An exponential time upper bound for
Quantum Merlin-Arthur games with unentangled provers

Martin Schwarz

Dahlem Center for Complex Quantum Systems
Freie Universität Berlin, 14195 Berlin, Germany

Abstract

We prove a deterministic exponential time upper bound for Quantum Merlin-Arthur games with $k$ unentangled provers. This is the first non-trivial upper bound of QMA($k$) better than NEXP and can be considered an exponential improvement, unless EXP = NEXP. The key ideas of our proof are to use perturbation theory to reduce the QMA(2)-complete SEPARABLE SPARSE HAMILTONIAN problem to a variant of the SEPARABLE LOCAL HAMILTONIAN problem with an exponentially small promise gap, and then to decide this instance using $\varepsilon$-net methods. Our results imply an exponential time algorithm for the PURE STATE N-REPRESENTABILITY problem in quantum chemistry, which is in QMA(2) but is not known to be in QMA. We also discuss the implications of our results on the BEST SEPARABLE STATE problem.
1 Introduction

Non-determinism is a fundamental concept of theoretical computer science and led to the definition of NP, kicking of modern computational complexity theory in the 1970’s [Coo71, Lev73]. Another powerful concept is interaction, where the prover/verifier interpretation of NP has been generalized to include randomness (MA, Merlin–Arthur games) [Bab85], multiple rounds of interaction [IP] [GMR85], and multiple provers [MIP] [BOGKW88]. The inherent power of these concepts became only manifest when IP and MIP were related to well-known complexity classes which are considered much more powerful than NP due to the seminal results that \( \text{MIP} = \text{NEXP} \) [BFL91] and \( \text{IP} = \text{PSPACE} \) [Sha92].

In the 1990’s, the formal foundations of quantum complexity theory have been laid [BV93] and analogous questions about the power of non-determinism (QMA) [Wat00, KSV02] and interaction with one or multiple provers (QIP, QMIP, and variants thereof) [Wat99, KM02] have been asked in a quantum context. In this setting, the proof becomes a quantum state and the verifier is a quantum computer. Some of these questions have been answered (QMA \( \subseteq \text{PP} \), QIP = PSPACE) [MW05, JIUW10]. But in addition to that, quantum theory turned out to offer new exciting possibilities, which have no classical counterpart!

In 2001, Kobayashi, Matsumoto, and Yamakami [KMY01, KMY03] first noticed the potential computational power that might be harnessed in Quantum Merlin–Arthur games from the promise of multiple unentangled quantum proofs, a concept which only makes sense in the quantum setting. This promise already hints at the very close relation to the problem of entanglement detection in quantum physics! [Bei10]

The resulting complexity class for \( k \) unentangled provers is called QMA(\( k \)), which was later shown to equal QMA(2) [HM13]. Liu, Christandl, and Verstraete [LCV07] showed that an important, natural problem in quantum chemistry, the pure state \( \mathcal{N} \)-representability problem, is in QMA(\( k \)) yet not obviously in QMA. Chailloux and Sattath [CS12] showed that the separable sparse Hamiltonian problem is QMA(2)-complete. Blier and Tapp [BT09] provided an example for the power of this class even if restricted to tiny proof states: they showed that NP is contained in a QMA\(_{\log}(2)\). In this setting, Merlin receives just two logarithmically sized quantum witness states relative to the input size, an exponential compression of the proof size compared to the classical case!

Aaronson et. al. [ABD+08] studied The Power of Unentanglement and raised the question whether the containment of NP in QMA\(_{\log}(2)\) might be “scaled up exponentially”, such that NEXP would be contained in QMA(2) with polynomially sized quantum proofs in turn. One obstacle to reach such a conclusion is the vanishing promise gap in known reductions of NP to QMA\(_{\log}(2)\), whereas a constant gap and \( O(\log(n)) \)-sized proofs would imply QMA(2) = NEXP. Indeed, [Per12] has shown that QMA(2) with exponentially small promise gap is indeed equal to NEXP. This question lead to two complementary lines of research: on the one hand, several researchers [Bei10, CF13, LGNN11] worked on improving the promise gap while maintaining a logarithmic witness size, whereas other groups started from the requirement of constant error and showed that witness sizes of \( \tilde{O}(\sqrt{n}) \) suffice to put NP into QMA\(_{\tilde{O}(\sqrt{n})}(2)\) with constant promise gap [ABD+08, HM13]. In certain restricted settings, a PSPACE upper bound for QMA(\( k \)) has been shown by [SW15]. Nevertheless, no non-trivial upper bound for the general class QMA(\( k \)) other than the trivial NEXP upper bound has been found so far.

We answer the QMA(\( k \)) \( \not\subseteq \) NEXP question in the negative (unless \( \text{EXP} = \text{NEXP} \)) by showing:

**Theorem 1.** QMA(\( k \)) \( \not\subseteq \) EXP

**Techniques** The key tool we use in our proofs is the application of matrix perturbation theory to a QMA(2)-complete separable sparse Hamiltonian in order to reduce the locality of its globally acting sparse terms while accepting the exponential cost in operator norm.

Perturbation theory has been introduced into quantum complexity theory before by the seminal works of [KKR06, OT08, BDLT08] and we refer to these works for a detailed introduction into this technique.
Its main application so far has been to reduce local Hamiltonian terms with high but constant locality to lower constant locality (e.g. 5-local to 2-local). One reason for this is the well-known fact, that these gadget constructions induce large operator norms that scale exponentially with the locality of the input terms. Thus, at most $O(\log(n))$-local terms can be reduced to constant locality while simultaneously maintaining polynomial scaling of the operator norm with the system size.

To our knowledge, this is the first work that explores the application of perturbative gadgets on globally acting Hamiltonian terms in the context of quantum complexity theory while accepting the exponential cost in operator norm. The resulting operators may be deemed unphysical, yet we can afford to work with them as our ultimate goal is a classical algorithm and not a physical Hamiltonian. The resulting separable local Hamiltonian instances with large norm or, equivalently, small promise gap, can then be directly solved using $\varepsilon$-net methods [SW15] in exponential time.

**Overview of the proof** To show Theorem 1, we start from a generic QMA$(k)$ verifier circuit. The proof then proceeds in four steps: first, we amplify the soundness and completeness bounds of the given QMA$(k)$ verifier using the amplification method of Harrow and Montanaro (Theorem 8) to the levels required for the reduction to separable sparse Hamiltonian by Chailloux and Sattath (Lemma 9). Second, we apply the Chailloux-Sattath construction yielding a separable sparse Hamiltonian instance that contains $k$ non-local but sparse Hamiltonian terms. Third, we apply our main technical lemma (Lemma 14) yielding a separable local Hamiltonian instance with exponentially small promise gap and polynomially increased system size. Fourth, we apply the $\varepsilon$-net methods of Shi and Wu (Corollary 6) to decide the instance in exponential time.

**Structure of the paper** In Section 1, we have motivated the study of QMA$(k)$ and its relation to other problems, reviewed related work, presented our key results and techniques, and gave an overview of the proof of our main theorem. In Section 2, we introduce all technical definitions and earlier results that we will use in Section 3 to prove our claims. In Section 4 we discuss the implications of our results on the best separable state problem, and finally we conclude in Section 5.

## 2 Preliminaries and definitions

Throughout the paper we use logarithms of base 2 and write $\tilde{O}(n) = O(n \text{ poly } \log(n))$. We say a pure state $|\psi\rangle$ is a product state, if it can be written as $|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$. More generally, a mixed state (or a general operator) $\rho$ is called separable, if it can be written as $\rho = \sum_i p_i \rho_i^A \otimes \rho_i^B$ with $p_i \geq 0$ and $\sum_i p_i = 1$.

**Definition 2 (QMA$(k)$).** A language $L$ is in QMA$_s(k)$, if there exists a polynomial-time quantum algorithm $A$ such that, for all inputs $x \in \{0,1\}^n$:

1. **Completeness:** If $x \in L$, there exist $k$ witnesses $|\psi_1\rangle, \ldots, |\psi_k\rangle$, each a state of $\ell$ qubits, such that $A$ outputs ACCEPT with probability at least $c$ on input $|x\rangle |\psi_1\rangle, \ldots, |\psi_k\rangle$.

2. **Soundness:** If $x \notin L$, then $A$ outputs ACCEPT with probability at most $s$ on input $|x\rangle |\psi_1\rangle, \ldots, |\psi_k\rangle$, for all states $|\psi_1\rangle, \ldots, |\psi_k\rangle$.

We use QMA$(k)$ as shorthand for QMA$_{\text{poly}(n)}(k)_{\frac{1}{4} \cdot \frac{2}{k}}$, and QMA as shorthand for QMA(1). We always assume $1 \leq k \leq \text{poly}(n)$. We also use the notation QMA$_s^{\text{SEP}}(k)$ where SEP indicates that $A$’s measurement $\{M, 1 - M\}$ is restricted to the set of separable operators $M = \sum_i \alpha_i \otimes \beta_i$ for some positive semidefinite matrices $\alpha_i$, $\beta_i$. 

3
Lemma 7 ([HM13, Lemma 6]). Harrow and Montanaro [HM13] prove the following results: the quantity\, \|H_i\|_{\text{op}} \leq w \) for each \( i \) as input. Assuming \( k, l = O(1) \), the quantity
\[
\text{OptSep}(Q) = \min \langle Q, \rho \rangle \text{ subject to } X \in \text{SepD}(A_1 \otimes \cdots \otimes A_k)
\]
where \( \text{SepD}(A_1 \otimes \cdots \otimes A_k) \) is the set of separable density operators over the space \( A_1 \otimes \cdots \otimes A_k \), can be approximated to precision \( \delta \) in
\[
\text{DTIME} \left( \exp \left( O \left( \log^2(d) \log \log(d) + \log(w/\delta) \right) \right) \times \text{poly}(d, w, 1/\delta) \right),
\]
which is quasi-polynomial in \( d, w, 1/\delta \). If \( n \) is considered as the input size and \( w/\delta = O(\text{poly}(n)) \), then \( \text{OptSep}(Q) \) can be approximated to precision \( \delta \) in \( \text{PSPACE} \).

Harrow and Montanaro [HM13] prove the following results:

Lemma 7 ([HM13, Lemma 6]). For any \( m, k, 0 \leq s < c \leq 1 \),
\[
\text{QMA}_m(k, s, c) \subseteq \text{QMA}^\text{SEP}_m(2s', c'),
\]
where \( c' = \frac{1 + c}{2} \) and \( s' = 1 - \left( \frac{1 - s}{100} \right)^2 \).

Theorem 8 ([HM13, Theorem 9]).

1. If \( s \leq 1 - 1/q(n), k = \text{poly}(n) \) and \( p(n), q(n) \) be arbitrary polynomials, then\(^2\)
\[
\text{QMA}_k(k, s, 1) = \text{QMA}^\text{SEP}_m(\text{poly}(n)q^2(n))(2\text{exp}(\text{poly}(n)), 1).
\]

\(^2\)We have explicitly included the asymptotic scaling of the proof sizes, which are implicit in the proof of the theorem in [HM13] but were not included in the original statement of the theorem.
2. If \( c - s \geq 1/q(n), c < 1, k = \text{poly}(n) \) and \( p(n), q(n) \) be arbitrary polynomials, then
\[
\text{QMA}_t(k)_{s,c} = \text{QMA}^{\text{SEP}}_{O(1/p^2(n)q^2(n))}(2)_{\exp(-p(n)\lambda),1-\exp(-p(n))}.
\] (4)

Lemma 9 (Separable Sparse Hamiltonian is \( \text{QMA}(2) \)-hard [CS12]). Let \( U = U_T U_{T-1} \ldots U_0 \) be the verifier circuit of a language \( L \in \text{QMA}^\text{SEP}(2) \) and \( \epsilon \) for some constant \( C \) with input \( x \), proof size \( \ell \), and \( \alpha \) ancillas. W.l.o.g., assume that \( U \) has been produced by Lemma 7 or Theorem 8. Then there exists a separable sparse Hamiltonian \( H_{\text{SSH}} \) that is a sum of \( O(T) \) sparse terms acting on at most \( 2\ell + O(T) + \alpha \) qubits, such that

1. if \( x \in L \) then there exists a \( |\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \), such that
\[
\langle \psi|H|\psi\rangle \leq \frac{C}{512(7+1)^s}.
\]
2. if \( x \notin L \) then for all \( |\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \), \( \langle \psi|H|\psi\rangle \geq \frac{C}{256(7+1)^s}.
\]

Theorem 10 (Eigenvalue-Approximating Gadget Theorem [OT08, Theorem 7]). Let \( ||V|| \leq \Delta/2 \) where \( \Delta \) is the spectral gap of \( H \) and \( \lambda(H) = 0 \). Let \( \tilde{H} |\Delta/2\rangle \) be the restriction of \( H = H + V \) to the space of eigenstates with eigenvalues less than \( \Delta/2 \). Let there be an effective Hamiltonian \( H_{\text{eff}} \) with \( \text{Spec}(H_{\text{eff}}) \subseteq [a, b] \). If the self-energy \( \Sigma(z) \) for all \( z \in [a - \epsilon, b + \epsilon] \) where \( a < b < \Delta/2 - \epsilon \) for some \( \epsilon > 0 \), has the property that
\[
||\Sigma(z) - H_{\text{eff}}|| \leq \epsilon,
\]
then each eigenvalue \( \tilde{\lambda}_j \) of \( \tilde{H} |\Delta/2\rangle \) is \( \epsilon \)-close to the \( j \)th eigenvalue of \( H_{\text{eff}} \). In particular
\[
|\lambda(H_{\text{eff}}) - \lambda(\tilde{H})| \leq \epsilon.
\]

Theorem 11 (Norm-Approximating Gadget Theorem [OT08, Theorem A.1]). Given is a Hamiltonian \( H \) such that no eigenvalues of \( H \) lie between \( \lambda_- = \lambda_+ - \Delta/2 \) and \( \lambda_+ = \lambda_- + \Delta/2 \). Let \( \tilde{H} = H + V \) where \( ||V|| \leq \Delta/2 \). Let there be an effective Hamiltonian \( H_{\text{eff}} \) with \( \text{Spec}(H_{\text{eff}}) \subseteq [a, b], a < b \). We assume that \( H_{\text{eff}} = \Pi_- H_{\text{eff}} \Pi_- \). Let \( D_\Delta \) be a disk of radius \( \epsilon \) in the complex plane centered around \( z_0 = \frac{b + a}{2} \). Let \( \epsilon \) be such that \( b + \epsilon < z_0 < r < \lambda_- \). Let \( w_{\text{eff}} = \sqrt{\frac{b - a}{2}} \). Assume that for all \( z \in D_\Delta \) we have
\[
||\Sigma(z) - H_{\text{eff}}|| \leq \epsilon.
\]

Then
\[
||\tilde{H}_{\lambda_-} - H_{\text{eff}}|| \leq \frac{3(||H_{\text{eff}}|| + \epsilon)||V||}{\lambda_- - ||H_{\text{eff}}|| - \epsilon} + \frac{r(\epsilon - \epsilon)}{(r - w_{\text{eff}})(r - w_{\text{eff}} - \epsilon)}.
\] (8)

Lemma 12 (Norm-approximating Parallel Subdivision Gadget [OT08]). Let \( H_{\text{target}} = H_{\text{else}} + \sum_{i=1}^k A_i \otimes B_i \) be an \( \ell \)-local Hamiltonian, where \( A_i \) and \( B_i \) are \( k \)-many pairs of \( (\ell/2) \)-local terms and \( H_{\text{else}} \) contains all terms that shall not be generated perturbatively. Let \( H_{\text{eff}} = H_{\text{target}} \otimes |0,..0\rangle\langle0,..0| \) be an effective Hamiltonian acting on a larger system extended by \( O(k) \) ancillas, i.e. \( H_{\text{target}} \) acting on the subspace where the ancillas are in their ground state. Let \( \Delta = \text{poly}(n,k)/\epsilon^2 \) for a sufficiently large \( \text{poly}(n,k) \). Then for any \( \epsilon \) there exists a \((\lceil \ell/2 \rceil + 1)\)-local Hamiltonian \( \tilde{H} \) with \( ||\tilde{H}|| = \Delta \) such that
\[
||\tilde{H}|\Delta/2 - H \otimes |0,..0\rangle\langle0,..0|| \leq \epsilon
\] (9)
where \( \tilde{H}|\Delta/2 \) indicates the restriction of \( \tilde{H} \) to the space of eigenvectors with eigenvalues less than \( \Delta/2 \).

Proof sketch. Lemma 12 is implicit in [OT08, Appendix A]: Consider the parallel subdivision gadget construction of [OT08, Section 3.1]. In the construction, Theorem 10 is applied to construct an operator \( \tilde{H} \) such that each eigenvalue \( \tilde{\lambda}_j \) of \( \tilde{H}|\Delta/2 \) is \( \epsilon \)-close to the \( j \)th eigenvalue of \( H_{\text{eff}} \). It is straightforward to check that the same assumptions (up to a larger polynomial in the choice of \( \Delta \)) suffice to apply Theorem 11 instead of Theorem 10 to the constructed gadget Hamiltonian, yielding the norm approximation
\[
||\tilde{H}|\Delta/2 - H \otimes |0,..0\rangle\langle0,..0|| \leq \epsilon
\] instead of an eigenvalue approximation. \(\square\)
3 Proof

We will now proceed to prove Theorem 1 by showing the following slightly more general result.

**Theorem 13** (main result). QMA(2)_{s,c} with c - s > 1/q, q = q(n), is decidable in

\[ \text{DTIME}(\exp(O(\text{poly}(k, \ell, q, \alpha, T, \log(n)))))) \subseteq \text{EXP}, \]  

where \( T \) is a bound on the size of the QMA(k) verifier circuit and \( \alpha \) a bound on the number of ancillas used.

We note that our upper bound is consistent with previously known hardness results, such as NEXP \( \subseteq \) QMA(2)_{s,c} with \( c - s \geq 2^{-O(\text{poly}(n))} \) [Per12] and \( \text{NP} \subseteq \text{QMA}(\log(2))_{s,c} \) with \( 1/(c - s) \leq O(\text{poly}(n)) \) [BT09, BT12], as well as the general lower bounds of [HM13, AIM14] relative to the Exponential-Time Hypothesis (ETH) of [IP99].

**Proof.** The proof proceeds in four steps: first, we amplify the soundness and completeness bounds of the given QMA(k) verifier using Theorem 8 to the levels required by Lemma 9. Second, we apply Lemma 9 yielding a separable sparse Hamiltonian instance. Third, this serves as input to Lemma 14 which yields a separable local Hamiltonian instance with exponentially small promise gap. Fourth, we apply Corollary 6 and decide the instance in EXP. Let us now discuss these steps in detail.

**Step one (QMA(k) amplification).** We apply Theorem 8 to the QMA(2)_{s,c} verifier circuit, yielding a QMA^{SEP}(2)_{\ell - p(n), 1 - e - p(n)} verifier circuit, where the proof size has been expanded to \( \ell' = \tilde{O}(k\ell p^2(n)q^2(n)) \). Choosing \( p(n) = 10\log(T) + D \) for a sufficiently large constant \( D \) satisfies the bounds \( s > \frac{1}{T+1} \) and \( c < \frac{1}{512(T+1)^{12}} \) required by Lemma 9. Thus we have \( \ell' = \tilde{O}(k\ell q^2(n)\log^2(T)) \).

**Step two (reduction to separable sparse Hamiltonian).** We apply Lemma 9 to the QMA^{SEP}(2)_{s,c} instance of step one, with \( c = \frac{1}{T+1} \), \( s = \frac{1}{512(T+1)^{12}} \), \( \ell' = \tilde{O}(k\ell log^2(T)q^2(n)) \), yielding a separable local Hamiltonian \( H_{SSH} \) with energy thresholds \( a \leq \frac{5C}{2048(T+1)^5} \) and \( b \geq \frac{7C}{2048(T+1)^5} \) in the YES and NO cases, respectively, where \( H_{SSH} \) acts on \( w = O(2\ell' + \alpha + T) = \tilde{O}(k\ell\log^2(T)q^2(n) + \alpha + T) \) qubits, where \( \alpha \) is the number of ancilla bits used by the original verifier.

**Step three (reduction to separable local Hamiltonian).** We apply our main technical Lemma 14 to \( H_{SSH} \) produced in step two. This yields a separable local Hamiltonian instance \( H_{SLH} \) with ground energies \( a \leq \frac{5C}{2048(T+1)^5} \) and \( b \geq \frac{7C}{2048(T+1)^5} \) in the YES and NO cases, respectively, where \( H_{SLH} \) acts on \( w' = O(2k\ell' + \alpha + T) = \tilde{O}(k\ell\log^2(T)q^2(n) + \alpha + T) \) qubits.

**Step four (enumeration of \( \varepsilon \)-net).** Finally, we apply Corollary 6 to \( H_{SLH} \) to approximate the ground energy of \( H_{SLH} \) over the set of separable states to precision

\[
\delta = \frac{C}{2048(T+1)^5} 2^{-O(\text{poly}(\ell')) \log(nkT))} \]

\[
\approx 2^{-O(\text{poly}(\ell') \log(nkT))}
\]

which suffices to decide the \( H_{SLH} \) instance. Since \( d = 2^{w'} \), this requires at most

\[
\text{DTIME}(\exp(O(\log^2(2^{w'})(\log log(2^{w'}) + \log(1/\delta)))) \times \text{poly}(2^{w'}, 1/\delta))
\]
Expanding all parameters we find an upper bound of
\[
\text{DTIME}(\exp(O(\text{poly}(k, \ell, q, T, \log(n)))))
\] (14)
Clearly, for QMA(\(k\)) with \(k, \ell, q, \alpha, T \in \text{O}(\text{poly}(n))\) we have
\[
\text{QMA}(k) \subseteq \text{DTIME}(\exp(O(\text{poly}(n)))) \subseteq \text{EXP}.
\] (15)
Note, that the condition in Corollary 6 implying containment in PSPACE is not satisfied in our setting. □

It remains to prove our main technical lemma.

**Lemma 14** (Separable Local Hamiltonian with exponentially small promise gap is QMA(2)-hard). Let \(H_{SSH}\) be a separable sparse Hamiltonian instance produced by Lemma 9. Then there exists a separable local Hamiltonian instance \(H_{SLH}\) such that

1. if \(x \in L\) then there exists a \(|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle\), such that \(\langle \psi | H | \psi \rangle \leq \frac{5C}{2048(T+1)^4} 2^{-O(\text{poly}(\ell) \log(nkT))}\)

2. if \(x \notin L\) then for all \(|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle\), \(\langle \psi | H | \psi \rangle \geq \frac{7C}{2048(T+1)^4} 2^{-O(\text{poly}(\ell) \log(nkT))}\)

\(H_{SLH}\) acts on an enlarged system of \(O(2k\ell + T + \alpha)\) qubits.

**Proof.** Note that \(H_{SSH}\) is a Feynman-Kitaev Hamiltonian [KSV02, section 14.4.1] of the standard form
\[H_{SSH} = H_{\text{in}} + H_{\text{prop}} + H_{\text{out}} + H_{\text{clock}}\] (16)
where \(H_{\text{prop}} = \sum_{t=1}^{T} H_t\). Since by the assumptions of Lemma 9 the verifier circuit encoded in \(H_{SSH}\) has been brought into a standard form by the Harrow-Montanaro construction (i.e. Lemma 7 or Theorem 8), we know that \(H_{SSH}\) contains exactly \(k\) non-local terms (one for each prover) of the form
\[H_t = -\frac{1}{2} |t\rangle \langle t - 1| \otimes U_t - \frac{1}{2} |t - 1\rangle \langle t| \otimes U_t^\dagger + \frac{1}{2} (|t\rangle \langle t| + |t + 1\rangle \langle t + 1|) \otimes 1\] (17)
These encode a simultaneous controlled-swap operation \(U_t = \text{CSWAP}\) on \(2\ell\) qubits. All other terms are 5-local. As noticed by Chailloux and Sattath [CS12], it’s necessary to perform this controlled-swap operation “in one time step” in the history state that constitutes the ground state of the Feynman-Kitaev Hamiltonian. This is required in order to ensure the separability of the ground state of \(H_{SSH}\) for satisfiable instances.

From this starting point, we show the lemma by reducing SEPARABLE SPARSE HAMILTONIAN to SEPARABLE LOCAL HAMILTONIAN with exponentially small promise gap. Our approach to deal with the non-local terms is to approximate them by local ones perturbatively using the parallel subdivision gadget of [OT08] as summarized in Lemma 12. Let us extract one representative non-local term \(H_{\text{CSWAP}}\) from \(H_t\). The lemma is later applied to all \(k\) of these terms in parallel. Note that the CSWAP operation across \(2\ell\) qubits exhibits a natural tensor product structure which is a crucial prerequisite to apply Lemma 12 iteratively. Using \(\text{CSWAP} = \text{CSWAP}^\dagger\) we write
\[H_{\text{CSWAP}} = \text{SWAP}_{1,\ell+1} \otimes \text{SWAP}_{2,\ell+2} \otimes \cdots \otimes \text{SWAP}_{\ell,2\ell} \otimes |1\rangle \langle 1| \otimes (|t\rangle \langle t + 1| + |t + 1\rangle \langle t|)\] (18)
We apply Lemma 12 to \(H_{SSH}\) iteratively \(O(\log(\ell))\) times in order to break down \(H_{\text{CSWAP}}\) into ultimately \(O(1)\)-local terms along its natural tensor product structure as illustrated in Figure 1. For one application of Lemma 12 a choice of
\[\Delta_1 = \frac{\text{poly}(n, k)}{\varepsilon^2}\] (19)
Figure 1: [CBBK15] Reduction tree diagram for parallel subdivision gadget acting on a $H_{\text{CSWAP}}$ term as defined in eq. (18) for the case $\ell = 5$. In our case, each $S_i$ is a two-qubit SWAP operator acting on qubits $(i, \ell + i)$, $C$ is the control operator, whereas $T$ is the time propagation operator. The vertical lines $|$ show where the subdivisions are made at each iteration to each term. The $X_{u_i}$ terms indicate the coupling to mediator qubit $u_i$ introduced in this step. Clearly, the number of mediator qubits scales as $O(\ell)$ and $O(\log(\ell))$ iterations suffice to arrive at $O(1)$-local terms.

Thus, after $O(\log(\ell))$ iterations, we find

$$\Delta_{i+1} = \frac{\text{poly}(n,k)}{\varepsilon^2} \Delta_i^3$$

(20)

It suffices. Iterating the gadget increases the required interaction strength by a polynomial factor [BDLT08].

$$\Delta \leq \left( \frac{\text{poly}(n,k)}{\varepsilon^2} \right)^{3\log(\ell)} \leq \left( \frac{\text{poly}(n,k)}{\varepsilon^2} \right)^{\text{poly}(\ell)} \leq 2^{O(\text{poly}(\ell) \log(nk/\varepsilon))}$$

(21)

To finish the proof of the lemma, it remains to show that the claimed completeness and soundness bounds are satisfied by $H_{\text{SLH}}$. Let us first verify completeness and soundness of $\tilde{H}_{\text{SSH}}$ explicitly which will in turn imply the exponentially rescaled bounds for $H_{\text{SLH}}$.

For the completeness bound we show that the separable witness state implied by Lemma 9, i.e. $|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$ with energy $\langle \psi | H_{\text{SSH}} | \psi \rangle \leq a$, implies that the separable state $|\psi\rangle | 0 \cdots 0 \rangle$ of the

$$H_{\text{SLH}} = \Delta^{-1} \tilde{H}_{\text{SSH}}$$

(24)
extended system satisfies \( \langle \psi | 0 \cdots 0 | \tilde{H}_{\text{SSH}} | \psi \rangle 0 \cdots 0 \leq \frac{5C}{2048(T+1)^5} \). Omitting the explicit tensoring of \( |0 \cdots 0 \rangle \) ancillas for readability, we have

\[
\langle \psi | \tilde{H}_{\text{SSH}} | \psi \rangle = \langle \psi | H_{\text{SSH}} | \psi \rangle + \langle \psi | (\tilde{H}_{\text{SSH}} - H_{\text{SSH}}) | \psi \rangle 
\leq \frac{C}{512(T+1)^5} + \left\| \tilde{H}_{\text{SSH}} - H_{\text{SSH}} \right\| \tag{25}
\]

\[
\leq \frac{4C}{2048(T+1)^5} + \frac{C}{2048(T+1)^5} \tag{26}
\]

\[
\leq \frac{5C}{2048(T+1)^5} \tag{27}
\]

where the first inequality follows from the assumptions of Lemma 9 and uses basic properties of the spectral norm, while the second inequality follows from eq. (23). Note that essentially the same separable state \(|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle\) (once extended to the larger space with \(|0 \cdots 0\rangle\) ancillas) satisfies the respective completeness bounds in both, the perturbed as well as the unperturbed setting.

Similarly, for the soundness bound, we know that for all states \(|\psi\rangle\)

\[
\langle \psi | H_{\text{SSH}} | \psi \rangle = \langle \psi | \tilde{H}_{\text{SSH}} | \psi \rangle + \langle \psi | (H_{\text{SSH}} - \tilde{H}_{\text{SSH}}) | \psi \rangle 
\leq \langle \psi | \tilde{H}_{\text{SSH}} | \psi \rangle + \left\| H_{\text{SSH}} - \tilde{H}_{\text{SSH}} \right\| \tag{29}
\]

\[
\leq \langle \psi | \tilde{H}_{\text{SSH}} | \psi \rangle + \frac{C}{2048(T+1)^5} \tag{30}
\]

\[
\leq \langle \psi | \tilde{H}_{\text{SSH}} | \psi \rangle + \frac{C}{2048(T+1)^5} \tag{31}
\]

where eq. (30) follows from basic properties of the spectral norm, and eq. (31) follows from eq. (23). Furthermore, since for all separable \(|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle\) we have \(\langle \psi | H_{\text{SSH}} | \psi \rangle \geq \frac{C}{256(T+1)^5}\) by the assumptions of Lemma 9, it follows that

\[
\langle \psi | \tilde{H}_{\text{SSH}} | \psi \rangle \geq \frac{7C}{2048(T+1)^5} \tag{32}
\]

Dividing eq. (28) and eq. (32) by \(\Delta\) already yields the lemma. Moreover, for \(H_{\text{SLH}}\) these bounds imply a promise gap of

\[
\gamma = \frac{1}{\Delta} \frac{7C - 5C}{2048(T+1)^5} = \frac{2C}{2048(T+1)^5} 2^{-O(\text{poly}(T\log(nkT))} \tag{33}
\]

Clearly, the inverse exponential scaling of \(\gamma\) in \(\ell\) dominates the scaling in all other parameters. Since in general QMA(2) instances \(\ell, k, T \in O(\text{poly}(n))\), the promise gap of \(H_{\text{SLH}}\) scales with \(2^{-O(\text{poly}(n))}\). \(\square\)

## 4 On the Best Separable State problem

In this section, we review the implications of our results on the related Best Separable State problem. The complexity of QMA(2) stems from essentially two sources: the search for a witness state over the set of separable states and the fact the verifier is a quantum computer. Removing the second aspect, one is lead to following related problem, which is often discussed in the context of QMA(2).

**Problem 1 (Best Separable State BSS).** Given as input an Hermitian matrix \(A \in \mathbb{C}^{d^2 \times d^2}\), with eigenvalues in \([0,1]\), and let

\[
\lambda_{\text{sep}}(A) := \max_{v,w \in \mathbb{C}^d, \|v\| = \|w\| = 1} (v \otimes w)^\dagger A (v \otimes w). \tag{34}
\]

Compute \(\tilde{\lambda}_{\text{sep}}(A)\), such that \(|\lambda_{\text{sep}}(A) - \tilde{\lambda}_{\text{sep}}(A)| \leq \varepsilon\). (Here \(\varepsilon\) is assumed to be an arbitrarily small constant if not specified otherwise.)
This problem has been related to 18 approximately equivalent problems by Harrow and Montanaro [HM13] emphasizing the importance of understanding the complexity of BSSε and its relation to QMA(2). For example, the BEST SEPARABLE STATE problem is equivalent to approximating the injective tensor norm of a 3-index tensor [DF92], a generic problem arising in several contexts (e.g. in [BBBH12] in relation to the Unique Games Conjecture [Kho02]), variants of the PLANTED CLIQUE problem, and – unsurprisingly – various problems in quantum information theory.

Note, that BSSε is clearly a generalization of QMA(2)ε,c as we have removed assumptions about the input. Indeed, by choosing d = 2poly(n) and ε = 1/poly(n), QMA(2) can be reduced to an exponentially large instance of BSSε in exponential time in n: It suffices to compute A classically by multiplying the matrices defining the verifier circuit V and choosing ε = 2−d = 1/poly(n).

**Hardness of BSSε** What is known about the hardness of the problem? BSS1/poly(d) is already known to be NP-hard [Gur03, Ioa07, Gha10] and is closely related to the problem of entanglement detection in quantum states. A long-standing open question is whether BSSε remains NP-hard in the regime of constant ε. [HM13, AIM14] give a lower bound of d3(log(d)) relative to the Exponential Time Hypothesis [IP99]. For the special case that ∥A∥F = (poly log(d)) or assuming that {A, 1 − A} is an LOCC measurement (which allows local operations and only classical communication across the subsystem boundary), quasi-polynomial time algorithms are known [BCY11, SW15].

**Impact of our results** Do our perturbative methods yield a quasi-polynomial time algorithm for the general case? Interestingly, this does not seem to be the case. Let us briefly discuss informally, why two of the most natural approaches fail to yield a better upper bound.

One natural approach is to reduce BSSε to an instance of QMAlog(2) and then apply Theorem 13. If after the reduction all parameters in the application of Theorem 13 turned out to be O(poly log(d)) this would result in a quasi-polynomial time upper bound for BSSε in terms of d. Such a reduction appears to introduce insurmountable overhead, though: Since we lack a O(log(d))-sized circuit decomposition of A, we can only recover a quantum circuit very generically by first diagonalizing A = UDU† in the eigenbasis, and then decomposing the unitary U over some gate set using the Solovay-Kitaev algorithm [NC11]. Assuming d a power of 2, ℓ = log(d), and an approximation error of ε, this will yield a quantum circuit of size

\[ T = O(4ℓ^2 4^{2\ell} \log^c(4\ell^2 4^{2\ell} / \varepsilon)) = O(poly(d, log(1/\varepsilon))) \]  

acting on 2ℓ qubits. Thus, even though k, q, ℓ, α, log(n) ≤ O(log(d)) in this case, Theorem 13 only yields a run-time bound exponential in d due to T scaling polynomial in d.\(^3\)

Another natural approach is to consider H = 1 − A as a (global) Hamiltonian with the goal to apply the perturbative gadgets immediately. Clearly, 0 ≤ H ≤ 1 and approximating 1 − λsep(A) to additive error ε is equivalent to approximating the ground energy of H over the set of separable states. Since the perturbative gadgets require a tensor product structure in the Hamiltonian terms, decompose H over the Pauli product operator basis, i.e. \( H = \sum c_{i_1, \ldots, i_n} \sigma_{i_1} \otimes \cdots \otimes \sigma_{i_n} \), where ℓ = log(d) respecting the tensor product structure of v ⊗ w. Without further assumptions, the sum consists of 4log(d^2) = d^4 terms. Using Lemma 12 iteratively O(log(ℓ)) times we can now break down each of the d^4 global terms in H into O(n) 3-local terms at the cost of increased operator norm. Furthermore, each non-local term induces O(log(d)) mediator qubits. Thus, after the reduction there are O(d4 log(d)) terms in a transformed local Hamiltonian \( \tilde{H} \) acting on an enlarged system of O(d4 log(d)) qubits or dimension d′ = 2O(d^4 log(d)). To approximate the operator in norm within ε it suffices to choose (cf. eq. (21))

\[ \Delta = 2^{O(poly(\ell) log(nk/\varepsilon))} = 2^{O(poly(\log(d)) log(d^4 log(d)/\varepsilon))} \leq 2^{O(poly(\log(d), log(1/\varepsilon)))} \]  

\(^3\)Note, that this rough bound does not even include the number of gates required for implementing D yet.
in Lemma 12. This yields the operator $\tilde{H}$ with $\|H - \tilde{H}\| \leq \varepsilon$ on the low-energy subspace of interest. Then we apply Corollary 6, which allows us to approximate $1 - \lambda_{\text{sep}}(A)$ to precision $O(\varepsilon)$ in

$$\text{DTIME}(\exp(O(\log^{O(1)}(d')(\log\log(d') + \log(\Delta/\varepsilon)))) \times \text{poly}(d', \Delta, 1/\varepsilon)),$$

which simplifies to

$$\text{DTIME}(\exp(O(\text{poly}(d, \log(1/\varepsilon))))).$$

(37)

In summary, we find that the lack of further structural information about $A$, such as a short circuit decomposition or a short Pauli decomposition (both in terms of $O(\text{poly log}(d'))$), is a significant obstacle for solving $\text{BSS}_\varepsilon$ in quasi-polynomial time using our methods. Only in the special case of $\text{QMA}(2)$, where such information is available, our method is able to effectively exploit it and yields a quasi-polynomial time upper bound (in terms of $d$), but not in the general case of $\text{BSS}_\varepsilon$. This is consistent with earlier quasi-polynomial time algorithms for $\text{BSS}_\varepsilon$ [BCY11, SW15] which require a bound on $\|A\|_F = O(\text{poly log}(d))$ as well.

5 Conclusion

We have shown the first non-trivial upper bound on $\text{QMA}(k)$. In fact, we have shown how to solve the class in deterministic exponential time and ruled out its equivalence with $\text{NEXP}$, unless $\text{NEXP} = \text{EXP}$. Our results imply an exponential time algorithm for the pure state $N$-representability problem in quantum chemistry, which is in $\text{QMA}(2)$ but is not known to be in $\text{QMA}$. Furthermore, we have discussed the implications of our results on the $\text{BSS}_\varepsilon$ problem and explained why no quasi-polynomial time algorithm for $\text{BSS}_\varepsilon$ follows. Rather, we found that the quantum circuit structure present in $\text{QMA}(k)$ but missing in $\text{BSS}_\varepsilon$ is necessary to apply our techniques effectively. In this paper, we were mainly concerned with proving an exponential time upper bound for $\text{QMA}(k)$ and leave the explicit determination of the polynomials in our upper bounds open for future work.

6 Acknowledgments

MS gratefully acknowledges the suggestion of Xiaodi Wu, during an informal presentation of an earlier, more lengthy version of the proof, to show step four of Theorem 13 by applying his previously published theorem [SW15, Corollary 6]. This helped shorten the paper and improved the presentation of the result. MS also acknowledges helpful discussions with Scott Aaronson, Yudong Cao, Aram Harrow, Daniel Nagaj, Frank Verstraete, and Thomas Vidick, as well as financial support by the Alexander-von-Humboldt Foundation. MS also thanks Jens Eisert and Daniel Nagaj for comments on earlier versions of this paper, and also the EU projects RAQUEL, SIQS, AQuS for supporting research at the Dahlem Center for Complex Quantum Systems.
References


