NEEXP is contained in MIP*

Anand Natarajan
Computing and Mathematical Sciences
Caltech
Pasadena, USA
Email: anand@natarajans.org

John Wright
Computing and Mathematical Sciences
Caltech
Pasadena, USA
Email: jswright@mit.edu

***Abstract*—We study multiprover interactive proof systems. The power of classical multiprover interactive proof systems, in which the provers do not share entanglement, was characterized in a famous work by Babai, Fortnow, and Lund (Computational Complexity 1991), whose main result was the equality MIP = NEXP. The power of quantum multiprover interactive proof systems, in which the provers are allowed to share entanglement, has proven to be much more difficult to characterize. The best known lower-bound on MIP* is NEXP ⊆ MIP* due to Ito and Vidick (FOCS 2012). As for upper bounds, MIP* could be as large as RE, the class of recursively enumerable languages.**

The main result of this work is the inclusion NEEXP = NTIME[$2^{2^{\mathrm{poly}(n)}}$] ⊆ MIP*. This is an exponential improvement over the prior lower bound and shows that proof systems with entangled provers are at least exponentially more powerful than classical provers. In our protocol the verifier delegates a classical, exponentially large MIP protocol for NEEXP to two entangled provers: the provers obtain their exponentially large questions by measuring their shared state, and use a classical PCP to certify the correctness of their exponentially-long answers. For the soundness of our protocol, it is crucial that each player should not only sample its own question correctly but also avoid performing measurements that would reveal the *other* player's sampled question. We ensure this by commanding the players to perform a complementary measurement, relying on the Heisenberg uncertainty principle to prevent the forbidden measurements from being performed.

I. Introduction

This paper is about the complexity class MIP* of multiprover interactive proof systems with entangled quantum provers—the quantum version of the classical class MIP. Classically, the study of MIP has had far-reaching implications in theoretical computer science. In complexity theory, the proof by Babai, Fortnow, and Lund [1] that MIP = NEXP was the direct antecedent of the PCP theorem [2], [3], a seminal result which is the foundation of the modern theory of hardness of approximation. In cryptography, the MIP model was introduced to allow for information-theoretic zero-knowledge proofs [4], and more recently MIP protocols have become essential building blocks in designing delegated computation schemes (see e.g. [5]). These implications alone would be a sufficient motivation for considering the quantum class MIP*, but remarkably, the study of MIP* is also deeply related to long-standing questions in the foundations of quantum mechanics regarding the nature of quantum entanglement. Indeed, the MIP* model itself was anticipated by the *nonlocal games* or *Bell tests* introduced in the work of John Bell [6], who was in turn inspired by the thought experiment proposed by Einstein, Podolsky, and Rosen [7]. These nonlocal games have had applications to quantum cryptography [8], [9], [10], delegated quantum computation [11], and more.

Even though the class MIP is now well-understood, it has proven difficult to determine the computational power of MIP*. A priori, it is not even clear that MIP* contains MIP, since adding entanglement could increase or decrease the power of the proof system. This is because the added resource of entanglement can make it easier for dishonest provers to cheat the verifier. Indeed, Cleve et al. [12] showed that for proof systems based on so-called XOR games (where the verifier's decision can only depend on the XOR of the provers' answer bits), the quantum class ⊕MIP* ⊆ EXP, whereas classically ⊕MIP = NEXP. In particular, this result implied that the classical ⊕MIP protocol for NEXP of Håstad [13] could not be sound against entangled provers. In spite of this, Ito and Vidick [14], [15] were able to show that NEXP ⊆ MIP*, by proving that a different classical protocol *is* sound against entanglement. Note that the protocol of [15] is *identical* to a protocol shown to be unsound by Cleve et al., except in that it uses 3 provers rather than 2 (the protocol is played by choosing a random subset of 2 provers from the 3). This illustrates the subtleties of dealing with entangled provers.

With the lower bound NEXP ⊆ MIP* established, a natural follow-up question is whether MIP* is *strictly* more powerful than MIP. Indeed, it was long known that some MIP* protocols possess a uniquely quantum property called *self-testing*, which has no direct analog

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00039

in the classical setting. Roughly speaking, an MIP^* protocol is a self-test for a particular entangled state $|\psi\rangle$ if only provers using states close to $|\psi\rangle$ can achieve close to optimal success in the protocol. In such a protocol, observing that the provers succeed with nearly optimal probability *certifies* that they share a state close to the target state $|\psi\rangle$. The germ of this idea came from the work of Bell [6], who studied the types of bipartite correlations that could be obtained from measuring an entangled state called the EPR state, which had been introduced by Einstein, Podolsky, and Rosen [7]. Bell gave a protocol where provers using the EPR state could succeed with a greater probability than purely classical provers, and subsequent works of Tsirelson [16], and Summers and Werner [17] showed that (a variant of) Bell's protocol certifies the EPR state in the sense of self-testing.

In order to prove stronger lower bonds on MIP^*, the post-Ito-Vidick phase of MIP^* research aimed to use this self-testing property to design protocols for problems in Hamiltonian complexity, the quantum analog of the theory of NP-completeness. In Hamiltonian complexity, the complexity class QMA plays the role of NP; it is the set of problems for which there exists a quantum witness state that can be efficiently checked by a polynomial-time quantum verifier. Problems in QMA seemed like a natural match for the powers of MIP^* as one could potentially construct a protocol for QMA by designing a self-test for accepting witness states of some QMA-complete problem. The connection between MIP^* and QMA was also well motivated from the point of view of the "quantum PCP" research program, which strives to find quantum analogues of the classical PCP theorem. In the classical setting, the PCP theorem can be viewed as a scaled-down version of $\mathsf{MIP}^* = \mathsf{NEXP}$, showing that there exists an MIP^* protocol for 3SAT (and thus for all of NP) with $O(\log(n))$-sized messages. Drawing inspiration from this, Fitzsimons and Vidick [18] stated a "quantum games PCP conjecture": that there should exist an MIP^* protocol with $\log(n)$-sized messages for the local Hamiltonian problem, and thus for the class QMA. This was proved by Natarajan and Vidick [19] in 2018 with a 7-prover protocol. Along the way to achieving this goal, [19] developed a highly efficient self-test for high-dimensional entangled states: their "quantum low-degree test" is a self-test for n EPR pairs with only $O(\log(n))$ communication.

Already, the result of [19] is strong evidence that $\mathsf{MIP}^* \neq \mathsf{MIP}$, since it is believed that $\mathsf{QMA} \neq \mathsf{NP}$. But, at the same time, several other works showed that even larger separations were possible in the regime of subconstant soundness gaps. Here there are results in two settings. For MIP^* with a soundness gap scaling inverse-exponentially (i.e. $1/\exp(n)$) in the instance size, Ji [20] showed a protocol for NEEXP: nondeterministic *doubly*-exponential time, and a subsequent work by Fitzsimons, Ji, Vidick, and Yuen [21] showed protocols for non-deterministic iterated exponential time (e.g. $\mathsf{NTIME}(2^{2^n})$) with a correspondingly small soundness gap (e.g. $2^{-C\cdot 2^n}$). In the "gapless" case, Slofstra [22], [23] showed that given a description of an MIP^* protocol, determining whether there exists an entangled strategy that succeeds with probability exactly 1 is undecidable by any Turing machine.

These results hint at the full power of MIP^* but are not conclusive, as it is not unusual for quantum complexity classes to increase significantly in power when a numerical precision parameter is allowed to shrink. For instance, QIP (quantum interactive proofs with a single prover) with an exponentially small gap is equal to EXP [24], while QIP with a polynomial gap is equal to $\mathsf{IP} = \mathsf{PSPACE}$. Likewise, QMA with exponentially small gap (known as PreciseQMA) is known to be equal to PSPACE [25], while QMA is contained PP, and $\mathsf{QMA}(k)$ (QMA with multiple unentangled Merlins) with exponentially small gap is equal to NEXP [26], whereas in the constant-gap regime the best known lower bound is that $\mathsf{QMA}(k) \supseteq \mathsf{QMA}$. Moreover, even the QMA lower bound for $\mathsf{MIP}^*_{\log}$ obtained by [27] holds for 7 provers only; with 2 provers, the best known lower bound for $\mathsf{MIP}^*_{\log}$ is $\mathsf{NP} = \mathsf{MIP}_{\log}$ [19]. Could it be that 2-prover MIP^* is equal to MIP, with entanglement providing no advantage at all?

This paper conclusively answers this question in the negative. Our main result (Theorem I.1) is to show that MIP^* contains NEEXP, with only two provers and with a constant completeness-soundness gap. This is establishes the first known unconditional separation between MIP^* and MIP in the constant-gap regime: previously, such a separation was known only assuming $\mathsf{QMA} \neq \mathsf{NP}$, and only in the scaled-down setting of logarithmic-sized messages.

Theorem I.1 (Theorem 17.12 in the full version)**.** *There is a two-prover, one-round* MIP^* *protocol for the* NEEXP*-complete problem* Succinct-Succinct-3Sat *with completeness* 1*, soundness* $1/2$*, and question and answer length* $\mathrm{poly}(n)$.

As a corollary of Theorem I.1, we obtain a lower bound on the hardness of approximation for the entangled value ω^* of a nonlocal game.

Corollary I.2. *There exists a constant* $c < 1$ *such that given a two-prover nonlocal game* $\mathscr{G}$ *of size* N*, the problem of deciding whether* $\omega^*(\mathscr{G}) = 1$ *or* $\omega^*(\mathscr{G}) \leq 1/2$*, promised one of the two holds, is* $\mathsf{NTIME}(2^{N^{\log^{-c} N}})$*-hard.*

For two-player games, the best prior lower bound was

NP [27]. The lower bound achieved in Corollary I.2 is stronger as for any $c < 1$, the function $2^{N^{\log^{-c} N}}$ is superpolynomial.

Techniques.: Our construction, inspired by [20] and [21], involves "compression": we show how to take an MIP protocol for NEEXP with exponentially-long questions and answers (the "big" protocol), and simulate it by an MIP^* protocol with polynomial-sized messages (the "small" protocol). However, the techniques we use to achieve our compression are quite different. We eschew the Hamiltonian-complexity ideas that were used in previous works, and in particular the use of history states. In our protocol, honest provers need only share a quantum resource state of (exponentially many) EPR pairs, together with a *classical* assignment to the NEEXP instance being tested. The use of history states was the main barrier preventing previous works from applying to the case of two provers.

We divide compression into two steps: *question compression* and *answer compression*. We achieve question compression by a technique which we call *introspection*, in which we command the provers to perform measurements on their shared EPR pairs whose outcomes are pairs of questions from the "big" protocol. To force the provers to sample their questions honestly, we use a variant of the quantum low-degree test from [19], which certifies Pauli measurements on exponentially many EPR pairs using messages of only polynomial size. A crucial challenge is to prevent each prover from learning the other prover's sampled question, since this would destroy the soundness of the "big" protocol. To achieve this, we use the *"data-hiding"* properties of quantum measurements in incompatible bases: if a set of qubits is measured in the Pauli X-basis, this "erases" all information about Z-basis measurements. This means that if Alice samples her question by measuring her half of a block of EPR pairs in the Z-basis, then her question can be hidden from Bob by forcing him (via self-testing) to measure his half of the EPR pairs in the X-basis. Interestingly, our data-hiding scheme does *not* operate in a black-box way on the "big" protocol, but rather makes essential use of its structure. In particular, we start with a "big" protocol based on a scaled-up version of a PCP construction using the low-degree test, where the question distribution consists of pairs of random points in a vector space and affine subspaces containing them. The linear structure of the vector space is essential for our data-hiding procedure to work.

Our approach to answer compression is more standard, essentially using composition with a classical PCP of proximity. Here, the verifier asks the provers to compute a PCP proof that their "big" answers satisfy the success conditions of the protocol, and verifies this PCP proof by reading an exponentially smaller number of bits. Care is needed to deal with entanglement between the provers. The first, fundamental challenge we face is that the success condition of the "big" protocol is a function of *both* provers' answers. Thus, to compute a PCP proof that the condition is satisfied, one of the provers must have access to both provers' answers. Classically, this is achieved using the technique of oracularization, in which one prover receives *both* provers' questions and is checked for consistency against the other prover, which only receives a single question. In the entangled setting, this oracularization procedure is sound, but not necessarily complete. This is because oracularization requires that each prover, if given the *other* prover's question, could predict its answer with certainty, even though this answer is obtained from a nondeterministic quantum measurement. In our protocol, we are able to use oracularization because honest provers always use a maximally entangled state, which they measure with projective measurements that pairwise commute for every pair of questions asked in the game. While this commutation requirement is restrictive, it still permits non-trivial quantum behavior; indeed, the linear system games used by Slofstra [23] involve similar commutation conditions.

The second challenge is to ensure that the PCP of proximity we use for composition is itself sound against entanglement. We achieve this by performing a further step of composition: we ask the provers to encode their PCP proof in the low-degree code and verify it with the low-degree test, which is known to be sound even against entangled provers [27]. This technique was introduced in the QMA protocol of [19] in order to perform energy measurements on the provers' state.

Implications and future work: We believe that our work opens up several exciting directions for further progress. For the complexity theorist, the most obvious future direction is to obtain even stronger lower bounds on MIP^* by iterating our protocol, as in [21]. At the most basic level, we could imagine taking our MIP^* protocol for NEEXP and performing a further layer of question compression and answer compression on it, thus obtaining an MIP^* protocol with logarithmic message size for NEEXP, or, scaling up, an MIP^* protocol with polynomial message size for $\mathsf{NTIME}(2^{2^{2^{\mathrm{poly}(n)}}})$. By further iterating question reduction and answer reduction k times, we could obtain potentially obtain lower bounds of $\mathsf{NTIME}(\underbrace{2^{\cdot^{\cdot^{\cdot^{n}}}}}_{k})$ on MIP^* while retaining a constant completeness-soundness gap. The main obstacle to achieving such results is that the question compression procedure developed in this paper is tailored to a special distribution of questions (that of the $\mathsf{MIP}_{\mathrm{exp}}$ protocol for NEEXP), whereas our answer compression procedure produces protocols whose question distribution is not

of this form.

Assuming that this obstacle can be surmounted, we could aspire to a more ambitious goal: a general "gap-preserving compression procedure" for some subclass of MIP^* protocols, which we may label "compressible" protocols. Such a procedure would consist of a Turing machine that takes as input any compressible MIP^* protocol $\mathscr{G}$, and generates a new compressible protocol $\mathscr{G}'$ with exponentially smaller message size, but approximately the same entangled value. It was shown by [21] that the existence of such a compression procedure for the set of *all* MIP^* protocols would imply that MIP^* contains the set of all computable languages, and moreover that there exists an undecidable language in MIP^*. These consequences would continue to hold as long as the set of compressible protocols contains a family of protocols solving problems in $\mathsf{NTIME}(f(n))$, where $f(n)$ is a growing function of n.

Showing that MIP^* contains undecidable languages would be significant not just for complexity theory but also for the foundations of quantum mechanics, as it would resolve a long-standing open problem known as *Tsirelson's problem*. Tsirelson's problem asks whether two notions of quantum nonlocality are equivalent: the *tensor-product model*, in which two parties Alice and Bob each act on their respective factor of a tensor-product Hilbert space $\mathcal{H}_{\mathsf{Alice}} \otimes \mathcal{H}_{\mathsf{Bob}}$, and the *commuting-operator model*, in which both parties act on a common Hilbert space $\mathcal{H}$, but the algebra of Alice's measurement operators must commute with Bob's, and vice versa. It was shown by Slofstra [22] that in the "zero-error" setting, these two models differ: there are quantum correlations which can be *exactly* achieved in the commuting-operator model but not in the tensor product model. Surprisingly, showing that MIP^* contains undecidable languages would imply that the two models are separated even in the bounded-error setting: it would imply that there exist correlations that can be achieved in the commuting-operator model that cannot even be approximated (up to constant precision) in the tensor-product model. The reason for this implication is that if the two models are indistinguishable up to bounded error, then there exists a Turing machine that can decide any language in MIP^* and is guaranteed to halt. This observation, which is folklore in the community, follows from the completeness of the non-commutative sum of squares hierarchy for the commuting-operator model, as documented in [21]. Showing a separation between the two models would have significant mathematical consequences as well, as it would yield a negative answer to the long-standing Connes' embedding problem.

In addition to these connections to complexity and mathematical physics, we hope that our results will have applications in other areas such as to delegated computation or quantum cryptography. In particular, our use of introspection is reminiscent of ideas used in quantum randomness expansion, where randomness generated by measuring EPR pairs is used to generate questions for a nonlocal game. Could our results improve on the infinite randomness expansion protocol of Coudron and Yuen [28]?

II. Overview of our proof

In this section we give a more detailed overview of the technical parts of the paper.

A. Basic quantum notation and qudits

While the full version of the paper contains a more complete set of quantum preliminaries in Section 4, for the purposes of this introduction we define some basic notation, aimed at the reader who is familiar with the standard quantum computing formalism over qubits but is less familiar with *qudits*: quantum systems of dimension not equal to 2. In this paper, we make extensive use of such qudits: in particular, for a finite field $\mathbb{F}_Q$, we will consider qudits of dimension Q, with a basis state $|i\rangle$ for every element $i \in \mathbb{F}_Q$. Under tensor product, we obtain a basis for the space of M qudits of dimension Q where each basis state $|x\rangle$ corresponds to a vector $x \in \mathbb{F}_Q^M$.

The basic resource state used in our protocols will be the EPR state over $2M$ qudits of dimension Q. The qudits are split into two registers of M qudits each, held by the two provers Alice and Bob, respectively.

$$|\mathrm{EPR}_Q^M\rangle = \frac{1}{\sqrt{Q^M}} \sum_{x \in \mathbb{F}_Q^M} |x\rangle_{\mathsf{Alice}} \otimes |x\rangle_{\mathsf{Bob}} .$$

This state is a *maximally entangled* state between Alice and Bob.

Acting on this state, we will ask the provers to perform measurements from a special class called *Pauli basis measurements*. To define these over a general field $\mathbb{F}_Q$ requires the introduction of some finite field technology, in particular the finite field trace function. For simplicity, in this overview we will imagine that Q is prime, allowing the addition in $\mathbb{F}_Q$ to be identified with the additive group $\mathbb{Z}_Q$, and simplifying the definition of the Paulis; in the full version of the paper, we will work with Q a power of 2. For a single qudit of dimension Q, the Pauli X and Z bases are the sets $\{|\tau_u^X\rangle\}_{u \in \mathbb{F}_Q}$ and $\{|\tau_u^Z\rangle\}_{u \in \mathbb{F}_Q}$ of vectors

$$|\tau_u^X\rangle = \frac{1}{\sqrt{Q}} \sum_{x \in \mathbb{F}_Q} \omega^{xu} |x\rangle , \quad |\tau_u^Z\rangle = |u\rangle ,$$

where $\omega = \exp(2\pi i/Q)$ is the Q-th root of unity. We denote the projectors onto these basis states by τ_u^X and τ_u^Z, respectively. For a system of M qudits, the Pauli X and Z observables are a set of *generalized observables*

indexed by elements of $\mathbb{F}_Q^M$: a generalized observable is a Hermitian matrix with eigenvalues that are Q-th roots of unity. They are given by

$$X(v) = \sum_{u \in \mathbb{F}_Q^M} \omega^{u \cdot v} \tau_{u_1}^X \otimes \ldots \otimes \tau_{u_M}^Z,$$

$$Z(v) = \sum_{u \in \mathbb{F}_Q^M} \omega^{u \cdot v} \tau_{u_1}^Z \otimes \ldots \otimes \tau_{u_M}^Z,$$

where $u_1, \ldots, u_M$ are the components of the vector u, and $u \cdot v$ is the dot product $\sum_{i=1}^M u_i \cdot v_i$. Measuring a generalized observable means performing a projective measurement onto the eigenvectors of the observable, with the outcome a corresponding to the eigenvector with eigenvalue ω^a.

B. Our starting point: a classical interactive proof for NEEXP

We start with a classical multiprover interactive proof protocol for NEEXP. The equality MIP = NEXP was originally shown by Babai, Fortnow, and Lund [1] using a protocol based on the *multilinearity test*: the idea is that an exponentially-long witness for a problem in NEXP is encoded in the truth-table of a multivariate polynomial function over a finite field, which is linear in each of the variables individually. The verifier is able to verify the witness by evaluating the multilinear polynomial over appropriately chosen points and subspaces. To scale up to NEEXP, we use a much more efficient version of the same idea, replacing the multilinearity test with the *low-degree test*, which works with multivariate polynomials of low total degree. This more efficient construction comes from the PCP literature. We give a relatively self-contained presentation of the protocol in Section 11. For the purposes of this overview, it is sufficient to know the following: any problem in NEEXP can be reduced to satisfiability for a doubly exponentially long 3Sat formula, succinctly encoded by a polynomial-sized circuit. (We refer to this problem as Succinct-Succinct-3Sat). Given a 3Sat formula ψ, we would like the provers to prove to us that they have a satisfying assignment a to this formula. Instead of reading the assignment directly, we will ask the provers to encode their assignment as a multivariate polynomial $g_a : \mathbb{F}_Q^M \to \mathbb{F}_Q$, where the number of variables M and the finite field size Q are appropriately chosen parameters, and return evaluations of this polynomial. To check that a satisfies ψ, the verifier first uses a technique called arithmetization to convert the formula ψ into a multivariate polynomial $g_\psi : \mathbb{F}_Q^{3M+k} \to \mathbb{F}_Q$. The polynomial g_ψ is chosen such that the assignment a satisfies ψ if and only if the expression

$$\begin{aligned}\mathrm{sat}_{\psi,a}(x,b,w) := g_\psi(x,b,w) \cdot (g_a(x_1) - b_1) \\ \cdot (g_a(x_2) - b_2)(g_a(x_3) - b_3)\end{aligned}$$

is equal to 0 at every point in a particular subset $H \subseteq \mathbb{F}_Q^{3M+k}$. Our classical protocol for NEEXP checks this condition:

Informal Theorem II.1 (Section 11 in the full version)**.** *There exists a protocol $\mathscr{G}_0$ for* Succinct-Succinct-3Sat *(and hence* NEEXP*), where the verifier's questions to the provers are constant-dimension subspaces of $\mathbb{F}_Q^M$, and the provers' responses are evaluations of degree-D M-variate polynomials on these subspaces. The parameters M, Q, D are all chosen to be $\exp(n)$, and hence the question and answer lengths as well as the runtime of the verifier in this protocol are $\exp(n)$.*

The distribution over subspaces sent to the provers in $\mathscr{G}_0$ is relatively simple, and in fact is independent of the instance of Succinct-Succinct-3Sat being tested. For the purposes of this overview, the reader can take the distribution over pairs of questions to be the *plane-point distribution* $\mathcal{D}$. A pair $(\boldsymbol{s}, \boldsymbol{u}) \sim \mathcal{D}$ consists of a uniformly random affine plane $\boldsymbol{s} \subseteq \mathbb{F}_Q^M$, which is sent to Alice, and a uniformly random point $\boldsymbol{u} \in \boldsymbol{s}$ which is sent to Bob. The full distribution over questions in $\mathscr{G}_0$ is more complicated than this but the essential ideas of our protocol will be illustrated by restricting to this case.

C. Restricting the strategies: registers and compilers

One of the main challenges in working with entangled provers is showing soundness against general entangled strategies. An important technique in this area is to force the provers to use a particular state and class of measurements by playing a type of game known as a *self-test*.

Informal Definition II.2. A game $\mathscr{G}_{\text{test}}$ is a *self-test* for a state $|\psi\rangle$ and measurements M^x if any strategy that succeeds in $\mathscr{G}_{\text{test}}$ with probability $1-\epsilon$ must use a state $|\psi'\rangle$ and measurements $(M')^x$ that are $\delta(\epsilon)$-close, in the appropriate metric, to $|\psi\rangle$ and M^x.

Some of the earliest self-tests include the famous CHSH game, which self-tests the Pauli X and Z operators on a single EPR pair (of qubits). Self-testing technology has greatly advanced over the years, and in this paper we design a highly efficient self-test based on the low-degree test of [19].

Informal Theorem II.3 (Theorem 6.2 in the full version)**.** *The Pauli basis test $\mathrm{Pauli}(n, q)$ is a self-test for the state $|\mathrm{EPR}_q^n\rangle$ and the Pauli X and Z basis measurements. This test sends the players questions of length $O(\log(n))$ and receives answers of length $O(\mathrm{poly}(n))$.*

The Pauli X and Z measurements are "complete" measurements, and as a consequence, there is no non-

trivial measurement on a set n qudits that can be measured jointly with both the Pauli X and Z measurements on those qudits. Using this property, we design a game called the *data-hiding game*, which certifies that a prover's measurements act trivially on a specified set of qudits.

Informal Theorem II.4 (Theorem 8.3 in the the full version)**.** *The data-hiding game $\mathscr{G}_{\mathrm{hide}}$ is a self test for states $|\psi\rangle = |\mathrm{EPR}_q^n\rangle \otimes |\mathrm{aux}\rangle$ and measurements M^x of the form $M^x = I \otimes (M')^x_{\mathsf{aux}}$. It has questions of length $O(\log(n))$ and answers of length $O(\mathrm{poly}(n))$.*

Together, the Pauli basis test and the data-hiding game allow us to restrict our analysis of our protocols to a class of strategies we call *register strategies*: strategies for which the shared state is a collection of ℓ registers, each in an EPR state, together with some auxiliary register:

$$|\psi\rangle = |\mathrm{EPR}_{q_1}^{n_1}\rangle \otimes \ldots \otimes |\mathrm{EPR}_{q_\ell}^{n_\ell}\rangle \otimes |\mathrm{aux}\rangle\,,$$

and where the provers can be commanded to perform either (1) Pauli basis measurements on specified subsets of the registers, or (2) measurements that do *not* act on specified subset of the EPR registers (but act on the auxiliary register or the remaining EPR registers). We formalize this by designing a *compiler*, which takes in a protocol $\mathscr{G}$ that is complete and sound for register strategies, and produces a new protocol $\mathscr{G}'$ which is complete and sound over all strategies.

Informal Theorem II.5 (Theorem 7.2 and Theorem 9.2 in the full version)**.** *Suppose $\mathscr{G}$ is a protocol for a computation problem for which completeness and soundness hold for register strategies, with $O(1)$ many registers of size n. (That is, for YES instances of the problem, there exists a register strategy achieving value 1, and for NO instances, no register strategy achieves value greater than $1/2$). Let the questions in $\mathscr{G}$ be of length Q and the answers be of length A. Then there exists a protocol $\mathscr{G}'$ which is complete and sound for general strategies, and for which the question length is $Q + \log(n)$ and the answer length is $A + \mathrm{poly}(n)$.*

The compiled protocol $\mathscr{G}'$ either runs the original protocol $\mathscr{G}$, or, with some probability, runs the Pauli basis test, the data-hiding game, or a consistency test.

D. Question reduction through introspection

With our compiler in place, we have now given the verifier the power to command the provers to perform Pauli basis measurements on a set of EPR pairs. We would like to use this to reduce the question size of the classical protocol $\mathscr{G}_0$ for NEEXP described above from $\exp(n)$ to $\mathrm{poly}(n)$. We will do so by forcing the provers, rather than the verifier, to sample the protocol's $\exp(n)$-length questions, a technique we call "introspection". That is, we would like to force the provers to sample pairs $(\boldsymbol{s}, \boldsymbol{u})$ from the plane-vs-point distribution $\mathcal{D}$, where $\boldsymbol{s}$ is a uniformly random affine plane in $\mathbb{F}_Q^M$, and $\boldsymbol{u}$ a uniformly random point on $\boldsymbol{s}$.

To design a scheme to sample from this distribution, let us first fix a representation of affine planes. We will represent an affine plane by an *intercept* $u \in \mathbb{F}_Q^M$ and two *slopes* $v_1, v_2 \in \mathbb{F}_Q^M$. The plane given by u, v_1, v_2 is the set $s_u^v = \{u + \lambda_1 v_1 + \lambda_2 v_2 : \lambda_1, \lambda_2 \in \mathbb{F}_Q\}$. As a first attempt, we may try the following scheme:

1) Alice and Bob share three registers, each of which contains an EPR state, so their shared state is
$$|\psi_0\rangle = |\mathrm{EPR}_Q^M\rangle_{R_0} \otimes |\mathrm{EPR}_Q^M\rangle_{R_1} \otimes |\mathrm{EPR}_Q^M\rangle_{R_2}\,.$$
2) Alice first measures her half of registers R_1 and R_2 in the Pauli Z-basis, to obtain uniformly random outcomes $\boldsymbol{v}_1, \boldsymbol{v}_2$. The shared state is now
$$|\psi_1\rangle = |\mathrm{EPR}_Q^M\rangle_{R_0} \otimes (|\boldsymbol{v}_1\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_1\rangle_{\mathsf{Bob}})_{R_1} \otimes (|\boldsymbol{v}_2\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_2\rangle_{\mathsf{Bob}})_{R_2}.$$
3) Now, Alice and Bob both measure register R_0 in the Pauli Z-basis, both obtaining the same outcome $\boldsymbol{u}$. The shared state is now
$$|\psi_2\rangle = (|\boldsymbol{u}\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{u}\rangle_{\mathsf{Bob}})_{R_0} \otimes (|\boldsymbol{v}_1\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_1\rangle_{\mathsf{Bob}})_{R_1} \otimes (|\boldsymbol{v}_2\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_2\rangle_{\mathsf{Bob}})_{R_2}.$$
Alice sets her plane $\boldsymbol{s}$ to be $s_{\boldsymbol{u}}^{\boldsymbol{v}}$ and Bob sets his point to be $\boldsymbol{u}$.

Indeed, the pair $(\boldsymbol{s}, \boldsymbol{u})$ generated by this procedure is distributed according to $\mathcal{D}$. However, there is a problem: through her measurement, Alice obtains additional side information, specifically the value of Bob's point $\boldsymbol{u}$. Can we command Alice to erase the side information? In fact, we can, using the *Heisenberg uncertainty principle*: if two observables anticommute, then measuring one completely destroys information about the other. Using this idea, we modify our protocol as follows:

1) As above.
2) As above. At this point, applying the definition of $|\mathrm{EPR}_Q^M\rangle$, we can write the shared state as
$$|\psi_1\rangle \propto \sum_{u \in \mathbb{F}_Q^M} (|u\rangle_{\mathsf{Alice}} \otimes |u\rangle_{\mathsf{Bob}})_{R_0} \otimes (|\boldsymbol{v}_1\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_1\rangle_{\mathsf{Bob}})_{R_1} \otimes (|\boldsymbol{v}_2\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_2\rangle_{\mathsf{Bob}})_{R_2}.$$
3) **New:** Intuitively, we would like Alice to be *prevented* from measuring the component of the intercept along the directions $\boldsymbol{v}_1, \boldsymbol{v}_2$. This information would be obtained by measuring the

observables[1] $Z(\boldsymbol{v}_1), Z(\boldsymbol{v}_2)$. To destroy it, we will ask Alice to measure the *complementary* Pauli observables $X(\boldsymbol{v}_1), X(\boldsymbol{v}_2)$ on register R_0, obtaining outcomes $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{F}_Q$. The shared state is now

$$|\psi_2'\rangle \propto \sum_{u,\lambda,\mu} \left(\omega^{\boldsymbol{\alpha}_1\lambda+\boldsymbol{\alpha}_2\mu} |\underbrace{u+\lambda\boldsymbol{v}_1+\mu\boldsymbol{v}_2}_{u'}\rangle_{\mathsf{Alice}} |u\rangle_{\mathsf{Bob}}\right)_{R_0} \otimes (|\boldsymbol{v}_1\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_1\rangle_{\mathsf{Bob}})_{R_1} \otimes (|\boldsymbol{v}_2\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_2\rangle_{\mathsf{Bob}})_{R_2}.$$

where, as above, $\omega = \exp(2\pi i/Q)$ is a Q-th root of unity. Alice and Bob's state on R_0 is now a uniform superposition over pairs u, u' of points lying on the same affine subspace with slopes $\boldsymbol{v}_1, \boldsymbol{v}_2$.

4) Alice and Bob both measure register R_0 in the Z basis, obtaining outcomes $\boldsymbol{u}$ and $\boldsymbol{u}'$, respectively. The shared state is now

$$|\psi_3'\rangle = (|\boldsymbol{u}\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{u}'\rangle_{\mathsf{Bob}})_{R_0} \otimes (|\boldsymbol{v}_1\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_1\rangle_{\mathsf{Bob}})_{R_1} \otimes (|\boldsymbol{v}_2\rangle_{\mathsf{Alice}} \otimes |\boldsymbol{v}_2\rangle_{\mathsf{Bob}})_{R_2}.$$

Alice sets her plane to be $s^{\boldsymbol{v}}_{\boldsymbol{u}}$ and Bob sets his point to be $\boldsymbol{u}'$.

Now, from the calculation performed above, it's clear that Bob's point $\boldsymbol{u}'$ is uncorrelated with Alice's intercept $\boldsymbol{u}$, apart from lying in the plane $s^{\boldsymbol{v}}_{\boldsymbol{u}}$, and hence there is no further information about Bob's point that Alice can learn by measuring her portion of the final state $|\psi_3'\rangle$. But Alice still obtains some additional information from her measurements along the way, in particular the outcomes α_1, α_2 of the X measurements. And moreover, how can we certify that the X measurements were performed correctly, since they are not Pauli basis measurements as given to us by the compiler? To answer these questions, we define a new game called the *partial data-hiding game* (Theorem 10.4 in the full version), which certifies that Alice and Bob perform the steps described above and that no extra information is leaked. Building on this game, we can now design a protocol for NEEXP with small question size:

Informal Theorem II.6 (Theorem 15.8 in the full version)**.** *There is an* MIP^* *protocol* $\mathscr{G}_1$ *for* NEEXP *with questions of length* $\mathrm{poly}(n)$*, and answers of length* $\exp(n)$*. The verifier can generate the questions in* $\mathrm{poly}(n)$ *time but needs* $\exp(n)$ *time to verify the answers.*

[1]Strictly speaking, this is only true when $\boldsymbol{v}_1 \cdot \boldsymbol{v}_1 \neq 0$ and $\boldsymbol{v}_2 \cdot \boldsymbol{v}_2 \neq 0$. A more rigorous treatment of this is given in Section 10 of the full version.

E. Answer reduction through PCP composition

We have succeeded in obtaining a game with short questions, but the answers are now exponentially long. In the last step, we will use composition with a classical probabilistically checkable proof (PCP) to delegate verification of the answers to the provers.

Schematically, the protocol $\mathscr{G}_1$ consists of the following steps:

1) The verifier sends Alice a question $\boldsymbol{x}$ and Bob a question $\boldsymbol{y}$.
2) Alice returns an (exponentially-long) answer $\boldsymbol{A}$ and Bob an exponetially-long answer $\boldsymbol{B}$.
3) The verifier computes a verification predicate $V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{A}, \boldsymbol{B})$ in exponential time.

We would like to delegate the last step to the provers by asking them to compute a PCP proof that $V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{A}, \boldsymbol{B}) = 1$, which the verifier can check by communicating only polynomially many bits with the provers. However, we face an obstacle: Alice cannot know $\boldsymbol{y}$ and $\boldsymbol{B}$, and neither can Bob know $\boldsymbol{x}$ and $\boldsymbol{A}$, and distributed PCPs (where neither party knows the entire assignment) are known to be impossible [29]. To proceed, we will first have to modify $\mathscr{G}_1$ by *oracularizing* it:

1) The verifier sends Alice the questions $\boldsymbol{x}, \boldsymbol{y}$, and Bob either $\boldsymbol{x}$ or $\boldsymbol{y}$, chosen uniformly at random.
2) Alice returns exponentially-long answers $\boldsymbol{A}, \boldsymbol{B}$, and Bob returns an answer $\boldsymbol{C}$.
3) The verifier computes a verification predicate $V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{A}, \boldsymbol{B})$ on Alice's questions and answers, and further checks that $\boldsymbol{A} = \boldsymbol{C}$, if Bob received $\boldsymbol{x}$, or that $\boldsymbol{B} = \boldsymbol{C}$, if Bob received $\boldsymbol{y}$.

The idea is that the new Alice simulates both Alice and Bob from the original protocol, and the new Bob certifies that the new Alice does not take advantage of her access to both questions to cheat. It is well-known that oracularization does not harm the soundness of interactive protocols, be they classical or quantum. However, in the quantum world, it is not necessarily the case that the oracularized protocol retains *completeness*. This is because Alice and Bob may have been asked to perform non-compatible measurements in the original protocol, rendering it impossible for the new Alice to simulate both the original Alice and Bob. Fortunately for us, the honest strategy for protocol $\mathscr{G}_1$ is such that completeness under oracularization.

Now that a single prover is in possession of all inputs to the verification predicate V, we can implement our idea of using a PCP proof. Classically, this idea is known as PCP *composition*, and is extensively used in the PCP literature. In the quantum case, the requirement to maintain soundness against entanglement makes composition technically difficult, and we defer

the details to Part V of the full version. Once the composition is performed, we reach our main result.

Informal Theorem II.7 (Theorem 17.12 in the full version)**.** *There is an* MIP^* *protocol* $\mathscr{G}_2$ *for* Succinct-Succinct-3Sat *(and hence for* NEEXP*) with question size, answer, and verifier runtime* $\mathrm{poly}(n)$.

F. Organization

The full version of this paper is organized into five parts. The first part is the introduction and this overview. The remaining parts are organized as follows.

- Part II contains two sections of preliminaries, one containing the classical background and another the quantum background.
- Part III contains the register compiler, i.e. the proof of Informal Theorem II.5. This involves designing the Pauli basis test (Section 6) and the data hiding test (Section 8). Section 5 serves as an introduction to this part and contains more details on the organization.
- Part IV contains the "introspection" question reduction step, i.e. the proof of Informal Theorem II.6. To begin, we sketch the classical MIP protocol for Succinct-3Sat in Section 11. Then we give the introspected, i.e. "big", low-degree test in Section 13, and finish by giving the entire small-question NEEXP protocol in Section 15. Section 14 contains a test necessary for the protocol called the "intersecting lines test". It allows us carry over the results of the low-degree test from one register to another.
- Part V contains the answer reduction, i.e. the proof of Informal Theorem II.7. The construction involves composing PCP protocols with error-correcting codes, and so Section 16 surveys the properties we need of an error-correcting code. Finally, Section 17 contains the actual proof of the answer reduction step.

Acknowledgement

We thank Henry Yuen for many useful conversations about the idea of "introspecting" interactive proof protocols, which inspired us to start this project. AN is also grateful to the Simons Institute for the hospitable environment of the Summer Cluster on Challenges in Quantum Computation during which these conversations where held. We thank Thomas Vidick for his guidance and advice. We thank Ryan O'Donnell and Ryan Williams for a succinct review of the literature on the complexity of succinct (succinct) 3Sat and NE(E)XP. We are also grateful to Zhengfeng Ji for several useful discussions, especially regarding the consequences of recursively composing our protocol with itself.

AN was partially supported by NSF grant CCF-1452616. JW was partially supported by ARO contract W911NF-17-1-0433. Both authors acknowledge funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant PHY-1733907).

References

[1] L. Babai, L. Fortnow, and C. Lund, "Non-deterministic exponential time has two-prover interactive protocols," *Computational complexity*, vol. 1, no. 1, pp. 3–40, 1991. I, II-B

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, "Proof verification and the hardness of approximation problems," *Journal of the ACM*, vol. 45, no. 3, pp. 501–555, 1998. I

[3] S. Arora and S. Safra, "Probabilistic checking of proofs: a new characterization of NP," *Journal of the ACM*, vol. 45, no. 1, pp. 70–122, 1998. I

[4] M. Ben-Or, S. Goldwasser, J. Kilian, and A. Wigderson, "Multi-prover interactive proofs: How to remove intractability assumptions," in *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, 1988, pp. 113–131. I

[5] Y. Kalai, R. Raz, and R. Rothblum, "How to delegate computations: the power of no-signaling proofs," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2014, pp. 485–494. I

[6] J. Bell, "On the Einstein Podolsky Rosen paradox," *Physics*, vol. 1, no. 3, pp. 195–200, 1964. I

[7] A. Einstein, B. Podolsky, and N. Rosen, "Can quantum-mechanical description of physical reality be considered complete?" *Physical review*, vol. 47, no. 10, p. 777, 1935. I

[8] A. Ekert, "Quantum cryptography based on Bell's theorem," *Physical review letters*, vol. 67, no. 6, p. 661, 1991. I

[9] D. Mayers and A. Yao, "Quantum cryptography with imperfect apparatus," in *Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science*, 1998, pp. 503–509. I

[10] R. Colbeck, "Quantum and relativistic protocols for secure multi-party computation," Ph.D. dissertation, University of Cambridge, 2006. I

[11] B. Reichardt, F. Unger, and U. Vazirani, "A classical leash for a quantum system: Command of quantum systems via rigidity of CHSH games," *Nature*, vol. 496, pp. 456–460, 2013. I

[12] R. Cleve, P. Hoyer, B. Toner, and J. Watrous, "Consequences and limits of nonlocal strategies," in *Proceedings of the 19th Annual IEEE Conference on Computational Complexity*, 2004, pp. 236–249. I

[13] J. Håstad, "Some optimal inapproximability results," in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1997, pp. 1–10. I

[14] T. Ito and T. Vidick, "A multi-prover interactive proof for NEXP sound against entangled provers," in *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, 2012, pp. 243–252. I

[15] T. Vidick, "Three-player entangled XOR games are NP-hard to approximate," *SIAM Journal on Computing*, vol. 45, no. 3, pp. 1007–1063, 2016. I

[16] B. Tsirelson, "Quantum generalizations of Bell's inequality," *Letters in Mathematical Physics*, vol. 4, no. 2, pp. 93–100, 1980. I

[17] S. Summers and R. Werner, "Maximal violation of Bell's inequalities for algebras of observables in tangent space-time regions," *Annales de l'IHP Physique théorique*, vol. 49, no. 2, pp. 215–243, 1988. I

[18] J. Fitzsimons and T. Vidick, "A multiprover interactive proof system for the local Hamiltonian problem," in *Proceedings of the 6th Innovations in Theoretical Computer Science*, 2015, pp. 103–112. I

[19] A. Natarajan and T. Vidick, "Low-degree testing for quantum states, and a quantum entangled games PCP," in *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science*, 2018. I, I, II-C

[20] Z. Ji, "Compression of quantum multi-prover interactive proofs," in *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, 2017, pp. 289–302. I, I

[21] J. Fitzsimons, Z. Ji, T. Vidick, and H. Yuen, "Quantum proof systems for iterated exponential time, and beyond," in *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, 2019. I, I, I

[22] W. Slofstra, "Tsirelson's problem and an embedding theorem for groups arising from non-local games," 2016, technical report, arXiv:1606.03140. I, I

[23] ——, "The set of quantum correlations is not closed," in *Forum of Mathematics, Pi*, vol. 7, 2019, p. e1. I, I

[24] T. Ito, H. Kobayashi, and J. Watrous, "Quantum interactive proofs with weak error bounds," in *Proceedings of the 3rd Innovations in Theoretical Computer Science*, 2012, pp. 266–275. I

[25] B. Fefferman and C. Y.-Y. Lin, "A complete characterization of unitary quantum space," in *Proceedings of the 9th Innovations in Theoretical Computer Science*, 2018, pp. 4:1–4:21. I

[26] A. Pereszlényi, "Multi-prover quantum Merlin-Arthur proof systems with small gap," 2012, technical report, arXiv:1205.2761. I

[27] A. Natarajan and T. Vidick, "Two-player entangled games are NP-hard," in *Proceedings of the 33rd Annual IEEE Conference on Computational Complexity*, 2018. I, I, I

[28] M. Coudron and H. Yuen, "Infinite randomness expansion with a constant number of devices," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2014, pp. 427–436. I

[29] A. Abboud, A. Rubinstein, and R. Williams, "Distributed PCP theorems for hardness of approximation in P," in *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, 2017. II-E