The status of the quantum PCP conjecture (games version)

Anand Natarajan[1] and Chinmay Nirkhe[2]

[1]Massachusetts Institute of Technology, *Cambridge, Mass.*
[2]IBM Quantum, Thomas J. Watson Research Center, *Yorktown Heights, N.Y.*

Abstract

In classical complexity theory, the two definitions of probabilistically checkable proofs – the constraint satisfaction and the nonlocal games version – are computationally equal in power. In the quantum setting, the situation is far less clear. The result $\mathsf{MIP}^* = \mathsf{RE}$ of Ji *et. al.* [JNV$^+$20a] and refinements by Natarajan and Zhang [NZ23] show that multiprover interactive proof systems with polylogarithmically long messages can solve any decision problem in RE, including undecidable problems like the halting problem. These results show that any connection between the "constraint satisfaction" or "Hamiltonian" quantum PCP conjecture and nonlocal games must involve restricting the players in the game to be *computationally efficient*. This note contains two main results: (1) we give a "quantum games PCP for AM" in the form of a new construction of a succinct MIP^* protocol with efficient provers for the canonical AM-complete problem, and (2) we explain an error in the energy amplification procedure of Natarajan and Vidick [NV18b] which invalidates their claim to have constructed a quantum games PCP for a QMA-complete problem. In surveying the obstacles remaining towards a quantum games PCP for QMA, we highlight the importance and challenge of understanding gap amplification for Hamiltonians even when locality is replaced by much weaker constraints, such as bounds on the "Pauli spectrum" of the Hamiltonian. We hope these questions will motivate progress towards new "baby versions" of Hamiltonian quantum PCP conjecture.

1 Introduction

1.1 Interactive proofs with entanglement

How powerful is an interactive proof system with provers sharing quantum entanglement? In 2020, Ji, Natarajan, Vidick, Wright, and Yuen proved that such an interactive proof system can be used to decide all recursively enumerable languages [JNV$^+$20a]. Equivalently, they proved the equality $\mathsf{MIP}^* = \mathsf{RE}$ between the respective complexity classes.

More specifically, they showed that RE is captured by *nonlocal games*, which can be thought of as an interaction between a verifier and two arbitrarily powerful devices, often denoted as Alice and Bob. Alice and Bob are spatially separated but allowed to share entanglement. The verifier (using private randomness) samples questions for Alice and Bob who then answer using their shared entanglement. The verifier then evaluates a binary relation on the question and answer pairs, and accepts or rejects accordingly.

A direct consequence of the $\mathsf{MIP}^* = \mathsf{RE}$ result [JNV$^+$20a] is the construction of a protocol with the following properties. A verifier — whose input is the description of a Turing machine, $\langle T \rangle$, and intention

is to decide whether T halts or not — sends messages q_1, q_2 to entangled and non-communicating Alice and Bob, respectively. The verifier receives answers a_1, a_2, respectively, and then performs a randomized computation $V_T(q_1, q_2, a_1, a_2)$. The property of the protocol is that there exists a strategy for Alice and Bob which causes the verifier to accept with probability 1 if T halts while *all* strategies for Alice and Bob cause the verifier to reject with high probability if T does not halt. Crucially, the algorithm V_T has a runtime which is poly(n) where $n = |\langle T \rangle|$ is the length of the description of the Turing machine; this implies that the questions and answers also have length at most poly(n). While the complexity of the verifier must be at least linear in n (by standard arguments), can the communication be shortened? More specifically, can the questions and answer lengths be made shorter?

The answer is – surprisingly – yes! Natarajan and Zhang [NZ23] improved the $\mathsf{MIP}^* = \mathsf{RE}$ result [JNV+20a] to prove that estimating the entangled value of a nonlocal game with either questions or answers of length $\operatorname{poly}\log(n)$ captures the RE complete problem of deciding if a Turing machine T halts. This result can be seen as partial progress towards resolving the *quantum games PCP conjecture* where PCP stands for Probabalistically Checkable Proofs. We will state this conjecture more precisely below in Section 1.4, but first, to put the result in context, let us take a detour to the classical PCP theorem, and the complexity of classical nonlocal games.

1.2 Probabilistically Checkable Proofs and Games

The power of nonlocal games—or multiprover interactive proof systems—with classical, unentangled players, was exactly characterized in a sequence of results [BFL91, FGL+96, AS98, H̊01][1] leading up to the PCP theorem. This theorem has several equivalent formulations, but in terms of multiprover interactive proofs (or more colloquially, nonlocal games), it states that the class NP is exactly equal to the class of problems that can be decided by one-round, two-prover MIP proof systems with $O(\log n)$-length questions and $O(1)$-length answers.

Note the following two key points of the classical games PCP theorem. Firstly, the parameters for the question and answer length immediately yield the more familiar "probabilistically checkable proofs" version of the PCP theorem. This is because any deterministic strategy to a 2-prover nonlocal game with length-q questions and length-a answers can be written down in a table of size $O(2^q \cdot a)$, and the interaction between the provers and verifier can be simulated by querying $2a$ bits in the table. This means that for any language in NP, there is a *probabilistically checkable proof system* for it: the verifier receives a string of length $O(2^q \cdot a) = \operatorname{poly}(n)$ (which in the honest case is the table corresponding to the optimal prover strategy), and makes $2a = O(1)$ queries to it. If the answer is YES, then the verifier will accept the honest proof string with high probability, whereas if the answer is NO, it will reject all proof strings with high probability. In fact, the implication goes the other way as well: any probabilistically checkable proof with

[1] The sequence of results stems from a result by Babai, Fortnow, and Lund [BFL91] which showed a multiround interaction between multiple provers and a verifier for NEXP with $O(\operatorname{poly}(n))$ sized total question length and $O(\operatorname{poly}(n))$ sized total answer length. Later results [FGL+96, AS98, H̊01] reduced the round complexity to 1, answer length to $O(1)$, and demonstrated a compression which gave the stated equivalence to NP.

polynomial-sized proofs implies a two-player interactive proof system with $O(\log n)$-length questions and $O(1)$-length answers. This is due to a standard transformation called the *clause-variable game*; a detailed treatment can be found in Thomas Vidick's lecture notes on the quantum PCP conjecture [Vid14]. Thus, the two formulations of the classical PCP theorem are equivalent.

Secondly, the protocol arising from the games version of the PCP theorem is *prover efficient*[2]. In the classical nonlocal games setting, we find that in reduction from any language $\mathcal{L} \in \mathsf{NP}$ to a family of nonlocal games, the *honest* classical provers in the game can generate (in polynomial time) the winning answers to the questions from the witness w to the original NP language. In other words, the honest provers only need to be P^w powerful. Meanwhile, the protocol is sound against arbitrarily powerful classical powers.

1.3 The quantum PCP conjectures

In the quantum case, we have no PCP theorem, but rather several formulations of quantum PCP conjectures, which are not known to be equivalent. A standard version of the conjecture is the "Hamiltonian qPCP conjecture", which states that a gapped version of the local Hamiltonian problem is QMA-complete under quantum polynomial-time reductions.

Conjecture 1 (Quantum PCP [AN02, AAV13])**.** *It is* QMA*-complete under quantum polynomial-time reductions to decide*[3] *whether a local Hamiltonian on n qubits and* $m = \Theta(n)$ *terms has ground energy (minimum eigenvalue)* $\leq m/10$ *(YES instance) or* $\geq m/5$ *(NO instance) even when promised that one of the cases holds.*

This conjecture can equivalently be stated in terms of the existence of probabilistically checkable *quantum* proofs for any language in QMA. This conjecture is discussed extensively in the survey by Aharonov, Arad, and Vidick [AAV13]. Another version of the conjecture concerns the "proof-checking" property of a quantum PCP. It is known that the proof-checking version of the conjecture is equivalent ot the Hamiltonian version under quantum polynomial-time reductions since the previously mentioned survey [AAV13]. A recent note by Burhman, Helsen, and Weggemans [BHW24] has resolved certain questions about the *adaptivity* of the proof-checking and connections to the complexity class QCMA – however, we do not discuss the proof-checking version any further in this note.

In addition to trying to generalize the proof-checking version of the classical PCP, attempts have been made to formulate a *quantum games PCP*, generalizing the statement of the classical PCP theorem in terms of games or MIP proof systems for NP. This direction was first proposed by Fitzsimons and Vidick [FV15], but in light of the subsequent progress in our understanding of MIP^*, it is worth revisiting it. Motivated by

[2]There is also a natural concern of the efficiency of a verifier. In some context a proof that is both verifier and prover efficient is called doubly efficient. However, in most such cases, such as the Goldwasser, Kalai, and Rothblum interactive proofs [GKR15], the notion of efficiency is of a fine grained nature concerning the efficiency of the verifer and prover in terms of the size or depth of the formula. Specifically, for [GKR15], the verifier should run in time $O(n\text{poly}(d))$ and space $O(\log n)$ and the prover in time $\text{poly}(n)$ where d is the depth of the circuit. However, in this note we are focused on the coarser perspective on complexity and therefore we are content with the verifier being efficient if they are a BPP or BQP device. Therefore, to emphasize the importance on making the prover efficient, we call these problems only prover efficient and not doubly efficient.

[3]The constants of $1/10$ and $1/5$ could be replaced with any other choice of constants.

our presentation of the classical games PCP, we propose that a reasonable statement of a quantum games PCP conjecture should satisfy the following requirements:

- It should put QMA into a class of MIP* proof systems with short questions and answers.
- The MIP* proof system should have question length $q = O(\log n)$ and answer length $a = O(1)$.
- Honest provers should be efficient, given (copies of) the QMA witness.

The purpose of these requirements is not to blindly mimic the classical case, but to preserve the hope that the resulting MIP* proof system will say something about the Hamiltonian qPCP conjecture. In particular, one may hope that these constraints will result in a proof system where the honest provers' strategy involves constructing some kind of polynomial-sized quantum probabilistically checkable proof, by performing an efficient transformation on the QMA witness.

1.4 Towards a quantum games PCP

We previously remarked that the Natarajan and Zhang result [NZ23] is only partial progress towards a quantum games PCP. This is because it fails to satisfy the second and third requirements given above (and arguably only satisfies the first requirement by a technicality, since the result applies to all of RE and not just QMA).

The fact that the Natarajan and Zhang result does not achieve the gold standard of $O(\log n)$ sized questions and $O(1)$ sized answers, and instead requires $\text{poly}\log(n)$ sized questions and answers, is the more minor of two reasons. We believe the roadblock to be more minor as the original games PCP for NP by Feige *et. al.* [FGL+96] involved $O(\text{poly}\log n)$ length messages and subsequent improvements yielded $O(\log n)$ length messages [AS98, ALM+98]. It seems plausible that similar improvements can be achieved in the MIP* setting by refining known techniques (although it is worth noting that, as remarked in [NZ23], achieving this improvement will require improving or replacing the analysis of the quantum soundess of the low-degree test in [JNV+20b]).

The second – and more substantial roadblock – is the complexity of the *honest* provers. The quantum analog of the efficient-provers property is to consider scenarios where the honest prover is not all-powerful but rather is only BQP-powerful but imbued with the solution to the particular problem. For example, in the case of a NP problem x, we can morally think of the honest provers, by analogy, as BQP^w powerful where w is the witness to the problem. More technically, the provers will need to share entanglement and potentially this entanglement may rely on the witness w. Recall that a nonlocal game consists of three phases: (a) a setup phase where Alice and Bob generate their specific entangled state ρ_{AB}, (b) an interaction phase where the game is played with the verifier, (c) and a grading phase done exclusively by the verifier. The technical definition of efficient provers will be one where (a) the setup phase can be performed by a quantum polynomial-time (QPT) device with access to w and (b) the interaction phases can be performed by QPT devices with access to their respective share of the state ρ_{AB}. To extend this

definition to QMA, a class where the witness is quantum, we define an efficient nonlocal game capturing a language in QMA like the NP definition except the classical proof w is replaced by $|\psi\rangle^{\otimes \mathrm{poly}(n)}$ in the setup phase, where $|\psi\rangle$ is the quantum witness for the problem. The rationale for multiple copies of the witness is that the efficient provers cannot (in general) clone $|\psi\rangle$, and many standard reductions between QMA protocols (e.g. amplification procedures) require multiple copies of the witness.

It is important to emphasize that prover efficiency is a property of a reduction from one language to another and not a property of a complexity class. This is because complexity class equalities and reductions between languages within a complexity class may not be prover efficient. The most pertinent example to keep in mind is that, even though QMA $\subseteq$ NEXP, and therefore, the QMA-complete local Hamiltonian problem can be reduced to a NEXP-complete problem such as Succinct-3-Coloring, the witness for the coloring problem is likely not efficiently computable from the quantum witness to the local Hamiltonian problem. Therefore, when discussing prover efficiency we will be careful to specify a language (and not a complexity class) and a specific model of the proof.

A priori, one might suspect that the construction of prover-efficient nonlocal games with short questions would be easiest for a language in the class QMA where the witness is a quantum state since the state of the honest provers could simply be the witness for the QMA problem. However, as we remark in this note, there are significant roadblocks to constructing such games and a construction is not known. We believe that this is the most practical question remaining in the pantheon of nonlocal game theory and, therefore, the appropriate question to be called the quantum PCP games conjecture.

Conjecture 2 (Quantum PCP (Games Version))**.** *There exists a prover-efficient* MIP* *protocol with* $O(\log n)$*-length questions and* $O(1)$*-length answers for* QMA*.*

More formally, for every language $\mathcal{L} \in$ QMA*, there exists a polynomial time verifier* V *and quantum polynomial time provers* $P_1, \ldots, P_k$ *such that:*

1. *If* $x \in \mathcal{L}$*, then there exists a* $\mathrm{poly}(n)$*-qubit state* $|\psi\rangle$ *such that* V *on input* x*, interacting with* $P_1, \ldots, P_k$ *on input* $(x, |\psi\rangle^{\otimes \mathrm{poly}(n)})$*, accepts with probability* $2/3$*. In particular, the provers* $P_1, \ldots, P_k$*, in the setup phase of the protocol, generate their shared entangled state in polynomial time from the input* $(x, |\psi\rangle^{\otimes \mathrm{poly}(n)})$*.*
2. *If* $x \notin \mathcal{L}$*, then for any (not necessarily efficient) provers* $P_1^*, \ldots, P_k^*$*, the verifier* V *given input* x *and interacting with* $P_1^*, \ldots, P_k^*$ *accepts with probability at most* $1/3$*.*

We remark that here, $|\psi\rangle$ *is allowed to be an arbitrary state depending on* x*. However, one may think of it as a* QMA *witness for* x*: indeed, any* MIP* *protocol of this form implies a* QMA *protocol for* $\mathcal{L}$*, where honest witness is* $|\psi\rangle^{\otimes n}$*, and the* QMA *verifier simulates the interaction between* V *and* $P_1, \ldots, P_k$ *(since the provers are efficient).*

The MIP* = RE result of [JNV$^+$20a] does not directly say anything about the efficiency of the provers, but by inspecting the strategy for the honest provers given in the completeness case of the protocol, we can

obtain explicit upper bounds on the prover runtime when the language $\mathcal{L}$ to be decided is in a time-bounded complexity class. Unfortunately, these bounds are not good enough for QMA, because they only depend on the *classical* nondeterministic time complexity of $\mathcal{L}$. Specifically, for any problem in $\mathsf{NTIME}[t(n)]$, the honest prover strategy requires quantum time $\mathrm{poly}(t(n))$, given a classical nondeterministic witness of size $t(n)$. This strategy requires the prover to compute a PCP-style encoding π of a classical tableau of the execution of a Turing machine solving the problem in $t(n)$ steps, and prepare on the order of $\mathrm{poly}\log t(n)$ EPR pairs; in response to a verifier question, the prover measures the EPR pairs to obtain indices into the encoded tableau, and reports the value of the tableau at those indices. For the specific case of BQP or QMA, the best known $\mathsf{NTIME}[t(n)]$ bounds on these classes are $t(n) = \exp(n)$. Moreover, for QMA, even given copies of the witness state, we do not know how an quantum prover could compute entries of π in time less than $t(n)$.

1.5 Contributions of this note

In this note, we prove two results and comment on the current state of affairs. The first is that prover-efficient nonlocal games exist for AM protocols with $\mathrm{poly}\log(n)$ sized questions and answers. Second, we describe the error in the incorrect result in Natarajan and Vidick [NV18b] claiming a reduction from the local Hamiltonian problems to quantum games PCPs. Lastly, we remark on the outstanding roadblocks for the games version of the QPCP conjecture – i.e. constructing prover-efficient nonlocal games for QMA.

2 Preliminaries

2.1 Nomenclature and notation

The majority of this note regards the computational complexity of computing the entangled value of a nonlocal game or the QMA-complete problem of computing the minimum eigenvalue of a local Hamiltonian. To simplify the reductions between these two problems, we will instead consider computing the *maximum* eigenvalue of a local Hamiltonian.

The problems can be expressed as optimization problems, but for connections to the pantheon of computational complexity classes, it is useful to consider the *promise-gapped* version of the decision problem. The problem of deciding, for a game G whether the entangled value, $\omega^*(G)$, is $\geq$ than c (YES instances) or $\leq s$ (NO instances) is the promise-gapped version of the problem with $\mathrm{pgap} = c - s$. Likewise, the promise-gapped version of the local Hamiltonian problem is to decide for a local Hamiltonian $\mathbf{H}$, whether (YES instances) $\lambda_{\max}(\mathbf{H}) \geq c$ or (NO instances) $\lambda_{\max}(\mathbf{H}) \leq s$.

The principle goal of this note is to perform reductions between games and Hamiltonians that amplify the promise gap. Notationally, we will discuss "amplifying the promise gap" of the game (or the Hamiltonian) when we precisely mean generate a reduction from a set of games (or Hamiltonian instances) to a new set of games (or Hamiltonian instances) with a promise gap between YES and NO instances greater

than the promise gap of the initial set. Lastly, unless specified otherwise, when discussing games we are always referring to the decision problem of estimating the entangled value of the game.

2.2 The computational power of prover *non-efficient* nonlocal games

As previously mentioned in the Introduction, stemming from the introspection (i.e. question and answer reduction) tools developed in [NW19, JNV+20a], Natarajan and Zhang [NZ23] proved that any Turing machine halting question can be transformed into a nonlocal game. This is the *state-of-the-art* in terms of succinct nonlocal games. Note, the following theorem states nothing about the efficiency of the honest provers.

Theorem 3 (Theorem 58 in [NZ23])**.** $\mathsf{MIP}^*[q = O(1), a = O(\text{poly}\log n)] = \mathsf{RE}$.

The appropriate interpretation is the every Turing machine halting problem, defined by input $\langle T \rangle$ with $n \stackrel{\text{def}}{=} |\langle T \rangle|$ can be reducing to deciding the quantum value $\omega^\star$ of a nonlocal game described by a table with constant sized questions and $O(\text{poly}\log n)$ sized answers – i.e. the tableau of the verification function[4] $V(q_1, q_2, a_1, a_2)$. Moreover, the table of the game can be computed from the description of the Turing machine in classical *quasi-polynomial time*. Equivalently, $\mathsf{MIP}^\star \supseteq \mathsf{RE}$ under quasi-polynomial time reductions since the Turing machine halting problem is RE-complete. The containment $\mathsf{MIP}^\star \subseteq \mathsf{RE}$ under quasi-polynomial time reductions is straightforward.

The nuance of the $O(\text{poly}\log n)$ sized questions versus the $O(\log n)$ sized questions in the classical games PCP stems from $O(\text{poly}\log n)$ sized questions required for the *quantum low individual degree test* [JNV+20b] which is a seminal step in the introspection arguments used in [JNV+20a] and [NZ23]. We believe (but it deserves a closer inspection through the entire lengthy proof) that a more efficient alternative to the quantum low individual degree test would close the gap on the polylogarithmic vs logarithmic difference between the quantum and classical variants.

3 An efficient nonlocal game for all AM languages

An AM protocol for deciding language membership $x \in \mathcal{L}$ is an interaction between a prover (Merlin) and a public-key verifier (Arthur) where in the first message Arthur sends a uniformly random $r \in \{0,1\}^{\text{poly}(n)} = R$ and then Merlin responds with $w(r)$ according to a witness $w : R \to \{0,1\}^{\text{poly}(n)}$. Arthur follows by running a deterministic computation $V(x, r, w(r))$.

It is well known that AM protocols can be constructed with completeness 1 and soundness $\leq 1/3$. We can call an honest prover efficient (or for brevity, efficient) if the honest prover can be simulated by a P^w machine – i.e. a polynomial-time machine which has access to w but no other "super-natural" computational abilities. In the entangled game setting, we are interested in understand efficient nonlocal games where the honest provers can be simulated by BQP^w machines. The amplification from P to BQP is

[4] This table has size $O(4^{q+a}) = \text{quasipoly}(n)$.

inherently necessary as the provers must share entanglement in order to have more computational power than NP — and as we don't know if AM = NP, we must consider entangled strategies.

To understand the efficient nonlocal game for AM at a high level, consider the following honest provers, Alice and Bob. They receive no input from the verifier, and instead sample their own questions by measuring $|r|$ EPR pairs in the same basis to generate the same question string r. Alice's answer to the verifier is simply r, and Bob's answer to the verifier is $(r, w(r))$. The verifier then checks that the answers have consistent strings r and that $V(x, r, w(r))$ accepts.

Notice, that if the randomness r is sampled correctly, then by the soundness of the original AM problem, the provers are incapable of any deceit as they must answer a witness for r. Of course, the principal issue is that we cannot trust nefarious provers to sample their own questions. This is where we can use the introspection tools of [JNV+20a]. With introspection, a verifier can use $\text{poly}\log(n)$ sized questions and force the provers to sample their own questions. This gives us a game with short questions and long answers for AM.

The next step is to have the provers grade their own witness — or equivalently, answer-reduction. Since Bob's proof has both r and w and x is public, Bob can generate a *probabilistically checkable proof of proximity* (PCPP) of the fact that $V(x, r, w)$ accepts. Applying this intuition into the answer reduction argument of [JNV+20a] gets us short questions and answers.

Theorem 4. *For every language $\mathcal{L} \in$ AM, there is a reduction from $\mathcal{L}$ to a language of nonlocal games with questions and answers of size $O(\text{poly}\log(n))$ such that (a) there is a reduction from problems in $\mathcal{L}$ to the entangled value of the game and (b) for any YES instance of the language $\mathcal{L}$ with witness $w = w(r)$, the honest provers only need be* BQP*-powerful with knowledge of w.*

Furthermore, one might ask whether prover-efficient succinct nonlocal games are achievable for languages in classes significantly larger than AM or QMA, such as NEXP, NEEXP, or even RE. While one can make such claims, the issue – as we suggested earlier – is in the subtelties of the definition of prover efficiency. Consider, for instance, a reduction from the NEXP-complete language of Succinct-3-Coloring to a nonlocal game. A consequence of Impagliazzo, Kabanets, and Wigderson [IKW02] shows that unless $\text{NEXP} \subseteq \Sigma_2$, the coloring function for Succinct-3-Coloring problems cannot be expressed succinctly. Therefore, assuming this monumental complexity class collapse is false, we can only prove prover efficiency in a model of *oracle access* to the witness instead of the prover actually holding the witness itself[5].

Secondly, its worth noting that we know of prover-efficient nonlocal games for the class $\text{PCP}[q = \text{poly}(n), a = O(1)]$ as the original proof of $\text{MIP}^\star \subset \text{MIP}$ by Ito and Vidick [IV12] is prover efficient. However, the transformation from a witness of a NEXP-complete language such as Succinct-3-Coloring to one for the PCP equivalent language may not be prover efficient. For this to be the case, we need that the witness w' for

[5] An astute reader might find this argument peculiar since we considered provers with access to the AM witness function $w(r)$ in the previous example. Indeed, if the function w were efficiently describable, the problem is actually in the class MA as the prover could send the description of the function w, upon which the verifier could sample randomness and verify that $w \cong 1$. The principle difference in these two examples is that the derandomization of $\text{AM} = \text{MA} = \text{NP}$ is well believed to be true, while the collapse of the polynomial hierarchy in $\text{NEXP} \subseteq \text{AM}$ is widely believed to be false.

the PCP version of the instance can be simulated with polynomial queries to the witness w of the NEXP version of the instance. To the best of our knowledge, applications of the known PCP transformations such as that of Dinur [Din07] require exponentially many queries to w. While we remark, but do not explicitly prove here, that a similar proof to the AM argument above yields a quantum games PCP for the canonical language for $\mathsf{PCP}[q = \mathrm{poly}(n), a = O(1)]$ that is prover efficient in the sense that the prover has oracle access to the witness for the PCP, we want to emphasize that this doesn't necessarily yield a prover-efficient argument for all NEXP languages. A similar remark extends for languages in complexity classes past NEXP.

4 The difficulty in constructing nonlocal games for languages in QMA

Given that MIP* allows for quantum provers, it seems natural to ask for efficient-prover protocols for a quantum problem: rather than a classical complexity class like AM, can we achieve such protocols for QMA? This is the "quantum games PCP" conjecture, first proposed by Fitzsimons and Vidick [FV15]. In 2018, Natarajan and Vidick claimed [NV18b] a resolution of this conjecture; however, this result relied on an earlier result of Vidick [Vid16] on the quantum soundness of the plane-vs-point low-degree test, whose proof turned out to have a bug. Nevertheless, the authors believed that a version of the result with weakened parameters—polylogarithmic instead of logarithmic questions—would hold using the weakened replacement for Vidick's low-degree test result obtained by Ji et al. [JNV+20a]. Unfortunately, it was discovered that the Natarajan and Vidick protocol for QMA is flawed in *another* way, so that it is now open whether QMA can be put in $\mathsf{MIP}^*[q = O(\log n), a = O(\log n)]$ with efficient provers even if Vidick's original low-degree result is recovered. In this section we explain Natarajan and Vidick's approach, why it fails, and state a corrected version of their faulty amplification lemma.

4.1 A template for MIP* protocols for QMA

The protocol in [NV18b] arose from a line of work originated by Fitzsimons and Vidick [FV15] in 2014. This work gave a prover-efficient protocol for the local Hamiltonian problem with short *quantum* messages, and subconstant soundness gap. The key idea was to distribute the ground state $|\psi\rangle$ amongst multiple provers using an error correcting code as a quantum "secret sharing" scheme. Specifically, in their protocol, there are five provers, and the ground state is encoded qubit-by-qubit with the $[[5, 1, 3]]$ error correcting code with a share being sent to each prover. To verify the energy, the verifier asks each prover for a small number of qubits from their share, and then jointly decodes the shares to measure a single local term of the Hamiltonian on the decoded state.

A series of works [Ji17, NV17] improved this protocol to one with classical messages, by using *self-testing*, a powerful tool in the nonlocal games literature, to force the provers to perform the measurements themselves on their shares, and honestly report their measurement outcomes to the verifier. The main new technical result of [NV18b], which now appears in a streamlined and self-contained form in the Appendix

of [JNV+20a] is the *quantum low-degree test*, a powerful self-test that can force the provers to perform tensor products of Pauli X- and Z-measurements, and has constant soundness gap and polylogarithmic question size.

Using this result, [NV18b] gave a "gap-preseving" protocol to approximate the ground energy of a Hamiltonian $\mathbf{H}$ consisting of a sum of (possibly high-weight) tensor products of Pauli X- and Z-operators.

1. Ask the provers to share the ground state using a qubit-by-qubit encoding, but using the $[[7, 1, 3]]$ Steane code. This means there are 7 provers.

2. Check that the provers share a valid code state, by commanding them to measure the stabilizers of the code. Since the Steane code is CSS, the stabilizers consist only of tensor products of X and Z, and so can be measured by the self-test.

3. Pick a term from the Hamiltonian $\mathbf{H}$. Measure the energy of the provers' state with respect to this term by asking them to measure the corresponding logical operator on the code state. Again, since the Steane code is CSS, and each term of $\mathbf{H}$ is a tensor product of Paulis, the logical operator is itself a tensor product of X- and Z-operators, and can be measured by the self-test.

This protocol is "gap-preserving" in that the optimal success probability of the provers is related linearly to the ground energy of the Hamiltonian independently of the number of qubits: an energy gap of δ in the Hamiltonian corresponds to a gap in acceptance probabilities of $O(\delta)$, independent of the number of qubits n.

In order to use this protocol to solve a QMA-complete problem with a constant soundness gap, it suffices to find a family of Hamiltonians $\mathbf{H}$ of the described form, for which finding the ground energy up to a constant factor is QMA-hard. Natrajan and Vidick attempted this by designing an amplification procedure for the known QMA-hard problem of approximating the ground energy of a local $XX + ZZ$ Hamiltonian up to inverse polynomial gap [CM16].

Unfortunately, this amplification procedure was incorrect[6]! The issue arises from the *normalization* of the energy of the Hamiltonian. Typically, one measures the energy on a scale set by the operator norm of $\mathbf{H}$, so that a "constant energy gap" means that the energy gap scales as $\Omega(\|\mathbf{H}\|)$ independently of the number of qubits n. However, a careful examination of the protocol described above reveals that the relevant scale is not the operator norm, but the 1-norm of the vector of coefficients of $\mathbf{H}$ in the Pauli basis, a quantity which we define below as the Pauli 1-norm. While the procedure of [NV18b] preserves the operator norm, it causes the Pauli 1-norm to grow exponentially, destroying the amplified energy gap upon renormalization. In the remainder of this section, we explain why this is the correct normalization to consider, and then describe the Natarajan and Vidick amplification procedure and give the correct scaling of the norms.

[6]The issue was first realized by Alex Lombardi who informed Natarajan and Vidick of it in a personal communication.

4.2 Measuring the energy of a state via Pauli tensor measurements

As we described in our summary of Natarajan and Vidick [NV18b], it is a general technique in nonlocal games to be able to force the provers to measure the energy of the specific state $|\psi\rangle$ they possess with respect to *a single* Pauli observable of the verifier's choice. The intention is to use this ability to estimate $\langle\psi|\mathbf{H}|\psi\rangle$ with the honest provers running the protocol with $|\psi\rangle$ being the state of maximal eigenvalue. Formally, consider a Hamiltonian $\mathbf{H}$ expressed in a decomposition as a summation over Paulis,

$$\mathbf{H} = \sum_{P \in S} \beta_P P. \tag{1}$$

Then the following procedure has an acceptance probability whose bias away from $1/2$ is proportional to the energy of the state.

1. Pick a Pauli P from S with probability $\mathbf{Pr}[P] \stackrel{\text{def}}{=} \dfrac{|\beta_P|}{\sum_{P \in S} |\beta_P|}$.
2. Measure P to get a bit $b \in \{\pm 1\}$.
3. Accept if $b = \text{sign}(\beta_P)$.

The acceptance probability can be calculated directly as

$$\begin{aligned}
\mathbf{Pr}[\text{accept}] &= \sum_{P \in S} \mathbf{Pr}[P] \cdot \mathbf{Pr}[b = \text{sign}(\beta_P)] && \text{(2a)}\\
&= \sum_{P \in S} \frac{|\beta_P|}{\sum_{P \in S} |\beta_P|} \mathbf{Pr}[b \cdot \text{sign}(\beta_P) = 1] && \text{(2b)}\\
&= \sum_{P \in S} \frac{|\beta_P|}{\sum_{P \in S} |\beta_P|} \cdot \left(\frac{1}{2} + \frac{1}{2} \text{sign}(\beta_P) \langle b \rangle_\psi \right) && \text{(2c)}\\
&= \frac{1}{2} + \frac{1}{2} \cdot \sum_P \frac{\beta_P}{\sum_{P \in S} |\beta_P|} \langle\psi|P|\psi\rangle && \text{(2d)}\\
&= \frac{1}{2} + \frac{\langle\psi|\mathbf{H}|\psi\rangle}{2\|\mathbf{H}\|_{\mathcal{P},1}}. && \text{(2e)}
\end{aligned}$$

In the last line we introduce the notion of the $\|\cdot\|_{\mathcal{P},1}$, the Pauli 1-norm, which is the minimum weight of $\sum_{P \in S} |\beta_P|$ when a Hamiltonian $\mathbf{H}$ is expressed in a decomposition as a summation over Paulis (as in eq. (1)).

Therefore, it is important to note that in *all known* constructions [FV15, NV18b] of nonlocal games from local Hamiltonians – including the proposal by Natarajan and Vidick [NV18b] – rely on a Hamiltonian energy measurement test with analysis analogous to the above calculation. The consequence of the success probability in eq. (2), is that the promise gap of the resultant nonlocal game family $\mathcal{L}_{\text{game}}$ is

$$\text{pgap}(\mathcal{L}_{\text{game}}) \geq \max_{\mathbf{H} \in \mathcal{L}_{\text{Hamiltonian}}} \left\{ \frac{1}{\|\mathbf{H}\|_{\mathcal{P},1}} \right\} \cdot \text{pgap}(\mathcal{L}_{\text{Hamiltonian}}). \tag{3}$$

Previously, we stated that the Natrajan and Vidick transformation [NV18b] generates a reduction from local Hamiltonians to nonlocal games with a promise gap of the game polynomial in the promise gap of the Hamiltonian. This is because they consider $XX + ZZ$ Hamiltonians which by definition have a Pauli 1-norm of at most $O(n^2)$. This is not the case for general Hamiltonians as the Pauli 1-norm can vastly vary from more "standard" norms as illustrated by the following example.

Remark 5. *The Pauli 1-norm is always an upper bound on the operator norm, but it can be much larger than it. Consider the Hadamard matrix*

$$H = \frac{1}{\sqrt{2}}(X + Z). \tag{4}$$

This has $\|H\| = 1$ *but* $\|H\|_{\mathcal{P},1} = \sqrt{2}$*. This gap can be exponentially amplified by tensor powers:*

$$\|H^{\otimes n}\| = 1, \|H^{\otimes n}\|_{\mathcal{P},1} = \sqrt{2^n}. \tag{5}$$

The principal error in Natarajan and Vidick [NV18b] is that they considered generating a nonlocal game for the tensor-product amplification of the Hamiltonian $\mathbf{H}$ and they incorrectly calculated the amplification of the Pauli 1-norm in this transformation. Therefore, they incorrectly concluded a nonlocal game of constant promise gap from the tensor-product amplification of a QMA-complete family of $XX + ZZ$ Hamiltonians. The following lemma provides the rectified amplification and accounts for the amplification of the Pauli 1-norm.

4.3 Hamiltonian promise gap and Pauli 1-norm amplification

Lemma 6 (Hamiltonian promise gap amplification)**.** *Consider a n-qudit ℓ-local Hamiltonian $\mathbf{H}$ such that $-1 \preccurlyeq \mathbf{H} \preccurlyeq 1$ and further assume a promise that $\lambda_{\max}(\mathbf{H}) \geq 1 - 1/p$ or $\lambda_{\max}(\mathbf{H}) \leq 1 - 1/q$. Then there exists an efficient transformation producing a $\ell \cdot k$-local Hamiltonian $\mathbf{H}'$ such that $-1 \preccurlyeq \mathbf{H}' \preccurlyeq 1$ and that*

$$\lambda_{\max}(\mathbf{H}') \geq 1 - \frac{k}{p} \quad or \quad \lambda_{\max}(\mathbf{H}') \leq 2\mathbf{e}^{-\frac{k}{2q}} - 1, \tag{6}$$

respective to the two promised cases. Furthermore, the Pauli 1-norm of $\mathbf{H}'$ can be bounded as

$$\|\mathbf{H}'\|_{\mathcal{P},1} \leq 1 + 2\left(\frac{1 + \|\mathbf{H}\|_{\mathcal{P},1}}{2}\right)^k. \tag{7}$$

Note that if $\|\mathbf{H}\|_{\mathcal{P},1} \leq 1$, then $\|\mathbf{H}'\|_{\mathcal{P},1} \leq 3$ and is, therefore, bounded. However, it is unknown whether there exists a family of local Hamiltonians which capture QMA for which the Pauli 1-norm is bounded and amplification is possible.

One might consider two strategies to adjust a Hamiltonian such that its Pauli 1-norm is manageable. However, neither of these will prove successful for achieving a nonlocal game with consant promise gap. The first would be to be to scale the Hamiltonian $\mathbf{H} \mapsto \theta\mathbf{H}$; for an appropriate $\theta \leq 1/\|\mathbf{H}\|_{\mathcal{P},1}$, this would certainly give a threshold on the Pauli 1-norm but the promise gap parameters of $1 - 1/p$ and $1 - 1/q$ would also move and make amplification difficult. This does not seem like a viable strategy.

The second is to consider applying a randomized restriction to the amplified Hamiltonian $\mathbf{H}'$ to reduce the Pauli 1-norm. Similar to Lemma 36 of [AN22], we can consider sampling $m = \Omega(n/\delta^2)$ terms from the Hamiltonian $\mathbf{H}'$ to generate a Hamiltonian $\mathbf{H}''$. Application of the operator Chernoff bound [Tro12, Lemma 2.8] gives that

$$\mathbf{Pr}\left[\left\|\mathbf{H}' - \mathbf{H}''\right\| \geq \delta\right] \leq 2^n \mathbf{e}^{-m\delta^2/32} \leq 1/3. \tag{8}$$

If we apply this after having amplified $\mathbf{H}'$ to a constant promise gap, we can select $\delta = O(1)$, we arrive at a Hamiltonian with manifestly fewer terms but a Pauli 1-norm that still scales with $O(n)$, at best. Therefore, the resulting nonlocal games promise gap will not be a constant.

Proof of Lemma 6. The intention is to amplify the promise gap by considering a tensor product of the original Hamiltonian. We exploit the following two basic facts:

1. If $0 \preccurlyeq \mathbf{M}$, then $\lambda_{\max}(\mathbf{M}^{\otimes k}) = \lambda_{\max}(\mathbf{M})^k$.
2. For small ϵ, $(1-\epsilon)^n \approx 1 - k\epsilon$.

These facts tell us that if $\mathbf{M}$ is a positive operator whose top eigenvalue is either 1 or $1 - \epsilon$, then the top eigenvalue of $\mathbf{M}^{\otimes k}$ is either 1 or bounded away from 1 by a constant, for k on the order of $1/\epsilon$.

To apply this idea to our situation, we first have to linearly shift the Hamiltonian so that the Hamiltonian's spectrum is non-negative. Specifically, we will shift the Hamiltonian by a multiple of identity to make it positive, amplify using by taking tensor product (shown above), and then shift back:

$$\mathbf{H}' \stackrel{\text{def}}{=} 2\left(\frac{\mathbb{I} + \mathbf{H}}{2}\right)^{\otimes k} - \mathbb{I}. \tag{9}$$

By construction $\mathbf{H}'$ has operator norm $\|\mathbf{H}'\| \leq 1$. It remains to bound the its top eigenvalue in the two cases, and to bound its Pauli 1-norm. Let $\lambda_{\max}(\mathbf{H}) = 1 - 1/a$. Then,

$$\lambda_{\max}(\mathbf{H}') = 2\left(\frac{1 + \lambda_{\max}(\mathbf{H})}{2}\right)^k - 1 = 2\left(1 - \frac{1}{2a}\right)^k - 1 = 2\left(1 - \frac{1}{2a}\right)^{2a \cdot \frac{k}{2a}} - 1. \tag{10}$$

Well-known bounds can be applied to lower- and upper-bound the top eigenvalue $\lambda_{\max}(\mathbf{H}')$.

$$2\left(1 - \frac{k}{2a}\right) - 1 \leq \lambda_{\max}(\mathbf{H}') \leq 2\mathbf{e}^{-\frac{k}{2a}} - 1 \tag{11}$$

Therefore, the promise gap of $1/q - 1/p$ is amplified to at least

$$\geq 2\left(\left(1-\frac{k}{2p}\right)-\mathbf{e}^{-\frac{k}{2q}}\right)\geq k\left(\frac{1}{2q}-\frac{1}{p}\right) \tag{12}$$

This is the "absolute" energy gap, prior to normalization by the Pauli 1-norm[7]. Next we compute the Pauli basis decomposition of $\mathbf{H}'$. Let S be the set of Pauli terms in H and let $m = |S|$.

$$\mathbf{H} = \mathop{\mathbf{E}}_{P\in S}[\alpha_P P] \tag{13a}$$

$$\frac{\mathbb{I}+\mathbf{H}}{2} = \frac{1}{2}\mathbb{I} + \frac{1}{m}\sum_{P\in S}\frac{1}{2}\alpha_P P \tag{13b}$$

$$= \frac{1}{2}\left(1+\frac{\alpha_I}{m}\right)\mathbb{I} + \frac{1}{m}\sum_{P\in S\setminus\{I\}}\frac{1}{2}\alpha_P P \tag{13c}$$

$$\|(I+\mathbf{H})/2\|_{\mathcal{P},1} \leq \frac{1}{2}+\frac{1}{2}\|\mathbf{H}\|_{\mathcal{P},1} \tag{13d}$$

$$\|((I+\mathbf{H})/2)^{\otimes k}\|_{\mathcal{P},1} \leq \frac{1}{2^k}(1+\|\mathbf{H}\|_{\mathcal{P},1})^k \tag{13e}$$

$$\|2((I+\mathbf{H})/2)^{\otimes k}\|_{\mathcal{P},1} \leq \frac{1}{2^{k-1}}(1+\|\mathbf{H}\|_{\mathcal{P},1})^k \tag{13f}$$

$$\|\mathbf{H}'\|_{\mathcal{P},1} \leq 1+2\left(\frac{1+\|\mathbf{H}\|_{\mathcal{P},1}}{2}\right)^k. \tag{13g}$$

□

5 Open problems

Aside from the obvious open conjecture of the quantum games PCP, the following are what we believe to be more accessible stepping stones.

1. Is there a QMA-complete family of Hamiltonians with bounded $\|\mathbf{H}\|_{\mathcal{P},1}$ for which Lemma 6 can be applied? This is an interesting question even if the Hamiltonian family does not have the $XX+ZZ$ form that makes it immediately amenable to a nonlocal game reduction.

2. Towards a negative answer to the previous question, for what classes of nonlocal Hamiltonians with $\|\mathbf{H}\|_{\mathcal{P},1} \leq 1$ can we show that finding the ground energy up to constant gap (relative to $\|H\|$) is *not* QMA-hard? So far, work on approximations for Hamiltonians has focused on the local case, where ansatzes such as product states are useful. Here we do not expect product states to serve as good ansatzes here, but perhaps more algebraic techniques (e.g. ncSoS) will yield fruitful results. This could be used to show that the local Hamiltonian problem is not as hard as it seems.

[7]For the QMA-complete Feynman-Kitaev circuit-to-Hamiltonian construction, the promise cases are constructed with $1/p$ as inverse exponentially small so for $k = \text{poly}(n)$, this amplification is roughly from $1/q$ to $e^{-k/2q}$. Picking $k = 2q$, gives us an amplification of $1/q$ to $1 - 1/e \approx 0.63$.

3. More sophisticated gap amplification schemes have been studied [Has07, ALV12, AKLV13, AHS20, AAG21, AN22, ABN23] that involve applying some polynomial f to $\mathbf{H}$. How can we bound $\|f(\mathbf{H})\|_{\mathcal{P},1}$, in terms of $\|\mathbf{H}\|_{\mathcal{P},1}$ and properties of f?

4. Is there a template for a 2-prover version of Natarajan-Vidick [NV18b]? That is, assuming any favorable conjecture in Hamiltonian complexity, can one find a two-prover succinct MIP^* for QMA, where the provers are efficient given copies of the ground state?

6 Acknowledgements

We acknowledge Alex Lombardi for finding the bug in [NV18a], as well as Tony Metger, Thomas Vidick, and Tina Zhang for many helpful discussions and sharing an unpublished result of theirs. We also thank Thomas Vidick for detailed comments on an early draft of this article. This work was partially completed while AN and CN were participants in the Simons Institute for the Theory of Computing workshop on *Quantum Algorithms, Complexity, and Fault Tolerance*.

References

[AAG21] Anurag Anshu, Itai Arad, and David Gosset. An area law for 2d frustration-free spin systems, 2021. doi:10.48550/ARXIV.2103.02492.

[AAV13] Dorit Aharonov, Itai Arad, and Thomas Vidick. Guest column: The quantum pcp conjecture. *SIGACT News*, 44(2):47–79, June 2013. doi:10.1145/2491533.2491549.

[ABN23] Anurag Anshu, Nikolas P. Breuckmann, and Chinmay Nirkhe. NLTS hamiltonians from good quantum codes. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1090–1096, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3564246.3585114.

[AHS20] Anurag Anshu, Aram W. Harrow, and Mehdi Soleimanifar. From communication complexity to an entanglement spread area law in the ground state of gapped local hamiltonians. *https://arxiv.org/abs/2004.15009*, 2020.

[AKLV13] Itai Arad, Alexei Kitaev, Zeph Landau, and Umesh Vazirani. An area law and sub-exponential algorithm for 1D systems, 2013. arXiv preprint arXiv: 1301.1162.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, May 1998. doi:10.1145/278298.278306.

[ALV12] Itai Arad, Zeph Landau, and Umesh Vazirani. Improved one-dimensional area law for frustration-free systems. *Physical Review B*, 85:195145, May 2012. doi:10.1103/PhysRevB.85.195145.

[AN02] Dorit Aharonov and Tomer Naveh. Quantum NP - A Survey, 2002, arXiv:arXiv:quant-ph/0210077.

[AN22] Anurag Anshu and Chinmay Nirkhe. Circuit Lower Bounds for Low-Energy States of Quantum Code Hamiltonians. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 6:1–6:22, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ITCS.2022.6.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of np. *J. ACM*, 45(1):70–122, January 1998. doi:10.1145/273865.273901.

[BFL91] László Babai, Lance Fortnow, and Carsten Lund. Non-deterministic exponential time has two-prover interactive protocols. *computational complexity*, 1(1):3–40, 1991. doi:10.1007/BF01200056.

[BHW24] Harry Buhrman, Jonas Helsen, and Jordi Weggemans. Quantum PCPs: on adaptivity, multiple provers and reductions to local hamiltonians, 2024, arXiv:2403.04841.

[CM16] Toby Cubitt and Ashley Montanaro. Complexity classification of local hamiltonian problems. *SIAM Journal on Computing*, 45(2):268–316, 2016, arXiv:https://doi.org/10.1137/140998287. doi:10.1137/140998287.

[Din07] Irit Dinur. The pcp theorem by gap amplification. *J. ACM*, 54(3):12–es, June 2007. doi:10.1145/1236457.1236459.

[FGL+96] Uriel Feige, Shafi Goldwasser, Laszlo Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *J. ACM*, 43(2):268–292, mar 1996. doi:10.1145/226643.226652.

[FV15] Joseph Fitzsimons and Thomas Vidick. A multiprover interactive proof system for the local hamiltonian problem. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ITCS '15, page 103–112, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2688073.2688094.

[GKR15] Shafi Goldwasser, Yael Tauman Kalai, and Guy N. Rothblum. Delegating computation: Interactive proofs for muggles. *J. ACM*, 62(4), sep 2015. doi:10.1145/2699436.

[Has07] Matthew B Hastings. An area law for one-dimensional quantum systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(08):P08024, 2007.

[H$\mathring{\text{o}}$1] Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, jul 2001. doi:10.1145/502090.502098.

[IKW02] Russell Impagliazzo, Valentine Kabanets, and Avi Wigderson. In search of an easy witness: exponential time vs. probabilistic polynomial time. *Journal of Computer and System Sciences*, 65(4):672–694, 2002. doi:https://doi.org/10.1016/S0022-0000(02)00024-7. Special Issue on Complexity 2001.

[IV12] Tsuyoshi Ito and Thomas Vidick. A multi-prover interactive proof for nexp sound against entangled provers. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 243–252, 2012. doi:10.1109/FOCS.2012.11.

[Ji17] Zhengfeng Ji. Compression of quantum multi-prover interactive proofs. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 289–302, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3055399.3055441.

[JNV+20a] Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen. MIP*=RE, 2020, arXiv:arXiv:2001.04383.

[JNV+20b] Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry S. Yuen. Quantum soundness of the classical low individual degree test. *ArXiv*, abs/2009.12982, 2020.

[NV17] Anand Natarajan and Thomas Vidick. A quantum linearity test for robustly verifying entanglement. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 1003–1015, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3055399.3055468.

[NV18a] Anand Natarajan and Thomas Vidick. Low-degree testing for quantum states, and a quantum entangled games PCP for QMA. In Mikkel Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 731–742. IEEE Computer Society, 2018. doi:10.1109/FOCS.2018.00075.

[NV18b] Anand Natarajan and Thomas Vidick. Retracted: Two-player entangled games are np-hard. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPICS.CCC.2018.20.

[NW19] Anand Natarajan and John Wright. *NEEXP is Contained in MIP**, page 510–518. IEEE, Nov 2019. Funding by NSF.

[NZ23] Anand Natarajan and Tina Zhang. Quantum free games. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1603–1616, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3564246.3585208.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, August 2012.

[Vid14] Thomas Vidick. CS286.2 lecture 1: The PCP theorem, hardness of approximation, and multi-player games. http://users.cms.caltech.edu/ vidick/teaching/286_qPCP/lecture1.pdf, 2014.

[Vid16] Thomas Vidick. Three-player entangled xor games are np-hard to approximate. *SIAM Journal on Computing*, 45(3):1007–1063, 2016, arXiv:https://doi.org/10.1137/140956622. doi:10.1137/140956622.