Simplified illustration of the Universal quantum path following lemma with f in black (smooth), $|\psi\rangle = f(0)$, $|\phi\rangle = f(1)$, and the path of intermediate states $|\psi_t\rangle$ in blue (piece-wise linear). In the actual construction, each linear segment is itself further subdivided and likewise approximately simulated.

► **Theorem 5.** SQCMASPACE with $2^{\text{poly}(n)}$ proof bits, $\text{poly}(n)$ ancilla qubits, completeness 1, and soundness $1/2$, equals NEXP, i.e. $\text{SQCMASPACE}(\text{poly}, \text{poly}, 1) = \text{NEXP}$.

For completeness, we also define the analogous class SQMASPACE (see full version [15]), which takes an exponential length streamed *quantum* proof, and show its equality to QMA_{EXP} (quantum analogue of NEXP):

► **Theorem 6.** SQMASPACE with $2^{\text{poly}(n)}$ proof bits, $\text{poly}(n)$ ancilla qubits, completeness $2/3$, soundness $1/3$, equals QMA_{EXP} , i.e. $\text{SQMASPACE}(\text{poly}, \text{poly}, 1) = \text{QMA}_{\text{EXP}}$. With $\text{poly}(n)$ proof bits, $O(\log(n))$ ancilla bits, it equals QMA, i.e. $\text{SQMASPACE}(\log, \log, 0) = \text{QMA}$.

2. Ground space traversal. Our second result reveals that Question 3 has an arguably surprising resolution – in strong contrast to the classical case, in which the solution space for a SAT instance can be connected or disconnected, in the quantum setting, ground spaces of local Hamiltonians are *always* connected.

At the heart of this result is a new technical lemma showing how to simulate any Lipschitz continuous path on the hypersphere by an exponentially long sequence of *local* quantum gates (i.e. gates on a typical gate-based quantum computer). For this, define a *path* between an initial state $|\psi\rangle$ to final state $|\phi\rangle$ as any Lipschitz continuous function on the unit hypersphere, i.e. $f : [0, 1] \mapsto S^{d-1}$, with $f(0) = |\psi\rangle$ and $f(1) = |\phi\rangle$ (illustration in Figure 1; formal definitions in Section 3). We show⁸:

► **Lemma 7** (Universal quantum path following lemma). Set $d := 2^n$, and let $f : [0, 1] \rightarrow S^{d-1}$ be a K -Lipschitz continuous path. For every $\varepsilon > 0$, there exists a sequence of $M \in O((\frac{n^2 d^2}{\varepsilon})^{2n})$ 2-local unitaries $U = U_M \cdots U_1$ which “ ε -approximately simulates” the path f as follows: For all $t \in \{0, \dots, M\}$, it holds that $\| |\psi_t\rangle - f(t/M) \|_2 \leq \varepsilon$, where $|\psi_t\rangle = U_t \cdots U_1 |\psi_0\rangle$ and $|\psi_0\rangle := f(0)$.

With Lemma 7 in hand, we resolve Question 3 by showing that in the quantum setting, ground spaces of local Hamiltonians are always connected in the following sense.

► **Theorem 8.** Let $H \in \text{Herm}(\mathbb{C}^d)$, $d = 2^n$ with $0 \preceq H \preceq I$, $|\psi\rangle, |\phi\rangle \in \mathbb{C}^d$ with $\langle \psi | H | \psi \rangle \leq \eta$ and $\langle \phi | H | \phi \rangle \leq \eta$. For any $\Delta \geq 2^{-\text{poly}(n)}$, there exists a sequence of 2-local unitary gates $U = U_m \cdots U_1$ with $m \leq 2^{\text{poly}(n)}$ such that

⁸ For simplicity in stating the bound on M in Lemma 7, we assume $K \in \Theta(1)$, as this suffices for our applications. However, Lemma 7 also holds for non-constant K with $M \in O(K(\frac{n^2 d^2}{\varepsilon})^{2n})$ if $0 < K \leq 1$ and $M \in O(2^{O(n)}(\frac{K^2 n^2 d^2}{\varepsilon})^{2n})$ if $K > 1$.

- (1) $\|U|\psi\rangle - |\phi\rangle\|_2 \leq \Delta$, and
 (2) for all $i \in [m]$, $\langle \psi_i | H | \psi_i \rangle \leq \eta + \Delta$, where $|\psi_i\rangle := U_i \cdots U_1 |\psi\rangle$.

In words, even if we wish to remain *exponentially* close to the ground space of H throughout the local evolution from $|\psi\rangle$ to $|\phi\rangle$, this can be achieved, at the expense of exponentially blowing up the length of the evolution. Returning to our motivating example of error correcting codes, we conclude: For any H , if the ground space of H encodes a quantum error-correcting code, and $|\psi\rangle$ and $|\phi\rangle$ are any pair of code words, then Theorem 8 says one can stealthily corrupt $|\psi\rangle$ into $|\phi\rangle$ via a sequence of 2-qubit gates, so that at any point in the evolution, we are exponentially close to the code space, and thus the corruption is unlikely to be caught via measurement of H . The trade-off is that, again, this evolution path “hugging” the code space is exponentially long.

As an immediate corollary, we are now able to answer an open question of [17].

► **Corollary 9** (Informal; formally Corollary 31). *GSCON with exponentially many gates and inverse polynomial promise gap is in P.*

This follows since by Theorem 8, all GSCON instances in the parameter regime above are YES instances. Thus, allowing an exponentially long proof *trivializes* GSCON, which is otherwise QCMA-complete in the setting of poly-length proofs [17].

As a sanity check, we also strengthen a result of [17] by showing that even an *unbounded* number of 1-local gates with constant promise gap do not suffice to trivialize GSCON.

► **Theorem 10** (Informal; see full version [15] for formal statement). *GSCON is PSPACE-complete for 1-local gates, constant promise gap, and an unbounded number of gates.*

The previous PSPACE-hardness result of [17] in the 1-local case required inverse exponential promise gap and an exponential bound on the number of gates.

We also obtain two additional results regarding GSCON.

► **Theorem 11.** *GSCON is PSPACE-hard for 2-local gates, inverse exponential promise gap, and an exponential number of gates.*

► **Theorem 12.** *GSCON is NEXP-complete in the same setting as above, but with a sparse Hamiltonian whose qubits are partitioned into registers L, R , and only unitaries are considered that act either only on L or only on R .*

3. Embedding streaming proofs into unentanglement. We next resolve Question 4 in the positive, showing that streaming proofs can be systematically compressed into exponentially smaller history states.

The formalization of this goes via the Sparse Separable Hamiltonian (SSH) problem (Definition 22), which informally is identical to the k -local Hamiltonian problem, except for two key differences: (1) H is sparse, rather than local, and (2) proofs are restricted to be in tensor product form. A bit more formally: Given a sparse Hamiltonian H (Definition 21) on n qubits and bipartition L versus R of $[n]$, does there exist $|\psi_1\rangle_L \otimes |\psi_2\rangle_R$ such that

$$\langle \psi_1 |_L \otimes \langle \psi_2 |_R H | \psi_1 \rangle_L \otimes | \psi_2 \rangle_R \tag{2}$$

is “small”, or does it hold that for all $|\psi_1\rangle_L \otimes |\psi_2\rangle_R$, $\langle \psi_1 |_L \otimes \langle \psi_2 |_R H | \psi_1 \rangle_L \otimes | \psi_2 \rangle_R$ is “large”? Note that, in general, optimizations over tensor product states $|\psi_1\rangle_L \otimes |\psi_2\rangle_R \in \mathbb{C}^{d^2}$ are harder than optimizations over *all* $|\psi\rangle \in \mathbb{C}^{d^2}$, i.e. without the tensor product requirement. For example, if H in Equation (2) had *polynomial* dimension, then maximizing Equation (2) is NP-hard [21], whereas maximizing $\langle \psi | H | \psi \rangle$ over all $|\psi\rangle \in \mathbb{C}^{d^2}$ is an eigenvalue problem,

and thus efficiently solvable in the dimension of H . In other words, the optimal solution to a tensor product optimization is *not* necessarily an eigenvector of H , and this makes the design and analysis of unentangled proof systems challenging.

We now state our main technical result. A key parameter is the *promise gap* of the Sparse Separable Hamiltonian problem. Chailloux and Sattath [6] show SSH is QMA(2)-complete (Definition 23) for inverse polynomial promise gap. We show:

► **Lemma 13** ((Informal) Embedding lemma; formally Lemma 32). *Let $p, q, r, m, \alpha, \beta : \mathbb{R} \mapsto \mathbb{R}$, where p, q, r are poly-bounded. Let Q be a quantum circuit consisting of m 2-qubit gates, taking in (1) input $x \in \Sigma^n$, (2) a classical streaming proof $y \in \{0, 1\}^{2^p}$, and (3) q ancilla qubits initialized to all zeroes. We are promised that either there exists a streaming proof y causing Q to accept with probability at least α , or all streaming proofs are accepted with probability at most β , for $\alpha - \beta \geq 2^{-r}$. Then, there exists a poly-time many-one reduction from (Q, x) to a Separable Sparse Hamiltonian H instance with norm $\|H\|_\infty \in \text{poly}(m, 2^r)$, and with thresholds α', β' , such that:*

1. H acts on $O(q + \log m)$ qubits.
2. The promise gap scales as $|\alpha' - \beta'| \in \Omega\left(\frac{1}{m2^r}\right)$.

In words, any quantum verification Q with q qubits as workspace and taking in a classical proof of length 2^p can be compressed to a Separable Sparse Hamiltonian instance on $O(q + p)$ qubits and with promise gap scaling⁹ as $1/2^p$. Moreover, the mapping (1) preserves the space required up to poly overhead, and (2) embeds the proof of length 2^p bits into $\sim p$ qubits. To the best of our knowledge, this is the first such systematic method for compressing arbitrary classical proofs via unentanglement.

Applications of the Embedding Lemma. Lemma 32 immediately applies to arbitrary SQCMASPACE verifiers. Here, however, we focus on the application to MIP:

► **Corollary 14** ((Informal) Reducing MIP to unentanglement; formally, Corollary 35). *There exists a poly-time many-one reduction from any classical multi-prover interactive protocol (MIP, Definition 24) with p provers, r rounds, u space, and t bits of communication per round, to an instance of Separable Sparse Hamiltonian on $\tilde{O}(u)$ qubits with promise gap scaling dominated by scaling 2^{-tr} . (The tilde in \tilde{O} hides polylogarithmic factors.)*

For context, recall that MIP with two provers, one round and polynomially many bits of communication equals NEXP [2, 13]. As for NP, it is contained in MIP with 2 provers, 1 round, and logarithmic bits of communication. In words, Corollary 14 says that any MIP protocol can be reduced to an SSH instance, with the key parameter being the number of bits t of communication; this is what dictates the promise gap of the SSH instance H we obtain. Note we also preserve the space used by the MIP protocol (which is important for Corollary 36 for the case of NP, where the MIP uses log-space).

With Lemma 32 in hand, we next show various QMA(2)-type containments. For this, we first show that the specific Hamiltonian construction H output by the Embedding Lemma can be decided in QMA(2) using appropriate parameters:

► **Lemma 15** (Informal; see Lemma 38). *Let H be the Sparse Separable Hamiltonian instance produced by the Embedding Lemma, acting on n qubits and with promise gap g . Then, H can be decided by a QMA(2) verifier acting on $O(n)$ qubits and with promise gap $O(g)$.*

⁹ This statement assumes the verification time m , proof length 2^p , and promise gap 2^r are polynomially related, which is a reasonable setting.

As an aside, at present we are curiously unable to show Lemma 38 *without* exploiting the specific structure of H from the Embedding Lemma.

Finally, by combining Lemma 32 and Lemma 38, we obtain the following two main corollaries:

► **Corollary 16** (Informal; see Corollary 40). SQCMASPACE with proof length 2^p , q ancilla qubits, and promise gap $1/2^r$ is contained in QMA(2) with $q + \log p$ proof and ancilla qubits, and promise gap $1/2^{p+r}$.

Above, note that p and r are polynomially bounded, i.e. logarithmic p and r are allowed.

► **Corollary 17** (Informal; see Corollary 41). MIP with t bits of communication per round, space u , v random bits, p provers, r rounds, and completeness/soundness c and s , respectively, is contained in QMA(2) with $u + v + \log(tr \log(pt))$ proof and ancilla qubits, and promise gap $2^{-tr \log(pt) + \log(c-s)}$.

Thus, we obtain the first systematic QMA(2)-type bounds on arbitrary multi-prover interactive protocols. Above, the QMA(2) verifier requires the same amount of ancilla space as the MIP, and the QMA(2) promise gap depends exponentially on the total amount of communication but only polynomially on the MIP promise gap. As a bonus, we also immediately rederive (Corollary 39) in a unified fashion the results $\text{NP} = \text{PQMA}_{\log}(2)$ [4] and $\text{NEXP} = \text{PreciseQMA}(2)$ [31].

Finally, as a last application of the Embedding Lemma, we return to our study of GSCON by showing NEXP-completeness for a variant of GSCON:

► **Theorem 18** (Informal; see full version [15] for formal statement). GSCON is NEXP-complete with a sparse Hamiltonian, an inverse exponential promise gap, and an exponential number of 2-local gates which may not act across a given L versus R cut of the qubits (i.e. all intermediate states are product across the L versus R cut).

1.2 Techniques

The proof of $\text{SQCMASPACE} = \text{NEXP}$ (Theorem 5) follows easily from the PCP characterization of NEXP [2] by using quantum computation to generate random numbers and streaming the PCP proof¹⁰. As for $\text{SQMASPACE} = \text{QMA}_{\text{EXP}}$ (Theorem 6), the obstacle is to show that (weak) error reduction holds for SQMASPACE. This is because with only poly-size ancilla space, the verifier seemingly can only repeat the verification a polynomial number of times, which is not enough to amplify an exponentially small promise gap. We overcome this by forcing the streamed proof itself to repeatedly replenish the verifier’s ancilla, and run a pair of counters to both ensure the prover sends correctly initialized ancilla qubits all set to zero, along with sufficiently many “good” proofs accepted with high probability.

The main technical contributions of this work, however, are the Universal quantum path following lemma (Lemma 7) and the Embedding lemma (Lemma 32), which we now discuss.

1. Universal Quantum Path Following Lemma. Recall Lemma 7 shows how to simulate any Lipschitz continuous path on the unit hypersphere via an exponentially long sequence of local gates. To show this, we first “discretize” the given path f into a dense enough

¹⁰For clarity, there is nothing specifically “quantum” about this proof. If one defines an analogous notion of “MASPACE”, meaning a *randomized* generalization of NPSPACE, an essentially identical proof would also yield $\text{MASPACE} = \text{NEXP}$.

sequence of points $|\psi_1\rangle, \dots, |\psi_N\rangle$ so that each consecutive pair of points $|\psi_j\rangle$ and $|\psi_{j+1}\rangle$ is “close”. Thus, if *global* (i.e. n -local) unitaries were allowed, a “small rotation” (i.e. close to identity) would suffice to exactly map $|\psi_j\rangle$ to $|\psi_{j+1}\rangle$. However, here we are restricted to 2-local gates, and the typical approach [29] to simulate global rotations using 2-local gates would yield intermediate states possibly very far from $|\psi_j\rangle$ and $|\psi_{j+1}\rangle$ (and more generally, from the desired path f). Hence, we devise a general decomposition of global unitaries close to identity into 2-local gates close to identity. Specifically, we give an approximate decomposition $e^{itH} \approx \prod_j e^{it_j H_j}$, where $\sum_j |t_j|$ is bounded by $2^{\text{poly}(n)} |t|^{1/2^n}$. Basically, we can decompose a unitary with a short *pulse time* into many local unitaries with short pulse times, which allows us to map a quantum state along each segment $|\psi_j\rangle$ to $|\psi_{j+1}\rangle$.

For that, we first write $H = \sum_j \alpha_j P_j$ in the Pauli basis (i.e. each P_j is a tensor product of the Pauli matrices and identity) and apply the Suzuki decomposition [33] to get $e^{iH} = \prod_j e^{i\alpha_j P_j} + O(t^2)$, where $\sum_j |\alpha_j| \leq t$. Clinton, Bausch, and Cubitt [9] give an exact 2-local decomposition for the $e^{i\alpha_j P_j}$ terms with bounded pulse times. We provide an alternative construction with a simpler analysis, and which requires a polynomial number of gates to implement a Hamiltonian interaction (compared to exponential in [9]), at the cost of a slightly worse pulse time bound compared to [9].

In terms of application, recall GSCON asks whether there exists a sequence of local unitaries mapping ground state $|\psi\rangle$ of H to orthogonal ground state $|\phi\rangle$, while remaining in low energy space. Since we can apply Lemma 7 to *arbitrary* Lipschitz continuous paths, we can apply it to path $f(t) = \cos(t\pi/2)|\psi\rangle + \sin(t\pi/2)|\phi\rangle$, where note that for all t , $f(t)$ is also a ground state¹¹ of H . Thus, Lemma 7 allows us to “follow” this path via 2-qubit gates, yielding a suitable gate sequence $U_m \cdots U_1$ for GSCON. In general, this sequence requires an exponential number of gates, and in return achieves exponential precision.

2. The Embedding Lemma. Lemma 32 shows how to compress any quantum verification Q with q qubits as workspace and taking in a streaming classical proof of length 2^p into a Separable Sparse Hamiltonian instance on $O(q+p)$ qubits and promise gap scaling as $1/2^p$. So, let $Q = V_m \cdots V_1$ be a quantum verifier taking in streaming proof y . Recall we formalized “streaming” by partitioning the gates $\{V_i\}$ into two sets: “Proof gates” indexed by $P \subseteq [m]$, and “computation gates” indexed by $[m] \setminus P$. Our goal is to design a Hamiltonian H so that, when there exists proof y accepted by Q , then an appropriately defined history state $|\psi_{\text{hist}}\rangle$ has low energy against H . The problem is that we do not know the proof gates $\{V_i\}$ with $i \in P$ while computing the reduction – the verifier Q only learns this information in the *future*. To overcome this, at a very high level, we instead demand an appropriately defined *unentangled* history state of form $|\psi_{\text{hist}}\rangle_L \otimes |\psi_{\text{hist}}\rangle_R$. We then exploit this “unentanglement” to logically simulate a quadratic Boolean function across the two copies of $|\psi_{\text{hist}}\rangle$, in turn allowing the history state to decide “on-the-fly” whether it wishes to stream proof bit 0 or 1 in step $t \in P$.

Formally, we define our Hamiltonian as (details in Section 4) $\tilde{H} = \Delta_{\text{in}} \tilde{H}_{\text{in}} + \Delta_{\text{prop}} \tilde{H}_{\text{prop}} + \Delta_{\text{sym}} \tilde{H}_{\text{sym}} + \tilde{H}_{\text{out}}$ for some weights $\Delta_{\text{in}}, \Delta_{\text{prop}}, \Delta_{\text{sym}} \geq 0$. Briefly, \tilde{H}_{in} and \tilde{H}_{out} ensure that in any candidate proof $|\psi\rangle_L \otimes |\phi\rangle_R$, both $|\psi\rangle_L$ and $|\phi\rangle_R$ are initialized correctly at time $t = 0$ and accept at time $t = m$. \tilde{H}_{sym} enforces that a low energy state is symmetric under exchange with respect to the cut L versus R . The key ingredient, however, is hiding in \tilde{H}_{prop} , and is the FLUX gadget,

$$(H_t^I)_L \otimes (H_t^{iX})_R + (H_t^{iX})_L \otimes (H_t^I)_R, \quad (3)$$

used to encode *future* streamed proof gates (i.e. for time steps $t \in P$).

¹¹ Note Theorem 8 also applies when $|\psi\rangle$ and $|\phi\rangle$ are not ground states, but just low energy states.

This gadget works as follows. A propagation term H_t^I or H_t^{iX} enforces that at time $t \in P$, the local proof $|\psi\rangle$ applies proof gate I (to simulate streaming bit 0) or proof gate iX (to simulate streaming bit 1), respectively. Since we do not know in advance which of these two gates will be applied, we run a thought experiment – imagine we had two parallel universes, where universe L streams bit 0, or universe R streams bit 1. This can be simulated via term $(H_t^I)_L \otimes (H_t^{iX})_R$ – namely, since the tensor product is multiplicative, this constraint is satisfied, i.e. $((H_t^I)_L \otimes (H_t^{iX})_R) |\psi\rangle_L \otimes |\phi\rangle_R = H_t^I |\psi\rangle_L \otimes H_t^{iX} |\phi\rangle_R = 0$, only if either universe L (i.e. $|\psi\rangle_L$) applies gate I or universe R (i.e. $|\psi\rangle_R$) applies gate iX , or both. The keyword here is “or”, in that the tensor product allows us to simulate the Boolean OR function between universes. Of course, we have not yet achieved anything, since neither universe has any choice in which bit it streams! This brings us back to the FLUX gadget – observe that the “+” in Equation (3) acts as a Boolean “AND”. In other words, to satisfy the gadget, universe L can apply I (this annihilates the first term, $(H_t^I)_L \otimes (H_t^{iX})_R$) and R can apply I (this annihilates the second term, $(H_t^{iX})_L \otimes (H_t^I)_R$). Similarly, both can instead choose to apply iX to satisfy the gadget. The conclusion is that both universes can freely decide which proof bit to stream at time $t \in P$, so long as they stream the *same* bit! Indeed, this works because we have exploited unentanglement to simulate the quadratic Boolean function EQUALS: $(x \vee \bar{y}) \wedge (\bar{x} \vee y) \leftrightarrow x = y$ for $x, y \in \{0, 1\}$.

The next challenge is to prove soundness of the construction, where recall tensor product optimizations are difficult to analyze since optimal solutions do not correspond to eigenvectors (and thus, standard techniques from the study of k -LH cannot be directly employed). Indeed, this step is rather involved (a step-by-step derivation of the construction is in Section 4.1). For example, the careful reader might wonder why we chose iX to stream bit 1 rather than simply X – it turns out use of X breaks soundness of the FLUX gadget. Even when we use iX , without the symmetry constraint \tilde{H}_{sym} , soundness again breaks via simultaneous cheating across *multiple* FLUX gadgets.

To overcome these obstacles, very briefly, our analysis first exploits the large weight Δ_{sym} to enforce any low energy state to look like $|\psi\rangle_L \otimes |\psi\rangle_R$ for some $|\psi\rangle$. To next force $|\psi\rangle$ to look like an actual history state, two ingredients smoothly fit together. First, since we use iX instead of X in the FLUX gadget, it turns out that for any choice of assignment $|\psi_1\rangle_L$ on L , a low energy state $|\psi_2\rangle_R$ on system R must implement at time $t \in P$ the operator

$$U(a_t, b_t) = \frac{1}{\sqrt{a_t^2 + b_t^2}} (a_t iX + b_t I). \quad (4)$$

for some $a_t, b_t \geq 0$. Now, due to the i in iX , $U(a_t, b_t)$ turns out to be *unitary*. Thus, conditioned on any fixed assignment on L , we can “invert” $U(a_t, b_t)$ by applying Kitaev’s change-of-basis operator [24], thus diagonalizing what we call the “residual propagation term on R ”,

$$\langle \psi_1 | H_t^I | \psi_1 \rangle (H_t^{iX})_R + \langle \psi_1 | H_t^{iX} | \psi_1 \rangle (H_t^I)_R. \quad (5)$$

The second ingredient is to show that by setting Δ_{prop} large enough, we can extract a “proper” propagation Hamiltonian hiding under this “residual operator on R ” over *all* time steps. This allows us to force any low energy state of \tilde{H} to indeed be of form $|\psi_{\text{hist}}\rangle_L \otimes |\psi_{\text{hist}}\rangle_R$ – which is *almost* what we want.

The final problem is that for any $t \in P$, $|\psi_{\text{hist}}\rangle$ is currently forced to apply a unitary of form $U(a_t, b_t)$ from Equation (4) for some a_t, b_t . What we *actually* want is for the FLUX gadget to act like a “switch” – either $a_t = 0$ and $b_t \gg 0$ (streaming proof bit 0) or $a_t \gg 0$ and $b_t = 0$ (streaming proof bit 1). By carefully exploiting the structure of $U(a_t, b_t)$ itself, we finally show that any low energy $|\psi_{\text{hist}}\rangle_L \otimes |\psi_{\text{hist}}\rangle_R$ can be “rounded” to obtain a state close-by which perfectly satisfies this desired “switch” behavior for all $t \in P$.

1.3 Related Work

GSCON. In the classical setting, Gopalan, Kolaitis, Maneva, and Papadimitriou [18] show the problem of determining whether two solutions of a Boolean formula are connected through its solution space is in P or PSPACE-complete, depending on the types of constraints allowed in the formula. The GSCON problem was introduced by Gharibian and Sikora [17], who show that GSCON with $m = \text{poly}(n)$ ($l = 2$)-local unitaries is QCMA-complete. For $m = \exp(n)$ and $l = 1$, it is PSPACE-complete. If the Hamiltonian is given as a succinct circuit description, GSCON is NEXP-complete. Gosset, Mehta, and Vidick [19] show the surprising result that QCMA-completeness holds even for *commuting* local Hamiltonians (an analogous result for QMA-completeness of k -LH on *commuting* Hamiltonians remains an open question). Nagaj, Hangleiter, Eisert, and Schwarz [28] next show QCMA-completeness for stoquastic Hamiltonians. Watson, Bausch, and Gharibian [36] study GSCON with a *translationally invariant Hamiltonian* on a 1D chain of qudits (i.e. there exists a single 2-local Hamiltonian acting on each pair of neighbors in the chain) and prove QCMA_{EXP}-completeness (QCMA_{EXP} is QCMA with exponentially large proof and exponential-time quantum verifier). We remark that the EXP in QCMA_{EXP} arises due to the translation-invariance, which causes the encoding size of the problem to be exponentially smaller than the size of the chain.

QMA(2). The complexity class QMA(k) (QMA with k unentangled proofs) was first introduced by Kobayashi, Matsumoto, and Yamakami [26]. Blier and Tapp [4] show that $\text{NP} \subseteq \text{QMA}_{\log}(2)$ (QMA(2) but with log-sized proofs and a log-space verifier) with perfect completeness and $1 - 1/\text{poly}$ soundness. Aaronson et al. [1] give a $\text{QMA}_{\log}(\tilde{O}(\sqrt{n}))$ protocol for 3-SAT with a constant soundness gap (as opposed to $1/\text{poly}$ in [4]). They further argue that assuming a weak version of the Additivity Conjecture from quantum information theory, $\text{QMA}(k) = \text{QMA}(2)$ for all $k \geq 2$ and QMA(2) can be amplified to exponentially small error. Harrow and Montanaro [22] prove this statement by developing a protocol for a *product test* that allows a quantum verifier to check if a state is a product state across n cuts, given two copies. It also follows that 3-SAT has a QMA(2) protocol with proof size $\tilde{O}(\sqrt{n})$. We remark that it remains an open problem whether QMA(2) is equal to NEXP, though an oracle separation to coNP exists [26]. Gharibian, Santha, Sikora, Sundaram and Yirka [16] define quantum generalizations of the Polynomial Hierarchy, QCPH and QPH (using classical and quantum proofs, respectively, and quantum verifiers in both cases), and show that (1) if QCPH = QPH, then QMA(2) is in the Counting Hierarchy, and (2) unless $\text{QMA}(2) = \text{Q}\Sigma_3$ (Q Σ_3 the third level of QPH), QMA(2) is strictly contained in NEXP.

Chen and Drucker [7] improve upon [1] with a BellQMA_{log}($\tilde{O}(\sqrt{n})$) protocol for 3-SAT, where BellQMA(k) is defined as QMA(k) without entangled measurements. QMA(2) permits an inverse polynomial gap, however with an exponentially small gap it is equal to NEXP as shown by Pereszlényi [31]. With a linear number of provers and an exponential soundness gap, BellQMA equals NEXP as well. Kinoshita [23] proves that QMA(2) with postselection also equals NEXP. Chiesa and Forbes [8] give a tight soundness analysis of the protocol of [4], showing a soundness gap $\Omega(n^{-1})$, notably without using a PCP. They further improve upon [7] by providing a smooth trade-off between the number of provers k and the soundness gap $\Omega(k^2/n)$. Chailloux and Sattath [6] show the Separable Sparse Hamiltonian problem with $1/\text{poly}$ promise gap is complete for QMA(2). Sparsity is crucial here, as [6] shows the Separable *Local* Hamiltonian problem is in QMA.

Space-bounded quantum computation. Watrous [34, 35] initiates the study of space-bounded quantum computation and shows $\text{BQSPACE}(s(n)) \subseteq \text{SPACE}(O(s(n)^2))$, where BQSPACE is the space-bounded variant of BQP with intermediate measurements. It follows that $\text{BQPSPACE} = \text{PSPACE}$. Fefferman and Lin [11] prove that QMA with an inverse exponentially small gap, denoted PreciseQMA , is equal to PSPACE , by showing that $\text{BQUSPACE}(s(n))$ (like BQSPACE but with only unitary gates) equals QMA with a poly-time verifier, $O(s(n))$ space and proof size, and $2^{-O(s(n))}$ soundness gap. Consequently, the precise local Hamiltonian problem (inverse exponential precision) is PSPACE -complete. Fefferman and Remscrim [12] improve upon these results by showing $\text{BQUSPACE}(s) = \text{BQSPACE}(s) = \text{QUMASPACE}(s) = \text{QMASPACE}(s)$. (For clarity, recall QMASPACE receives a *poly*-sized *quantum* proof, whereas in this work SQCMASPACE takes an *exponential* size *classical* proof.) Notably, they are able to eliminate intermediate measurements, which is nontrivial in the space-bounded setting as deferred measurements require a fresh ancilla for each measurement.

1.4 Open questions

First, while we have given characterizations for both SQCMASPACE and SQMASPACE , our focus has primarily been on *classical* streamed proofs. Discovering further properties of *quantum* streamed proofs is thus left as a natural open question.

Next, via the Universal Quantum Path Following Lemma (Lemma 7), we showed that GSCON with exponentially many gates and inverse poly promise gap is in P (Corollary 31). However, what remains unclear is the complexity of GSCON with exponentially many gates and inverse *exponential* promise gap. Then, depending on the exact size of the gap and number of unitaries allowed, Lemma 7 does not necessarily apply, and indeed, in Theorem 11 we show that GSCON in this setting is PSPACE -hard. The only progress we are able to make here is Theorem 12, which requires a *sparse* (versus local) Hamiltonian and predefined L versus R cut across which gates may not act (whereas originally GSCON has no such restriction). Second, whereas the classical analogue of GSCON , $S, T\text{-CONN}$, satisfies a dichotomy theorem (i.e. is either in P or PSPACE -complete depending on the constraints allowed) [18], a similar result remains unknown for GSCON .

In terms of unentanglement, the Embedding Lemma (Lemma 32) recovers the result of [4] for NP with *log*-size $\text{QMA}(2)$ proofs, and in particular, also with an inverse poly promise gap. Whether this gap can be improved to *constant* while maintaining a *log*-size proof remains open. Next, can an analogue of Lemma 38 be shown *without* assuming the structure on H guaranteed by the Embedding Lemma? Recall our proof of Lemma 38 crucially leveraged the latter. Finally, the complexity of $\text{QMA}(2)$ remains frustratingly open – is $\text{QMA}(2) = \text{NEXP}$? What other natural complete problems are there for $\text{QMA}(2)$ beyond the (inverse poly-gapped) Separable Sparse Hamiltonian [6]?

1.5 Organization

Section 2 begins with definitions. Section 3 gives our first main result, the Universal Quantum Path Following Lemma. Section 4 gives our second main result, the Embedding Lemma, with applications in Section 5. Due to space constraints, we sketch our main results. We refer the reader to the full version [15] for formal proofs of all results.

2 Definitions

We begin by defining SQCMASPACE.

► **Definition 19** (Streaming classical proof). *Let U be a quantum circuit acting on an n_1 -qubit input register R_1 , n_2 -qubit ancilla register R_2 , and 1-qubit proof register R_3 , for some $n_1, n_2 > 0$. Registers R_2 and R_3 are initialized to all zeroes. At a high level, the idea is to stream a classical proof in register R_3 one bit at a time. To do so, we view the entire execution of U as a sequence of 1- and 2-qubit gates, but where certain 1-qubit gates on R_3 are a priori unknown. Formally, there are two main phases in the circuit, which repeat until the circuit completes. In iteration i :*

1. (Computation phase) A sequence of 1- and 2-qubit gates acts on registers R_1 and R_2 .
2. (Proof phase)
 - a. (Compute) Single-qubit gate $W_i \in \{I, X\}$ is applied to R_3 , for X the Pauli NOT gate.
 - b. (Copy) R_3 is classically copied into R_2 via CNOT gate (controlled from R_3 onto R_2).
 - c. (Uncompute) W_i is applied to R_3 to return R_3 to $|0\rangle$.

Remarks. We view each gate W_i as being applied dynamically by the prover, i.e. each time the computation phase ends, the prover supplies the next bit. Each computation phase may have a different gate sequence. Without loss of generality, in Step 2b we assume there is a fixed qubit in R_2 , say q , to which the content of R_3 is copied each time. If U does *not* wish to use the next proof bit, it may set q to $|+\rangle$ just before Step 2b, so that the CNOT gate of 2b acts invariantly.

► **Definition 20** (Streaming-QCMASPACE (SQCMASPACE(p, q, r))). *A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in SQCMASPACE(p, q, r) for polynomially-bounded functions p, q, r , if there exist thresholds $\alpha(n), \beta(n)$ satisfying $\alpha(n) - \beta(n) \geq 2^{-r(n)}$, and a poly(n)-time and $q(n)$ -space uniform family¹² of quantum circuits $\{Q_n\}$ with properties as follows. Q_n takes as input a string $x \in \Sigma^n$, a classical streaming proof $y \in \{0, 1\}^{2^{p(n)}}$, and $q(n)$ ancilla qubits in state $|0\rangle^{\otimes q(n)}$. We say Q_n accepts (x, y) with probability p if on input (x, y) , measuring Q_n 's dedicated output wire in the standard basis yields 1 with probability p . Then:*

- (Completeness) If $x \in A_{\text{yes}}$, $\exists y \in \{0, 1\}^{2^{p(n)}}$ s.t. Q_n accepts (x, y) with probability $\geq \alpha$.
- (Soundness) If $x \in A_{\text{no}}$, $\forall y \in \{0, 1\}^{2^{p(n)}}$, Q_n accepts (x, y) with probability $\leq \beta$.

Finally, let the input, ancilla, and proof registers be denoted R_1, R_2, R_3 respectively. To enforce that R_1 and R_3 are not used as ancilla, we require that Q_n only acts on R_1 and R_3 via CNOTs with the control in R_1 or R_3 and the target in R_2 .

We next state existing definitions.

► **Definition 21** (Sparse Hamiltonian (e.g. [6])). *A Hermitian operator $H \in \text{Herm}((\mathbb{C}^2)^{\otimes n})$ is row-sparse if each row of H has at most poly(n) non-zero entries, and if there exists an efficient classical algorithm mapping row index $i \in [2^n]$ to a sequence of all non-zero entries H_{ij} of H .*

► **Definition 22** (Separable Sparse Hamiltonian (SSH(g)) [6]). *Let $g : \mathbb{N} \mapsto \mathbb{R}$ be an efficiently computable function. Given as input a sparse Hamiltonian H , a bipartition L versus R of the n qubits H acts on, and threshold parameters α, β satisfying $\beta - \alpha \geq 1/g(n)$, decide:*

- (YES case) $\exists |\psi_1\rangle_L |\psi_2\rangle_R : \langle \psi_1 |_L \langle \psi_2 |_R H | \psi_1 \rangle_L | \psi_2 \rangle_R \leq \alpha$.
- (NO case) $\forall |\psi_1\rangle_L |\psi_2\rangle_R : \langle \psi_1 |_L \langle \psi_2 |_R H | \psi_1 \rangle_L | \psi_2 \rangle_R \geq \beta$.

¹²To clarify, there exists a Turing machine that outputs gate i on input $(i, 1^n)$ in poly(n)-time and $q(n)$ space. These bounds match those used in the definition of QU-SPACE [11, Def. 4].

Chailloux and Sattath show that the Separable Sparse Hamiltonian problem with inverse polynomial gap, SSH(1/poly), is QMA(2)-complete, for QMA(2) defined next.

► **Definition 23** (QMA(2, p, q, r) [26]). *A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QMA(2, p, q, r) for polynomially bounded functions p, q, r (may also be, e.g., logarithmic) if there exist thresholds $\alpha(n), \beta(n)$ satisfying $\alpha(n) - \beta(n) \geq 2^{-r(n)}$, and a poly-time uniform family of quantum circuits $\{Q_n\}$ with properties as follows. Q_n takes as input a string $x \in \Sigma^n$, a quantum proof $|\psi_1\rangle_L \otimes |\psi_2\rangle_R \in \mathbb{C}^{2^{p(n)}} \otimes \mathbb{C}^{2^{q(n)}}$, and $q(n)$ ancilla qubits in state $|0\rangle^{\otimes q(n)}$. We say Q_n accepts (x, y) with probability p_{acc} if on input $(x, |\psi_1\rangle_L |\psi_2\rangle_R)$, measuring Q_n 's dedicated output wire in the standard basis yields 1 with probability p_{acc} . Then:*

- (Completeness) *If $x \in A_{\text{yes}}$, $\exists |\psi_1\rangle_L |\psi_2\rangle_R$ s.t. Q_n accepts $(x, |\psi_1\rangle_L |\psi_2\rangle_R)$ w.p. $\geq \alpha$.*
- (Soundness) *If $x \in A_{\text{no}}$, $\forall |\psi_1\rangle_L |\psi_2\rangle_R$, Q_n accepts $(x, |\psi_1\rangle_L |\psi_2\rangle_R)$ w.p. $\leq \beta$.*

Harrow and Montanaro [22] have shown that error reduction holds for QMA(2), i.e. we may assume α and β are exponentially close to 1 and 0, respectively.

► **Definition 24** (MIP($t(n), u(n), v(n), p(n), r(n), c(n), s(n)$) (introduced in [3], as stated in [14])). *A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in MIP(t, p, r, c, s) if there exist polynomial t and polynomially bounded functions u and v , and a classical verifier V using $\text{poly}(n)$ time, $u(n)$ space, $v(n)$ bits of randomness, and interacting with p non-communicating provers via r rounds of interaction, where each round consists of $t(n)$ bits of communication between verifier and provers, and where $n = |x|$ is the size of input x , such that*

- *If $x \in A_{\text{yes}}$, then there exists a strategy for the provers that is accepted by the verifier with probability $\geq c$.*
- *If $x \in A_{\text{no}}$, any strategy of the provers is accepted by the verifier with probability $\leq s$.*

► **Definition 25** (GSCON($H, k, \eta_1, \eta_2, \eta_3, \eta_4, \Delta, l, m, U_\psi, U_\phi$) [17]).

1. *Input:*

- *A k -local Hamiltonian $H \in \text{Herm}(\mathcal{B}^{\otimes n})$, where $\mathcal{B} := \mathbb{C}^2$.*
- *$\eta_1, \eta_2, \eta_3, \eta_4, \Delta \in \mathbb{R}$ and integer $m \geq 0$, such that $\eta_2 - \eta_1 \geq \Delta$ and $\eta_4 - \eta_3 \geq \Delta$.*
- *Polynomial size quantum circuits U_ϕ, U_ψ generating “starting” and “target” states $|\phi\rangle$ and $|\psi\rangle$ (on input $|0^n\rangle$), respectively, satisfying $\langle \psi | H | \psi \rangle \leq \eta_1$ and $\langle \phi | H | \phi \rangle \leq \eta_1$.*

2. *Output:*

YES: *There exists a sequence of l -local unitaries U_1, \dots, U_m such that:*

- (a) *(Intermediate states remain in low energy space) For all $i \in [m]$ and intermediate states $|\psi_i\rangle := U_i \cdots U_1 |\psi\rangle$, it holds that $\langle \psi_i | H | \psi_i \rangle \leq \eta_1$, and*
- (b) *(Final state is close to target state) $\| |\psi_m\rangle - |\phi\rangle \|_2 \leq \eta_3$.*

NO: *For all l -local sequences of unitaries U_1, \dots, U_m , either:*

- (a) *(Intermediate state obtains high energy) There exists $i \in [m]$ and an intermediate state $|\psi_i\rangle$ such that $\langle \psi_i | H | \psi_i \rangle \geq \eta_2$, or*
- (b) *(Final state far from target state) $\| |\psi_m\rangle - |\phi\rangle \|_2 \geq \eta_4$.*

We assume U_ψ and U_ϕ to be given as sequences of gates from a universal gate set. The numeric parameters are specified with rational entries using $O(\text{poly}(n))$ bits of precision. Note that $|\psi\rangle$ and $|\phi\rangle$ are not necessarily required to be ground states.

3 Universal Quantum Path Following Lemma

In this section, we give a general construction for simulating any Lipschitz continuous path f on the unit hypersphere via a sequence of 2-local gates, thereby sketching the proofs of Lemma 7 and Theorem 8. Full proofs are deferred to the full version [15]. We begin by formally defining paths.

► **Definition 26** (Paths and Lipschitz continuity). *For any $d \geq 2$, consider the unit hypersphere $S^{d-1} := \{|\psi\rangle \in \mathbb{C}^d \mid \|\psi\|_2 = 1\}$. A path is any function $f : [0, 1] \rightarrow S^{d-1}$. We say f is K -Lipschitz continuous if for all $a, b \in [0, 1]$, $\|f(a) - f(b)\|_2 \leq K|a - b|$. The distance between two paths is defined by the metric $d(f, g) := \max_{t \in [0, 1]} \|f(t) - g(t)\|_2$ ($\|\cdot\|_2$ Euclidean norm).*

Suppose we are given such a path f that we wish to approximate with some precision ε_0 . First, we decompose f into a sequence of points $|\psi_1\rangle, \dots, |\psi_N\rangle$ with $|\psi_i\rangle := f((i-1)/N)$. Due to Lipschitz continuity, $\|\psi_i - \psi_{i+1}\|_2 \leq \varepsilon_1$ for a suitable ε_1 with $N = O(1/\varepsilon_1)$. Hence, the angle between $|\psi_i\rangle$ and $|\psi_{i+1}\rangle$ is $O(\varepsilon_1)$ and there exists a (global) unitary rotation from $|\psi_i\rangle$ to $|\psi_{i+1}\rangle$ of the form e^{iH} with $\|H\|_\infty = O(\varepsilon_1)$ ($\|\cdot\|_\infty$ denotes the spectral norm).

Now we have a sequence of small, but global, rotations, and we wish to simulate each such rotation by a sequence of 2-local unitaries. The problem is that standard decompositions of large unitaries (even encoding such small rotations!) into 2-local gates may use 2-local gates which themselves are large rotations, thus causing us to temporarily deviate significantly from our desired path f . To get around this, we show the following lemma, which gives decompositions into 2-local gates, each of which is *itself* a small rotation (i.e. close to identity). This lemma enables us to approximate each small rotation with a precision of $O(\varepsilon_1^2) = O(1/N^2)$. Thus, we can apply this approximation to all N rotations and still achieve a total precision of $O(1/N)$.

► **Lemma 27.** *Let $U = e^{iH}$ for Hermitian $H \in \text{Herm}(\mathbb{C}^d)$, $d = 2^n$ with $\|H\|_\infty =: \varepsilon < (\pi/16)^{2n}$. There exists an approximate decomposition $U = U_m \cdots U_1 + O(d^2\varepsilon^2)$ into $m \leq 2^{O(n)}$ 2-local unitaries, such that*

$$\sum_{j=1}^m \|I - U_j\|_\infty = O\left(n^2 d^2 \varepsilon^{1/2n}\right). \quad (6)$$

Equation (6) bounds the maximum distance an intermediate state may have in the transition from $|\psi_i\rangle$ and $|\psi_{i+1}\rangle$ (applying the gates U_1, \dots, U_m one by one), scaling with ε . As argued above, the map between two close vectors can also be viewed as a small rotation, and so the lemma below follows.

► **Lemma 28.** *Let $|\psi\rangle, |\phi\rangle \in \mathbb{C}^d$ be unit vectors with $d = 2^n$. Let $\|\psi - \phi\|_2 \leq \varepsilon < (\pi/16)^{2n}$. There exists a sequence of 2-local unitaries $U = U_m \cdots U_1$ with $m \leq 2^{O(n)}$, such that*

- (1) $\|\phi - U|\psi\rangle\|_2 = O(d^2\varepsilon^2)$, and
- (2) for all $i \in [m]$, $\|\psi - U_i \cdots U_1|\psi\rangle\|_2 = O(n^2 d^2 \varepsilon^{1/2n})$.

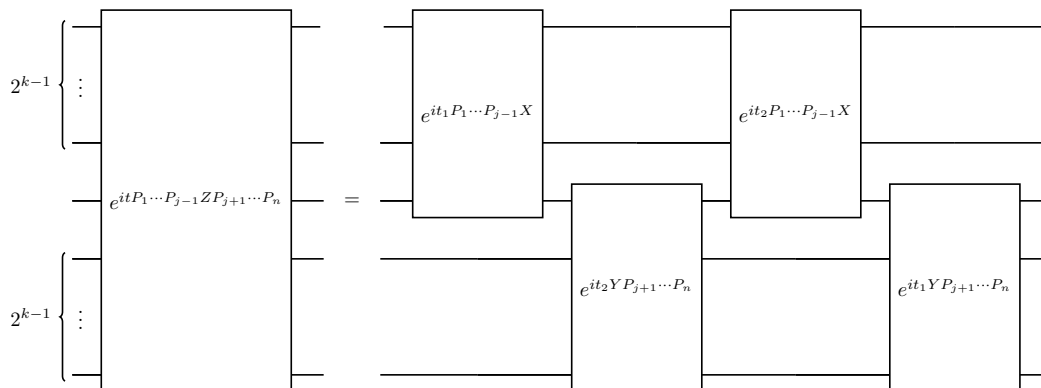
Hence, choosing ε proportional to ε_0^{2n} , we can approximate the path f with ε_0 precision by concatenating the approximations from Lemma 28 of each segment $|\psi_i\rangle, |\psi_{i+1}\rangle$, yielding Lemma 7. Next, we sketch the proof of Lemma 27.

3.1 Decomposition of Pauli Interactions

To approximate e^{iH} in Lemma 27 with 2-local unitaries, we can write $H = \sum_j \alpha_j H_j$ in the Pauli basis (i.e. each H_j is a tensor product of I, X, Y, Z) with small α_j . Applying a result due to Suzuki [33, Theorem 3], we have $e^{iH} = \prod_j e^{\alpha_j H_j} + O(\varepsilon^2)$, with $\varepsilon = \|H\|_\infty$. Next, a construction of Clinton, Bausch, and Cubitt [9] is used to decompose the $e^{\alpha_j H_j}$ terms into 2-local gates of the form $e^{it_j H_j}$, such that the total evolution time $\sum_j |t_j|$ is bounded by $O(|\alpha_j|^{1/n})$.¹³ Here, we give an alternative construction of their decomposition with a simpler analysis of pulse time bounds, and with an exponential improvement in the number of gates required. The key insight is the lemma below.

¹³Decompositions of arbitrary unitaries into 2-local gates are well known (e.g., [30]), but to the best of our knowledge, they do not provide bounds on the distance from I .

► **Lemma 29** ([9, Lemmas 7 and 9]). *Let $U = e^{itH}$ for a Hamiltonian $H = \frac{1}{2i}[h_1, h_2]$, where h_1 and h_2 anti-commute and square to identity. For $0 \leq t \leq \pi/2$, there exist $t_1, t_2 \in \mathbb{R}$ with $|t_1| + |t_2| \leq \sqrt{2}t$, and $U = e^{it_1 h_1} e^{it_2 h_2} e^{it_2 h_1} e^{it_1 h_2}$.*



■ **Figure 2** Decomposition of Pauli interactions.

Since the Pauli matrices pairwise anti-commute and square to identity, we can apply Lemma 29 to (exactly) decompose e^{itH} for $H \in \{I, X, Y, Z\}^{\otimes n}$ as depicted in Figure 2. By applying this decomposition recursively, we obtain the following result.

► **Lemma 30.** *Let $H \in \{I, X, Y, Z\}^{\otimes n}$ with $n \in (2^{k-1} + 1, 2^k + 1]$, and $t \in \mathbb{R}$ with $8|t|^{1/2^k} \leq \pi/2$. There exists a decomposition of $e^{itH} = \prod_{j=1}^m e^{it_j H_j}$, where the H_j are 2-local Pauli matrices, $m \leq 4^k = O(n^2)$, and $\sum_{i=1}^m |t_i| = O(n^2 |t|^{1/2^k}) = O(n^2 |t|^{1/2^n})$.*

To prove Lemma 27, we concatenate the gate sequences from Lemma 30 for each $e^{i\alpha_j H_j}$.

3.2 Applying Quantum Path Following to $\text{GSCON}_{\text{exp}}$

We apply the quantum path following approach to construct unitary sequences for GSCON instances. Recall, the sequences obtained from Lemma 7 have exponential length. Suppose we are given a GSCON instance, where $l = 2$, and for simplicity the starting state $|\psi\rangle$ and the target state $|\phi\rangle$ are orthogonal ground states of H (as opposed to just low energy states). To determine whether we have a YES-instance, we need to check whether there exists a sequence of 2-local unitaries that maps $|\psi\rangle$ to $|\phi\rangle$ but keeps the energy of intermediate states low. Certainly, states in the span of $|\psi\rangle$ and $|\phi\rangle$ are also ground states. Hence, we can apply Lemma 7 to the path $f(t) := \cos(t\pi/2)|\psi\rangle + \sin(t\pi/2)|\phi\rangle$ to obtain a suitable unitary sequence.

To prove Theorem 8, we also have to consider the case where $|\psi\rangle$ and $|\phi\rangle$ are *not* orthonormal ground states. The idea is to first rotate $|\psi\rangle$ onto a ground state $|\mu\rangle$ of H , and then use the same method to rotate $|\mu\rangle$ to $|\phi\rangle$. Intermediate states along this path have an energy of at most η . Let $|\lambda_1\rangle, \dots, |\lambda_d\rangle$ be an orthonormal eigenbasis of H with $|\mu\rangle = |\lambda_1\rangle$. We can write $|\psi\rangle = \cos(\theta)|\mu\rangle + \sin(\theta)|\nu\rangle$ for $|\nu\rangle \in \text{Span}\{|\lambda_2\rangle, \dots, |\lambda_d\rangle\}$. Let $f(t) := \cos((1-t)\theta)|\mu\rangle + \sin((1-t)\theta)|\nu\rangle$. Then $f(t)$ is Lipschitz continuous and $f(t)^\dagger H f(t) \leq f(0)^\dagger H f(0) \leq \eta$. Hence, Lemma 7 gives a suitable sequence for Theorem 8.

► **Corollary 31.** *GSCON with $m = 2^{\text{poly}(n)}$, $l = 2$, and subexponential δ does not have no-instances for sufficiently large n .*

4 Embedding streaming proofs into unentanglement

The main result of this section is the following.

► **Lemma 32** (Embedding lemma). *Let $p, q, r, m, \alpha, \beta : \mathbb{R} \mapsto \mathbb{R}$ be efficiently computable functions, where p, q, r are polynomially bounded. Let Q_n be a quantum circuit consisting of $m(n)$ 1- and 2-qubit gates, taking in (1) input $x \in \Sigma^n$, (2) a classical streaming proof $y \in \{0, 1\}^{2^{p(n)}}$, and (3) $q(n)$ ancilla qubits in state $|0\rangle^{\otimes q(n)}$, such that $m(n) \geq 2^{p(n)}$ and $q(n) \geq p(n)$ for all sufficiently large n . Define thresholds $\alpha(n), \beta(n)$ satisfying $\alpha(n) - \beta(n) \geq 2^{-r(n)}$. We are promised that either:*

- (YES) \exists a streaming proof $y \in \{0, 1\}^{2^{p(n)}}$ s.t. Q_n accepts (x, y) with probability at least α .
- (NO) \forall streaming proofs $y \in \{0, 1\}^{2^{p(n)}}$, Q_n accepts (x, y) with probability at most β .

There exists a poly(n)-time mapping from (Q_n, x) to a sparse Hamiltonian H on $O(q(n) + \log(m(n)))$ qubits, partition (L, R) of the qubits H acts on, and threshold parameters $\alpha'(n)$ and $\beta'(n)$ satisfying $\alpha'(n) - \beta'(n) \geq ((m(n) + 1)2^{r(n)})^{-1}$ such that:

- If (Q_n, x) is a YES case, there exists $|\psi_1\rangle_L |\psi_2\rangle_R$ such that $\langle \psi_1 | \langle \psi_2 | R H | \psi_1 \rangle_L | \psi_2 \rangle_R \leq \alpha'$.
- If (Q_n, x) is a NO case, then for all $|\psi_1\rangle_L |\psi_2\rangle_R$, $\langle \psi_1 | \langle \psi_2 | R H | \psi_1 \rangle_L | \psi_2 \rangle_R \geq \beta'$.

The norm of H scales as $\|H\|_\infty \in \text{poly}(m(n), 2^{r(n)})$.

In the remainder of this section, we give a brief overview of the main ingredients of the proof of Lemma 32. (See full version [15] for a formal proof.)

4.1 Preliminary ingredients

Let $Q_n(y) = V_m \cdots V_1$ be the quantum circuit in Lemma 32 for input size n given streaming proof y , which recall acts on registers R_1 (input of size n), R_2 (ancilla of $q(n) \in \text{poly}(n)$ qubits), R_3 (streaming classical proof, single qubit).

We first slightly adapt the definitions of history state and the Feynman-Kitaev circuit-to-Hamiltonian construction [24]. We define the history state as

$$|\psi_{\text{hist}}(y)\rangle = \frac{1}{\sqrt{m+1}} \sum_{t=0}^m V_t \cdots V_1 |0 \cdots 0\rangle_{R_2} |0\rangle_{R_3} |t\rangle_{R_4}, \quad (7)$$

where R_4 denotes the clock register. We write $|\psi_{\text{hist}}(y)\rangle$ to stress the proof y is now embedded into the circuit Q_n . Also, since $m(n) \in \Omega(2^{p(n)})$ necessarily (otherwise the circuit does not have time to see each bit of proof y), the clock register R_4 is encoded in binary.

Next, we define the Feynman-Kitaev circuit-to-Hamiltonian construction elements as

$$H_{\text{in}} := (I - |0 \cdots 0\rangle\langle 0 \cdots 0|)_{R_2} \otimes |1\rangle\langle 1|_{R_3} \otimes |0\rangle\langle 0|_{R_4} \quad (8)$$

$$H_{\text{out}} := |0\rangle\langle 0|_{\text{out}} \otimes |m\rangle\langle m|_{R_4} \quad (9)$$

$$H_{\text{prop}} := \sum_{t=1}^m H_t, \text{ where } H_t \text{ is defined as} \quad (10)$$

$$H_t := -V_t \otimes |t\rangle\langle t-1|_{R_4} - V_t^\dagger \otimes |t-1\rangle\langle t|_{R_4} + I \otimes (|t\rangle\langle t| + |t-1\rangle\langle t-1|)_{R_4}, \quad (11)$$

where in H_{out} , $|0\rangle\langle 0|_{\text{out}}$ projects onto the dedicated output wire of Q_n .

Finally, define for 1- or 2-qubit unitary U the operator H_t^U as H_t , but with V_t replaced with U . Let $P \subseteq [m]$ denote the set of time steps for which $V_i = W_i$ or $V_i = W_i^\dagger$ (corresponding to Steps 2a and 2c of Definition 19, respectively), i.e. in which a proof bit is written or uncomputed. We shall refer to such V_i as *proof gates*.

4.2 Proof sketch

We now sketch the proof of Lemma 32.

Proof. We assume all notation and definitions of Section 4.1. Define $\tilde{H} = \Delta_{\text{in}}\tilde{H}_{\text{in}} + \Delta_{\text{prop}}\tilde{H}_{\text{prop}} + \Delta_{\text{sym}}\tilde{H}_{\text{sym}} + \tilde{H}_{\text{out}}$, where

$$\tilde{H}_{\text{in}} := (H_{\text{in}})_L \otimes I_R + I_L \otimes (H_{\text{in}})_R \quad (12)$$

$$\tilde{H}_{\text{prop}} := \sum_{t=1}^m \tilde{H}_t, \quad \text{where } \tilde{H}_t \text{ is defined as} \quad (13)$$

$$\tilde{H}_t := \begin{cases} (H_t^I)_L \otimes (H_t^{iX})_R + (H_t^{iX})_L \otimes (H_t^I)_R & \text{if } t \in P \\ (H_t)_L \otimes I_R + I_L \otimes (H_t)_R & \text{if } t \notin P \end{cases} \quad (14)$$

$$\tilde{H}_{\text{out}} := (H_{\text{out}})_L \otimes I_R + I_L \otimes (H_{\text{out}})_R \quad (15)$$

$$\tilde{H}_{\text{sym}} := I - P_{LR}^{\text{sym}} \quad \text{for} \quad P_{LR}^{\text{sym}} := \frac{1}{2} \left(I_{LR} + \sum_{xy} |xy\rangle\langle yx|_{LR} \right), \quad (16)$$

and $\Delta_{\text{in}}, \Delta_{\text{prop}}, \Delta_{\text{sym}}$ are set as follows. Set $M := (m+1)2^r$. Then, define $\Delta_{\text{in}} = M^{31}$, $\Delta_{\text{prop}} = 72M^{31}$, and $\Delta_{\text{sym}} = M^{66+2k}$, where $q(n) \in O(n^k)$ for some $k \in O(1)$. Next, set $\alpha' = 2\frac{1-\alpha}{m+1}$ and $\beta' = 2\frac{1-\beta}{m+1} - \frac{1}{M}$, where recall $\alpha - \beta \geq 2^{-r}$ by assumption. Observe \tilde{H} acts on $O(q(n) + \log(m(n)))$ qubits (workspace and clock register encoded in binary, respectively). Importantly, \tilde{H} is sparse (in the sense of Definition 21). For clarity, this means our reduction does *not* output the explicit Hamiltonian \tilde{H} , but rather the classical algorithm of Definition 21 which produces entries of \tilde{H} on demand. Finally, the norm of \tilde{H} is $\|\tilde{H}\|_{\infty} \in \text{poly}(m, 2^r)$.

Correctness. Assume (Q_n, x) is a YES case. Let $Q_n = V'_m \cdots V'_2 V'_1$. For each $t \in P$ with $V'_t = X$ (i.e. a proof bit of 1 is streamed at time t), define $V_t := iX$, and for all $t \notin P$, define $V_t := V'_t$. It is straightforward to verify $(\tilde{H}_{\text{in}} + \tilde{H}_{\text{prop}} + \tilde{H}_{\text{sym}})|\psi_{\text{hist}}\rangle \otimes |\psi_{\text{hist}}\rangle = 0$, and $\langle \psi_{\text{hist}} | \otimes \langle \psi_{\text{hist}} | \tilde{H}_{\text{out}} | \psi_{\text{hist}} \rangle \otimes |\psi_{\text{hist}}\rangle \leq \frac{2(1-\alpha)}{m+1} = \alpha'$.

Assume next that (Q_n, x) is a NO case. Assume, for sake of contradiction, there exists $|\psi_1\rangle_L |\psi_2\rangle_R$ such that $\langle \psi_1 | \langle \psi_2 | \tilde{H} | \psi_1 \rangle_L | \psi_2 \rangle_R \leq \beta'$. Soundness proceeds in steps:

Step 1. Use the $\Delta_{\text{sym}}\tilde{H}_{\text{sym}}$ term to show that, up to small additive error, $|\psi_1\rangle = |\psi_2\rangle$.

Step 2. Use the $\Delta_{\text{in}}\tilde{H}_{\text{in}} + \Delta_{\text{prop}}\tilde{H}_{\text{prop}}$ term to show that $|\psi_1\rangle \approx |\psi_{\text{hist}}\rangle$. This step is rather involved, since we must “dynamically” consider a “residual” operator *conditioned* on the choice of $|\psi_1\rangle$. Formally, for any time $t \in P$, the energy penalty of proof $|\psi_1\rangle |\psi_2\rangle$ against \tilde{H}_{prop} is

$$\langle \psi_1 | \langle \psi_2 | (H_t^I \otimes H_t^{iX} + H_t^{iX} \otimes H_t^I) | \psi_1 \rangle | \psi_2 \rangle =: \langle \psi_2 | G(a_t, b_t) | \psi_2 \rangle, \quad (17)$$

for $a_t := \langle \psi_1 | H_t^I | \psi_1 \rangle$ and $b_t := \langle \psi_1 | H_t^{iX} | \psi_1 \rangle$. While the right side of this equation is minimized by the smallest eigenvalue of $G(a_t, b_t)$, the spectrum of the latter *depends on the assignment* $|\psi_1\rangle$. We address this via a series of lemmata, which culminate in the following. For this, define unitary $U(a_t, b_t) := \frac{1}{\sqrt{a_t^2 + b_t^2}}(aiX + bI)$.

► **Lemma 33.** *Assume $\langle \psi_1 | \langle \psi_1 | \Delta_{\text{prop}}\tilde{H}_{\text{prop}} | \psi_1 \rangle | \psi_1 \rangle \leq 2$. Suppose that $\delta' \geq 0$ and $\Delta_{\text{prop}} \geq 1$ satisfy $\Delta_{\text{prop}} > \max(36\sqrt{2}\delta', (8m^4)/c)$. For all $t \in P$, define F_t to be the Feynman-Kitaev propagation term (Equation (11)) for unitary $U(a_t, b_t)$. Then,*

$$2\Delta_{\text{prop}} \sum_{t \notin P} H_t + \sum_{t \in P} G(a_t, b_t) \succeq \delta' \left(\sum_{t \notin P} H_t + \sum_{t \in P} F_t \right). \quad (18)$$

This lemma allows us to relate a “residual” operator $G(a_t, b_t)$ to the standard Kitaev propagation Hamiltonian terms F_t , *albeit* for unitaries $U(a_t, b_t)$, which are *not* necessarily the correct unitary encoding streaming of a proof bit.

Step 3. Thus far, we have that for any $t \in P$, there exist scalars $a_t, b_t \geq 0$, such that $|\psi_1\rangle \approx |\psi_{\text{hist}}\rangle$ applies unitary $U(a_t, b_t)$ at time t . In the honest case, for all $t \in P$ the history state would choose $|\psi_1\rangle$ on system L so that either $a_t = 0$ and $b_t = 1$ (corresponding to streaming proof bit 0 in step t) or $a_t = 1$ and $b_t = 0$ (corresponding to streaming proof bit 1 in step t). Step 3 argues that for any low-energy $|\psi_{\text{hist}}\rangle$, this must *approximately* hold.

Finally, by combining these steps, one arrives at the desired contradiction. ◀

5 Applications of the Embedding Lemma

In this section, we apply the Embedding Lemma (Lemma 32) to obtain various corollaries. First, we reduce various complexity classes to the Separable Sparse Hamiltonian (SSH) problem. Then, we exploit the exact structure of the SSH instances from Lemma 32 to obtain various upper bounds of form $\text{QMA}(2, p, q, r)$ for appropriate p, q, r .

5.1 Reductions to Separable Sparse Hamiltonian (SSH)

The first corollary is immediate, as an SQCMASPACE circuit has $m \in \Theta(2^p)$ without loss of generality.

► **Corollary 34.** *There exists a poly-time many-one reduction from any SQCMASPACE(p, q, r) instance to an instance of Separable Sparse Hamiltonian on $O(q + \log p)$ qubits with promise gap $\Omega(2^{-p-r})$.*

The second follows by applying the standard trick of concatenating the answers of the provers for all sequences of questions from the verifier V into the streamed proof y of length $pt2^{tr}$.

► **Corollary 35.** *There exists a poly-time many-one reduction from any MIP(t, u, v, p, r, c, s) protocol to an instance of Separable Sparse Hamiltonian on $O(u + v + \log(tr \log(pt)))$ qubits with promise gap scaling as $\Omega\left([2^{tr \log(pt)}(c - s)]^{-1}\right)$.*

The next corollaries follow since $\text{NP} \subseteq \text{MIP}(\log, \log, 2, 1, 1, 1 - 1/\text{poly}(n))$, and $\text{NEXP} = \text{MIP}(\text{poly}, \text{poly}, \text{poly } 2, 1, 1, 2^{-r})$ for any desired polynomial r [2, 13], respectively.

► **Corollary 36.** *There exists a poly-time many-one reduction from any NP language to an SSH instance on $O(\log(n))$ qubits with completeness 1 and soundness $1 - 1/\text{poly}(n)$.*

► **Corollary 37.** *There exists a poly-time many-one reduction from any NEXP language to an SSH instance on $O(\text{poly}(n))$ qubits with completeness 1 and soundness $1 - 1/\exp(n)$.*

5.2 Containment in QMA(2, p, q, r)

Together with the next lemma, the above corollaries show containment of SQCMASPACE, NP, and NEXP in $\text{QMA}(2, p, q, r)$ for various appropriate p, q, r by describing a $\text{QMA}(2)$ verifier to solve the SSH instances from Lemma 32.

► **Lemma 38.** *Assume the notation of Lemma 32, and let \tilde{H} be the SSH instance produced by the latter. Then, \tilde{H} can be decided in $\text{QMA}(2, q + \log m, q + \log m, r \log m)$, i.e. with proof and ancilla space scaling as $O(q + \log m)$, and promise gap as $O(1/(2^r m))$.*

Proof sketch. We use Kitaev’s original approach for placing the k -local Hamiltonian problem in QMA [24, Proposition 14.2]: Pick a random “term” (defined shortly) of \tilde{H} and measure it against the claimed proof $|\psi\rangle = |\psi_1\rangle_L |\psi_2\rangle_R$. The catch is that the “terms” of \tilde{H} are not k -local, so a slight bit more work is required to ensure V can implement these measurements. We define “terms” of \tilde{H} as precisely the set of summands (with appropriate weights) on Equation (12) (e.g. $\Delta_{\text{in}} H_{\text{in}} \otimes I$ is a term), Equation (14) (e.g. for any $t \in P$, $\Delta_{\text{prop}} H_t^I \otimes H_t^{iX}$ and $\Delta_{\text{prop}} H_t^{iX} \otimes H_t^I$ are each terms), and Equation (15) (e.g. $I \otimes H_{\text{out}}$), as well as $\Delta_{\text{sym}} \tilde{H}_{\text{sym}}$. By construction, each term is a projector Π_i up to scaling w_i . We thus write $\tilde{H} = \sum_{i=1}^K w_i \Pi_i$ with $0 \leq w_i \leq \text{poly}(m)$.

Constructing V . Given proof $|\psi\rangle = |\psi_1\rangle_L |\psi_2\rangle_R$, V now selects index i with probability $p_i = w_i/W$ (for $W := \sum_{i=1}^K w_i$), applies measurement $(\Pi_i, I - \Pi_i)$, and rejects on outcome 0. The probability that V accepts $|\psi\rangle$ is $1 - \frac{1}{W} \langle \psi | \tilde{H} | \psi \rangle$. Therefore, V accepts with probability at least $1 - \alpha'/W$ in the YES case and at most $1 - \beta'/W$ in the NO case.

Efficiency of V . It remains to argue that V can be implemented efficiently, i.e. using $O(q + \log m)$ ancilla qubits and $\text{poly}(n)$ gates. Efficient sampling from distribution p_i can be implemented by sampling a number in $[W]$ uniformly at random. Since $W \in \text{poly}(m) \in \text{exp}(n)$, we require $O(\log(m))$ ancillas and time.

Regarding the measurement step, we first note the terms corresponding to \tilde{H}_{sym} can be implemented efficiently using the well-known SWAP test [5], which outputs 0 with probability $\langle \psi_1 | \langle \psi_2 | P^{\text{sym}} | \psi_1 \rangle | \psi_2 \rangle = (1 + |\langle \psi_1 | \psi_2 \rangle|^2)/2$. For the remaining terms of \tilde{H} , we note that up to a permutation U , which can be implemented efficiently classically, we have $U H_t U^\dagger = |0 \cdots 0\rangle \langle 0 \cdots 0|_A \otimes \left(-\frac{1}{2} V_t \otimes |1\rangle \langle 0|_B - \frac{1}{2} V_t^\dagger \otimes |0\rangle \langle 1|_B + \frac{1}{2} I \otimes (|1\rangle \langle 1| + |0\rangle \langle 0|)_B \right) = |0 \cdots 0\rangle \langle 0 \cdots 0|_A \otimes H_t'$. This measurement can now be implemented efficiently since H_t' is 3-local. \blacktriangleleft

With Lemma 38 in hand, we obtain the following corollaries, the first of which recovers the results of Blier and Tapp [4] for NP and Pereszlényi for NEXP [31]. Below, recall $\text{PQMA}_{\log}(2) = \text{QMA}(2, \log n, \log n, \log n)$, i.e. $\text{QMA}(2)$ with log-size proof and ancilla and 1/poly promise gap (technically, $\text{PQMA}_{\log}(2)$ also has perfect completeness by definition, which also matches the result we obtain below).

► **Corollary 39.** $\text{NP} = \text{PQMA}_{\log}(2)$ (cf. [4]) and $\text{NEXP} = \text{PreciseQMA}(2)$ (cf. [31]).

► **Corollary 40.** $\text{SQCMASPACE}(p, q, r) \subseteq \text{QMA}(2, q + \log p, q + \log p, p + r)$.

► **Corollary 41.** $\text{MIP}(t, u, v, p, r, c, s) \subseteq \text{QMA}(2, u + v + \log(tr \log(pt)), u + v + \log(tr \log(pt)), tr \log(pt) + \log(c - s))$.

References

- 1 Scott Aaronson, Salman Beigi, Andrew Drucker, Bill Fefferman, and Peter Shor. The Power of Unentanglement, November 2008. [arXiv:0804.0802](https://arxiv.org/abs/0804.0802).
- 2 L. Babai, L. Fortnow, and C. Lund. Nondeterministic exponential time has two-prover interactive protocols. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 16–25 vol.1, October 1990. doi:10.1109/FSCS.1990.89520.
- 3 Michael Ben-Or, Shafi Goldwasser, Joe Kilian, and Avi Wigderson. Multi-prover interactive proofs: How to remove intractability assumptions. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, pages 113–131, New York, NY, USA, 1988. Association for Computing Machinery. doi:10.1145/62212.62223.
- 4 Hugue Blier and Alain Tapp. A quantum characterization of np. *computational complexity*, 21(3):499–510, September 2012. doi:10.1007/s00037-011-0016-2.

- 5 Harry Buhrman, Richard Cleve, John Watrous, and Ronald de Wolf. Quantum fingerprinting. *Physical Review Letters*, 87(16), September 2001. doi:10.1103/physrevlett.87.167902.
- 6 André Chailloux and Or Sattath. The Complexity of the Separable Hamiltonian Problem. In *2012 IEEE 27th Conference on Computational Complexity*, pages 32–41, June 2012. ISSN: 1093-0159. doi:10.1109/CCC.2012.42.
- 7 Jing Chen and Andrew Drucker. Short multi-prover quantum proofs for sat without entangled measurements, 2010. arXiv:1011.0716.
- 8 Alessandro Chiesa and Michael A. Forbes. Improved soundness for qma with multiple provers. *Chicago Journal of Theoretical Computer Science*, 2013(1), January 2013. doi:10.4086/cjtcs.2013.001.
- 9 Laura Clinton, Johannes Bausch, and Toby Cubitt. Hamiltonian simulation algorithms for near-term quantum hardware. *Nature Communications*, 12(1), August 2021. doi:10.1038/s41467-021-25196-0.
- 10 Stephen A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, STOC '71, pages 151–158, Shaker Heights, Ohio, USA, May 1971. Association for Computing Machinery. doi:10.1145/800157.805047.
- 11 Bill Fefferman and Cedric Yen-Yu Lin. A Complete Characterization of Unitary Quantum Space. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 4:1–4:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.ITCS.2018.4.
- 12 Bill Fefferman and Zachary Remschrin. Eliminating intermediate measurements in space-bounded quantum computation. *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, June 2021. doi:10.1145/3406325.3451051.
- 13 Uriel Feige and László Lovász. Two-prover one-round proof systems: Their power and their problems (extended abstract). In *Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing (STOC)*, pages 733–744. Association for Computing Machinery, 1992. doi:10.1145/129712.129783.
- 14 J. Fitzsimons and T. Vidick. A multiprover interactive proof system for the local hamiltonian problem. In *2015 Conference on Innovations in Theoretical Computer Science (ITCS 2015)*, pages 103–112, 2015.
- 15 Sevag Gharibian and Dorian Rudolph. Quantum space, ground space traversal, and how to embed multi-prover interactive proofs into unentanglement, 2022. arXiv:2206.05243.
- 16 Sevag Gharibian, Miklos Santha, Jamie Sikora, Aarthi Sundaram, and Justin Yirka. Quantum Generalizations of the Polynomial Hierarchy with Applications to QMA(2). In *43rd International Symposium on Mathematical Foundations of Computer Science (MFCS 2018)*, volume 117, pages 58:1–58:16, 2018.
- 17 Sevag Gharibian and Jamie Sikora. Ground State Connectivity of Local Hamiltonians. *ACM Transactions on Computation Theory*, 10(2):8:1–8:28, April 2018. doi:10.1145/3186587.
- 18 Parikshit Gopalan, Phokion G. Kolaitis, Elitza N. Maneva, and Christos H. Papadimitriou. The Connectivity of Boolean Satisfiability: Computational and Structural Dichotomies. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 346–357, Berlin, Heidelberg, 2006. Springer. doi:10.1007/11786986_31.
- 19 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.
- 20 D. Gottesman. Stabilizer codes and quantum error correction. Available at arXiv.org quant-ph/9705052, 1997.
- 21 L. Gurvits. Classical deterministic complexity of Edmond’s problem and quantum entanglement. In *35th Symposium on Theory of computing (STOC 2003)*, pages 10–19. ACM Press, 2003.

- 22 Aram W. Harrow and Ashley Montanaro. Testing product states, quantum Merlin-Arthur games and tensor optimisation. *Journal of the ACM*, 60(1):1–43, February 2013. doi:10.1145/2432622.2432625.
- 23 Yusuke Kinoshita. Qma(2) with postselection equals to nexp, 2018. arXiv:1806.09732.
- 24 A. Yu. Kitaev, A. H. Shen, and M. N. Vyalyi. *Classical and Quantum Computation*. American Mathematical Society, USA, 2002.
- 25 Alexei Y. Kitaev and John 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*, STOC '00, pages 608–617, Portland, Oregon, USA, May 2000. Association for Computing Machinery. doi:10.1145/335305.335387.
- 26 Hirotada Kobayashi, Keiji Matsumoto, and Tomoyuki Yamakami. Quantum merlin-arthur proof systems: Are multiple merlins more helpful to arthur? In Toshihide Ibaraki, Naoki Katoh, and Hirotaka Ono, editors, *Algorithms and Computation*, pages 189–198, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- 27 L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973.
- 28 Daniel Nagaj, Dominik Hangleiter, Jens Eisert, and Martin Schwarz. Pinned quantum merlin-arthur: The power of fixing a few qubits in proofs. *Physical Review A*, 103(1), January 2021. doi:10.1103/physreva.103.012604.
- 29 M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.
- 30 Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th edition, 2011.
- 31 Attila Pereszlényi. Multi-prover quantum merlin-arthur proof systems with small gap, 2012. arXiv:1205.2761.
- 32 Walter J. Savitch. Relationships between nondeterministic and deterministic tape complexities. *Journal of Computer and System Sciences*, 4(2):177–192, April 1970. doi:10.1016/S0022-0000(70)80006-X.
- 33 Masuo Suzuki. Generalized Trotter’s formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems. *Communications in Mathematical Physics*, 51(2):183–190, 1976. URL: <https://projecteuclid.org/euclid.cmp/1103900351>.
- 34 John Watrous. Space-Bounded Quantum Complexity. *Journal of Computer and System Sciences*, 59(2):281–326, October 1999. doi:10.1006/jcss.1999.1655.
- 35 John Watrous. On the complexity of simulating space-bounded quantum computations. *computational complexity*, 12(1):48–84, June 2003. doi:10.1007/s00037-003-0177-8.
- 36 James D. Watson, Johannes Bausch, and Sevag Gharibian. The complexity of translationally invariant problems beyond ground state energies, 2020. arXiv:2012.12717.