2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)

Succinct arguments for QMA from standard assumptions via compiled nonlocal games

Tony Metger
ETH Zurich
Zurich, Switzerland
tmetger@ethz.ch

Anand Natarajan
MIT
Cambridge, MA, USA
anandn@mit.edu

Tina Zhang
MIT
Cambridge, MA, USA
tinaz@mit.edu

2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS) | 979-8-3315-1674-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/FOCS61266.2024.00078

***Abstract*—We construct a succinct classical argument system for QMA, the quantum analogue of NP, from generic and standard cryptographic assumptions. Previously, building on the prior work of Mahadev (FOCS '18), Bartusek et al. (CRYPTO '22) also constructed a succinct classical argument system for QMA. However, their construction relied on post-quantumly secure indistinguishability obfuscation, a very strong primitive which is not known from standard cryptographic assumptions. In contrast, the primitives we use (namely, collapsing hash functions and a mild version of quantum homomorphic encryption) are much weaker and are implied by standard assumptions such as LWE. Our protocol is constructed using a general transformation which was designed by Kalai et al. (STOC '23) as a candidate method to compile any quantum nonlocal game into an argument system. Our main technical contribution is to analyze the soundness of this transformation when it is applied to a *succinct* self-test for Pauli measurements on maximally entangled states, the latter of which is a key component in the proof of $\mathsf{MIP}^* = \mathsf{RE}$ in quantum complexity.**

***Index Terms*—quantum complexity, interactive protocols, succinct arguments, quantum cryptography, post-quantum cryptography**

I. Introduction

Succinct verification of computation is a notion that has been extensively studied in the classical setting. A weak classical client may delegate a classical computation to a powerful server, and may then wish to check whether the server performed the computation correctly without having to compute the answer for itself. In this case, the client can ask the server to execute a *succinct interactive argument*, in which the server (efficiently) convinces the client beyond reasonable doubt that the computation was performed correctly, and the client only has to do work scaling with $\operatorname{poly}\log T$ in order to be convinced, where T is the time that it took to do the computation itself. The messages in this succinct interactive protocol should also be $\operatorname{poly}\log T$ in length.

Not long after it came to light that quantum algorithms could outperform the best known classical algorithms in certain computational tasks, the question was posed of whether a quantum prover could convince a classical verifier of the answer to a problem in BQP without requiring the classical verifier to simulate the computation itself. For certain problems, like factoring, a classical verifier can check correctness by exploiting the fact that the problem lies in NP; however, NP is not known to contain BQP, and for some problems this may be infeasible. This line of inquiry was initiated by Gottesman in 2004 [2], and has led to a long line of work on the problem now known as *quantum verification*.

Mahadev's work in 2018 [3] showed that it is indeed possible for an efficient quantum prover to convince a classical verifier of the answer to any problem in BQP, given that the quantum prover is subject to certain (post-quantum) cryptographic assumptions. (In fact, her work also showed that it is possible for an efficient quantum prover to convince a classical verifier of the answer to any problem in QMA, assuming the prover is given polynomially many copies of the witness state for the QMA problem.) Mahadev's quantum verification protocol inspired a slew of followup work in which her techniques were used to design other cryptographic quantum verification protocols with desirable additional properties, e.g. the property of being non-interactive [4] or composable [5] or linear-time [6]. In 2022, Bartusek et al. [7] showed, *assuming post-quantum* iO, that some version of Mahadev's protocol can be made *succinct*, in the same sense that we described in the opening paragraph: the classical verifier only needs to read messages that are $\operatorname{poly}\log n$ bits long, where n is the size of the instance, and do work scaling with $\operatorname{poly}\log T + \tilde{O}(n)$, where T is the time required to execute the verification circuit.

iO, or indistinguishability obfuscation, is an immensely powerful and subtle primitive that has recently been constructed from a combination of several standard assumptions [8]. However, some of these assumptions are *not* post-quantum, and there is currently no construction of post-quantum iO from standard assumptions. Post-quantum iO is known to imply other elusive cryptographic objects, e.g. public-key quantum money [9], and constructing it from standard post-quantum assumptions remains a difficult and important open problem.

The essential difficulty, and the reason for the use of iO in [7], is that Mahadev's approach to verification is in some sense a qubit-by-qubit approach, and requires $\Omega(\lambda)$ bits of communication (where λ is the security parameter) *for every qubit* in the prover's witness state, because the verifier needs to send the prover as many $\Omega(\lambda)$-sized public keys as the

This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing. TM acknowledges support from the ETH Zurich Quantum Center and an ETH Doc.Mobility Fellowship. The full version of this paper with technical details is available at [1].

2575-8454/24/$31.00 ©2024 IEEE
DOI 10.1109/FOCS61266.2024.00078

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

witness has qubits. As a result, Mahadev's approach is difficult to make succinct, since setting up the keys already requires at least $n \cdot \Omega(\lambda)$ bits of communication, where n is the number of qubits in the prover's witness. The authors of [7] use iO in a clever way to compress the keys and thus reduce the amount of required communication to $\text{poly}(\lambda) \cdot \text{poly}\log n$; this is the bulk of their work. More specifically, the authors begin by constructing a *question-succinct* (short questions, long answers) protocol for verifying QMA using iO. Then they present a general compiler which uses a recent post-quantum analysis of Killian's succinct arguments of knowledge [10], [11] to turn any question-succinct protocol that satisfies certain properties into a fully succinct protocol.

Our result

Our main contribution in this paper is to construct succinct classical-verifier arguments for QMA from standard (and even relatively general) assumptions, without relying on post-quantum iO. More specifically, we prove the following theorem:

Theorem I.1. *Assume that a quantum levelled homomorphic encryption scheme exists which specialises to a classical encryption scheme when it is used on classical plaintexts.*[1] *Assume also that post-quantum succinct arguments of classical knowledge exist.*[2] *Then a constant-round classical-verifier argument system for any promise problem in* QMA *exists, in which:*

1) *the honest quantum prover runs in quantum polynomial time, given polynomially many copies of an accepting* QMA *witness state,*
2) *the completeness-soundness gap is a constant, and*
3) *the total communication required is of length* $\text{poly}\log n \cdot \text{poly}\,\lambda$*, where* n *is the instance size and* λ *is the security parameter. The verifier runs in time* $\text{poly}(\log T, \lambda) + \tilde{O}(n)$*, where* T *is the size of the* QMA *verification circuit.*

The main advantage of our protocol compared with Bartusek et al.'s protocol [7] is that our protocol does not use post-quantum iO, which at this time cannot be instantiated from standard assumptions. Even setting aside the issue of post-quantum iO, however, we remark that the non-iO assumptions that our approach relies on are more generic than the Learning With Errors (LWE)–based assumptions which Bartusek et al. use. For example, our approach avoids using the delicate 'adaptive hardcore bit' property of LWE-based trapdoor claw-free functions (TCFs), which was introduced in [13] and used in Mahadev's original verification protocol (as well as Bartusek et al.'s protocol). The main primitive we rely on, quantum homomorphic encryption (QHE), can be constructed in its usual form from LWE without the adaptive hardcore bit assumption [14]. Moreover, we do not in fact need all the properties of standard QHE: for example, we do not use the standard notion of *compactness*, which says that decryption time cannot depend on the size of the circuit being evaluated. Instead, we only need a weak notion of compactness which says that classical ciphertexts encrypted under the QHE scheme should be classically decryptable (in any polynomial time, even if the decryption time depends on the evaluated circuit). This more general notion of *non-compact QHE with classical decryption for classical ciphertexts* plausibly exists from assumptions other than LWE: for instance, [12] represents recent progress in this direction. As such, our approach shows that the important primitive of quantum verification—and even succinct verification—may exist from a wider range of assumptions than LWE only. (In contrast, a large number of post-quantum primitives that use techniques from Mahadev's original verification protocol can only, as far as we can see, be constructed from LWE.)

We achieve Theorem I.1 by combining powerful information-theoretic tools which originate in the study of nonlocal games (e.g. those found in [15]) with tools that cryptography offers (in particular, cryptographic succinct arguments of knowledge turn out to be very useful for us). The resulting protocol is (compared with the protocol designed by Bartusek et al.) a remarkably clean object which has a natural intuitive interpretation. The tool that allows us to combine self-testing techniques with cryptographic techniques is a *compilation procedure* introduced by Kalai, Lombardi, Vaikuntanathan, and Yang [16], which Natarajan and Zhang [17] recently exploited in order to achieve classical-verifier quantum verification using a different approach from Mahadev's original approach.

A. A different approach to verification based on nonlocal games

Since Bell's historical observation [18] that there are certain *nonlocal games* which quantum entangled players can win with higher probability than classical players, the *entangled two-prover* model of computation has been a model of great interest in quantum complexity theory and quantum foundations [19]. A nonlocal game is a game played between a single efficient classical referee (or verifier) and two or more unbounded players (or provers) who cannot communicate with each other but are allowed to share entanglement. The study of the computational power of nonlocal games (i.e. what can the verifier compute efficiently with the help of the provers, if the verifier doesn't trust the provers?) has led to a fruitful line of work which, in particular, has shown that the verifier in this setting can decide any problem in RE [20]. In addition, it is known [21] that, even if the honest provers are required to be efficient, the verifier can still decide any problem in BQP (or QMA, if one of the provers gets access to polynomially many

[1] In fact, we do not need all the properties of a typical QHE scheme: for example, we do not use the standard notion of *compactness*, which says that decryption time cannot depend on the size of the circuit being evaluated. Instead, we only need a weak notion of compactness which says that classical ciphertexts encrypted under the QHE scheme should be classically decryptable (in any polynomial time, even if the decryption time depends on the evaluated circuit). We also expect that the weaker primitive of classical-client quantum blind delegation (in which interaction is allowed) would likely suffice. These weaker primitives could plausibly be instantiated from weaker assumptions than LWE, since they do not imply classical FHE, which is only known assuming LWE to date. For recent progress towards this, see [12].

[2] These can be constructed from any collapsing hash function.

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

copies of a witness). Put another way, quantum verification in the *entangled two-prover* setting is known to exist.

In 2023 Natarajan and Zhang [17] presented a reproof of Mahadev's result which took a different approach to her original approach, building on previous work on quantum blind delegation [14] and the work of Kalai, Lombardi, Vaikuntanathan, and Yang [16]. Kalai, Lombardi, Vaikuntanathan, and Yang used quantum blind delegation (in particular, quantum homomorphic encryption) in order to design a *compilation* scheme which maps any entangled two-prover proof system to a *single-prover argument*, using cryptography to enforce the no-communication assumption between the provers. Kalai et al. showed that their compilation scheme preserves quantum completeness and classical soundness, and Natarajan and Zhang showed that it also preserves quantum soundness for a certain restricted class of two-prover nonlocal games, which was sufficient to compile a two-prover quantum verification protocol into a single-prover cryptographic protocol and thus recover Mahadev's result.

From the point of view of designing succinct arguments, this approach is more attractive than Mahadev's original approach as a starting point, because the verifier only needs to send the prover a *single* public key of length $\mathrm{poly}(\lambda)$ in order to allow it to do homomorphic evaluations. One might then hope to construct a succinct cryptographic verification protocol for QMA in the following way: start with a succinct *two-prover* quantum verification protocol, pass it through the KLVY compiler, and prove soundness using similar techniques to those which Natarajan and Zhang used in [17]. This approach avoids using iO entirely, because the KLVY compiler is 'naturally' succinct when applied to a succinct protocol.

B. Succinct quantum verification in the entangled two-prover setting

It is therefore natural to ask whether *succinct* quantum verification in the entangled two-prover setting is known. The answer to this question is—unfortunately—no, but for surprisingly complicated reasons. Below is a list of the partial results in this area which are known:

1) If the honest provers are allowed to be inefficient, and if the (classical) verifier is allowed to take $\mathrm{poly}\, n$ time, then there is a protocol with $\mathrm{poly}\log n$ total communication in the entangled two-prover setting to decide QMA (in fact, to decide all of RE). This was shown by [22].
   Unfortunately, this result is not useful to us if our goal is to compile a succinct two-prover proof system into a succinct one-prover quantum verification protocol, since we want the honest prover to be efficient.
2) In a setting where the verifier interacts with *seven* provers instead of two, [23] claimed to show that efficient-prover quantum verification of QMA is possible. However, the proof of this result had two substantial errors in it. One of these errors has been resolved by [24]. The other one remains unresolved: see this erratum notice with an explanation of the error [25].
   Even assuming the errors in [23] can be fixed, a seven-prover protocol is not useful to us because the techniques from [16], [17] were only designed for nonlocal games with two provers. It seems difficult to extend these techniques to a larger number of provers, which would be necessary to compile the seven-prover protocol from [23].
3) Examining the proof of $\mathsf{MIP}^* = \mathsf{RE}$ from [20] shows that it relies on two so-called *compression theorems*: a *question reduction theorem* which takes a two-prover nonlocal game with long questions (messages from the verifier to the provers) and maps it to a nonlocal game with exponentially smaller questions while preserving most other properties of the game, and an *answer reduction theorem* which takes a two-prover nonlocal game with long answers (messages from the provers to the verifier) and maps it to a nonlocal game with exponentially smaller answers.
   One would think that these theorems would make proving succinctness in the nonlocal setting easy. Unfortunately, these compression theorems come with caveats: in particular, the answer reduction theorem can only be applied to so-called *oracularisable* protocols, and no one has come up with a two-prover verification protocol for QMA with efficient honest provers which satisfies this property. Moreover, even supposing that we had a protocol to which we could apply answer reduction, the answer reduction procedure itself happens to be so complicated and delicate that there is no clear way to analyse its soundness in the compiled setting, even given the techniques from [17] and the additional techniques for compiling nonlocal games which have been developed since then [26].
   Question reduction is both simpler and more lenient, however: while it has never been published, *question-succinct* quantum verification for QMA in the two-prover setting can be elegantly obtained from known results [15], [21].

C. The best of both worlds

The essential reason that two-prover succinct verification remains an open problem is that nonlocal *answer reduction* is hard. The only known way to make the answers in a nonlocal game shorter is to use an 'entanglement-sound' classical PCPP, and constructing this object is arguably the most technical and delicate part of the proof that $\mathsf{MIP}^* = \mathsf{RE}$. On the other hand, one can make the *questions* in certain (useful) classes of nonlocal games shorter using only the elegant machinery of de la Salle [15], who simplified the question reduction theorems of [20] by rephrasing them in terms of sampling from ϵ-biased sets. Therefore, in the nonlocal world, *question reduction* is now considered to be relatively easy, and *answer reduction* remains hard.

In Bartusek et al.'s approach to succinct verification, meanwhile, the situation was just the opposite: shortening the questions in the Mahadev protocol using only cryptography was a significant challenge, and shortening the answers could

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

be done using known techniques in a relatively black-box manner. Given that this is the case, one might hope to combine the Bartusek et al. approach with the compilation approach in order that the strengths of each might cancel out the weaknesses of the other.

This is precisely what we do in this work. We construct a succinct verification protocol for QMA by firstly compiling, using the KLVY compiler, a *question-succinct* two-prover protocol for QMA, and then compressing the answers in a generic way using Bartusek et al.'s Killian-based compiler.

The success of this approach makes a case for using the KLVY compiler as a general way to translate techniques that are well-understood in the entangled two-prover world into the single-prover cryptographic world. Once this has been done, they can be combined with 'natively' cryptographic techniques in order to marry the desirable properties of both. It seems plausible that many of the existing results in the sphere of classical-client quantum delegation and verification could have been obtained in a more unified way and from milder or more generic assumptions if the KLVY compiler had been known at the time of their genesis, because many tasks that appear difficult in the cryptographic single-prover setting are well-studied already in the nonlocal setting (and vice versa).

II. Technical overview

We focus here on how we obtain question-succinct quantum verification in the single-prover cryptographic setting, since the Killian-based answer compression protocol and its analysis were already presented in [7, Section 9].

1) The basic template from [17]: Like [17], our starting point is a basic framework for QMA verification in the two-prover setting due to Grilo [21]. The verifier and the two provers (who we will call Alice and Bob) receive as input an instance of the QMA-complete promise problem *2-local XZ Hamiltonian* [27]. In other words, the problem that the verifier is trying to decide is whether a certain Hamiltonian H on n qubits, expressed as a sum of polynomially many 2-local X/Z Pauli terms (where each term is a tensor product of n operators, each of which is chosen from $\{\mathrm{id}, \sigma_X, \sigma_Z\}$, such that all but 2 factors in the tensor product are id), has lowest eigenvalue $\leq \alpha$ or $\geq \beta$ for two real numbers (α, β), where we are promised that $\beta - \alpha \geq \frac{1}{\mathrm{poly}(n)}$.

Honest Alice and Bob start out by sharing n EPR pairs. The two-prover protocol underlying [17] for deciding whether H has lowest eigenvalue $\leq \alpha$ or $\geq \beta$ consists of two subtests, the *Pauli braiding test* and the *energy test*:

Protocol 1 (informal)**.**

1) **Pauli braiding.** Alice and Bob execute a version of the Pauli braiding protocol from [28], in which they play interleaved copies of CHSH (or another similar game, like Magic Square) and a simple game known as the 'commutation test'. This protocol is a *robust self-test* for the n-qubit Pauli group[3], in the sense that entangled players who win with high probability in this game must both be playing with measurement operators that are close (up to local isometries) to actual Pauli measurements. In other words, the Pauli braiding test allows the verifier to 'force' entangled provers to perform Pauli measurements when requested to do so, even without trusting the provers. The most modular analysis of this protocol proceeds via a theorem from approximate representation theory that was first proven by Gowers and Hatami [29].
2) **Energy testing via teleportation.** Alice is asked to *teleport* the n-qubit witness state to Bob using their n shared EPR pairs. She then reports the teleportation corrections to the verifier. Bob is asked to measure certain Pauli operators and report the outcomes. The verifier corrects Bob's reported outcomes using Alice's reported teleportation corrections, and interprets the result as a measurement of a term from H. It accepts or rejects depending on whether this measurement indicates that the state which Alice was meant to teleport to Bob is low-energy or high-energy.

The intuition for the soundness of Protocol 1 is as follows: the Pauli braiding test guarantees in some sense, through the use of the Gowers-Hatami theorem [29], that all successful Bobs are in fact equivalent to honest Bob; and the energy test is straightforward to analyse if Bob is honest. In order to translate the intuition into reality, we have to make sure that Bob uses the *same* strategy in both subtests so that the guarantee on Bob in the Pauli braiding test also applies to Bob in the energy test. That is, we must make sure he cannot play honestly only in the Pauli braiding test and then deviate however he likes in the energy test.

Suppose for the moment that the two subtests can be made *perfectly indistinguishable* to Bob: that is, suppose that Bob's questions in both subtests are drawn from the same distribution. This would ensure that he does the same measurements in both subtests, since he does not know which subtest is being performed. The Pauli braiding subtest then guarantees that these measurements are 'close' to honest measurements, and the soundness of subtest (ii) follows directly from the soundness of subtest (ii) with an honest Bob.

In [17], following a template laid out by Vidick in [30], the two subtests were indistinguishable because Bob's questions are very simple: in both subtests, Bob only ever receives one of two questions, each with $\frac{1}{2}$ probability. One of these two questions is an instruction to measure all of his qubits in the Z basis (and report all n outcomes), and the other is an instruction to measure all his qubits in the X basis. Slightly more formally, honest Bob will in one case apply the projective measurement $\{|z\rangle\langle z| : z \in \{0,1\}^n\}$, and in the other case he will apply the projective measurement $\{H^{\otimes n}|x\rangle\langle x|H^{\otimes n} : x \in \{0,1\}^n\}$.

[3] More precisely the Heisenberg-Weyl group, the group consisting of tensor products of $\mathrm{id}, \sigma_X, \sigma_Z$ with ± 1 signs, but we ignore this distinction in this introduction.

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

Measurements of this form, as it turns out, are particularly 'compatible' with the Gowers-Hatami-based analysis of the Pauli braiding test, in a sense that we will make somewhat more precise later (when we explain our 'mixed-vs-pure basis test' later in this overview). It would therefore be convenient if this question structure was also sufficient for the energy test. Fortunately, this happens to be the case in the non-succinct setting: it turns out that 2-local X/Z Hamiltonian with inverse polynomial gap is complete for QMA even if we restrict the 2-local terms to XX and ZZ terms, i.e., terms where the two non-identity components of the n-fold tensor product are always of the same type (σ_X or σ_Z). Note that the verifier can reconstruct a measurement of any XX-type term from the outcomes of an all-X measurement performed by Bob, and any ZZ-type term from the outcomes of an all-Z measurement performed by Bob. This means that in [17], it was sufficient in both subtests to ask Bob the same two questions (all-X and all-Z), each with $\frac{1}{2}$ probability. Perfect indistinguishability of the two subtests in Protocol 1 then follows.

2) Obtaining succinctness: In designing a (question-)succinct protocol with two entangled provers (which we will later compile into a crypographically secure single-prover protocol), we are faced with two new challenges compared with [17]:

1) The Pauli braiding test (subtest (i) of Protocol 1) does not have succinct questions. In particular, while *Bob's* questions can easily be made succinct (as we just described, it suffices to have only two Bob questions), *Alice's* questions are more complicated.
2) In the non-succinct setting, $\frac{1}{\mathrm{poly}(n)}$ completeness-soundness gap is generally tolerated because it is assumed that $\mathrm{poly}(n)$ many rounds of sequential repetition can be performed in order to boost the gap. In the succinct setting, this is not feasible, since repeating a succinct protocol $\mathrm{poly}(n)$ times results in $\mathrm{poly}\log(n) \cdot \mathrm{poly}(n)$ communication; therefore, in the succinct setting, we must design a protocol which has *constant* soundness gap even without any repetition. This means that we cannot start with a 2-local XX/ZZ Hamiltonian, since it is not known whether this problem is QMA-complete with a constant promise gap. If we are to take the same approach of starting from some Hamiltonian problem, then it has to be a Hamiltonian problem with *constant* promise gap such that the terms can be grouped into a small number of subsets (at most $2^{\mathrm{poly}\log n}$ subsets), each of which contains only terms that commute. If this is the case, then Bob can measure all the terms in a single subset simultaneously and report all the outcomes together, and the verifier only needs to use $\mathrm{poly}\log n$ bits to tell Bob which subset to measure. If this is not the case, then the energy testing template from subtest (ii) of Protocol 1 will not work, since the verifier will not be able to tell Bob which terms he should measure in a succinct way.

a) Subsampling Hamiltonians: We take a similar approach to Bartusek et al. [7] in order to deal with the second issue. We use naïve QMA parallel amplification (first written down in [31]; the procedure simply repeats the QMA verifier in parallel a polynomial number of times) in order to boost the promise gap to a constant; this results in a Hamiltonian that is a sum of exponentially many terms, each of which can be efficiently measured by measuring each of the n qubits of the witness in either the X or the Z Pauli basis (with potentially different basis choices for different qubits). We then use a generic *PRG with soundness against adversaries with quantum advice* in order to 'subsample' these terms and emerge with a Hamiltonian that is a sum of $2^{\mathrm{poly}\log n}$ terms, each of which can be efficiently measured by measuring each of the n qubits of the witness in either the X or the Z Pauli basis.

b) The mixed-vs-pure basis test and a new self-testing-oriented proof of Gowers-Hatami: At this point we have created a new problem: the terms of the Hamiltonian we want to use in the energy subtest can no longer be measured by a Bob who only ever measures every qubit of his state in either the X basis or the Z basis. This is because the amplified Hamiltonian contains tensor products of arbitrary combinations of XX and ZZ terms from the original Hamiltonian, and not only tensor products of terms in the same basis. These mixed terms *can* be measured by a Bob who does what we call *mixed basis* measurements (measurements that involve measuring each of n qubits in either the X or the Z basis, with potentially different basis choices for different qubits). However, if the verifier picks the mixed bases depending on the distribution induced by the constant-gap Hamiltonian, the resulting distribution over Bob questions is not necessarily 'compatible' with even the regular Pauli braiding test. Moreover, it becomes even more difficult to use anything other than the all-X and all-Z measurements when we consider the *succinct* version of Pauli braiding, for reasons that we will elaborate on shortly (in the section 'Succinct Pauli braiding').

The natural solution is to use the all-X and all-Z measurements when we play Pauli braiding, use the mixed basis measurements when we do the energy test, and introduce some sort of *consistency test* to ensure that the operators that Bob uses in the energy test are in some sense the same ones as the ones he uses in Pauli braiding. We call this test the mixed-vs-pure basis test. Such tests have been analysed in the nonlocal setting before [32], but we are the first to attempt to analyse such a test in the compiled setting, and the compilation introduces unforeseen difficulties (see 'Difficulties in the analysis of the mixed-vs-pure basis test' below).

The easiest solution to the difficulties that we were able to come up with involves reproving the Gowers-Hatami theorem (or, rather, the parts of the theorem relevant for self-testing) in a way that supports arbitrary non-uniform expectations. The (informal) theorem statement for our version of Gowers-Hatami is as follows:

Theorem II.1 (informal)**.** *Let $f : G \to U(\mathcal{H})$ be a function from a finite group G to the set of unitaries on some Hilbert space $\mathcal{H}$. Then there exists a finite-dimensional Hilbert space*

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

$\mathcal{H}'$, an isometry $V : \mathcal{H} \to \mathcal{H}'$, and a unitary representation $\pi : G \to U(\mathcal{H}')$ of G such that for all measures μ over G,

$$\mathop{\mathbb{E}}_{g\sim\mu,h\sim W_n} \|f(h)f(g) - f(hg)\|^2 \leq \epsilon \\ \implies \mathop{\mathbb{E}}_{g\sim\mu} \|f(g) - V^\dagger\pi(g)V\|^2 \leq \epsilon$$

where we are being purposefully vague about the norm.

The difference between this theorem and the more typical formulation is that the typical formulation has uniform expectations over the group everywhere. A version of Gowers-Hatami similar to Theorem II.1 is often needed in the self-testing setting when μ is in particular the uniform distribution over $\{\sigma_Z(a) : a \in \{0,1\}^n\}$ or $\{\sigma_X(b) : b \in \{0,1\}^n\}$, and it is plausible that Theorem II.1 could also be proven by modifying in some way Gowers and Hatami's original proof of their theorem. Nonetheless, the proof that we present for Theorem II.1 is an entirely different proof that only uses basic tools from quantum information, namely Stinespring dilation (instead of matrix Fourier analysis on non-Abelian groups [29]). We emphasise that our proof is *not* a reproof of the full Gowers-Hatami theorem, because the original theorem gets bounds on the dimension of the 'post-rounding' Hilbert space $\mathcal{H}'$ (which one typically does not need in self-testing-related applications of Gowers-Hatami). However, our proof has the advantage that it is completely elementary and self-contained. We believe this proof may be of independent interest, because the fact that it is simple and self-contained makes it easier to modify the statement when necessary to incorporate additional desirable properties (such as, for example, the tolerance for non-uniform expectations that we needed for this work). Together with a 'distribution-switching' trick (see [1] for details), we are able to use this version of Gowers-Hatami to work out an analysis of the mixed-vs-pure basis test. We give more details about how we did this at the end of the following section.

c) Difficulties in the analysis of the mixed-vs-pure basis test: Now we elaborate more thoroughly on the nature of the difficulties that we encountered in analysing the mixed-vs-pure basis test. We firstly justify the sense in which the all-X and all-Z measurements are particularly 'compatible' with Pauli braiding, in order to clarify why the consistency test is necessary in the first place.

Why the mixed-vs-pure basis test is necessary. The reason why the all-X and all-Z measurements are particularly suitable for use in the Pauli braiding protocol is that the all-Z question can be interpreted as a simultaneous measurement of the 2^n binary observables $\{\sigma_Z(a) : a \in \{0,1\}^n\}$, where $\sigma_Z(a)$ is the binary observable that is the tensor product of σ_Z on all the qubits i where $a_i = 1$ and identity otherwise; and, similarly, the all-X question can be interpreted as a simultaneous measurement of the 2^n binary observables $\{\sigma_X(b) : b \in \{0,1\}^n\}$. Another (more precise) way to say this is that, given the (potentially cheating) projective measurement $\{P^Z_u : u \in \{0,1\}^n\}$ that Bob applies when he receives the instruction to measure everything in the Z basis, we can construct a set of 2^n binary observables $\{Z(a) : a \in \{0,1\}^n\}$ which are *exactly linear*, in the sense that

$$Z(a)Z(a') = Z(a + a'), \tag{II.1}$$

even if Bob is dishonest: simply take

$$Z(a) \coloneqq \sum_{u\in\{0,1\}^n} (-1)^{u\cdot a} P^Z_u.$$

A similar statement holds true for the all-X measurement: we can define a set of 2^n binary observables $\{X(b) : b \in \{0,1\}^n\}$ such that

$$X(b)X(b') = X(b + b'). \tag{II.2}$$

We can use the CHSH game and the commutation test in order to certify that these $2 \cdot 2^n$ binary observables $\{Z(a), X(b) : a, b \in \{0,1\}^n\}$ satisfy the commutation relations that would hold if they were genuine Paulis, i.e.

$$\|Z(a)X(b) - (-1)^{a\cdot b}X(b)Z(a)\|_2 \leq O(\epsilon). \tag{II.3}$$

Taking the linearity (Equation (II.1) and Equation (II.2)) and commutation (Equation (II.3)) relations together, we can prove that $\{Z(a) : a \in \{0,1\}^n\}$ and $\{X(b) : b \in \{0,1\}^n\}$ approximately satisfy the relations satisfied by the corresponding elements of the Pauli group. Moreover, by taking products, we can extend $Z(a)$ and $X(b)$ to a matrix-valued function $f(s, a, b) = (-1)^s Z(a)X(b)$ that approximately obeys the multiplication law of the Pauli group. The Gowers-Hatami theorem then implies that there is a *rounding* of f which *exactly* satisfies the Pauli group relations (up to isometry). That is, there exists a representation ρ of the Pauli group such that, on average over s, a, b, $f(s, a, b)$ is close to $\rho(s, a, b)$ conjugated by the isometry.

Zooming back out to the level of designing Bob's questions, note that the all-Z and all-X questions were particularly nice for the Pauli braiding test because (1) the sets $\{\sigma_Z(a) : a \in \{0,1\}^n\}$ and $\{\sigma_X(b) : b \in \{0,1\}^n\}$ taken together generate the entire n-qubit Pauli group, and (2) the trick of constructing many binary observables from a single projective measurement gave us *exact linearity* on the Z side and the X side individually almost for free: that is, $\{Z(a) : a \in \{0,1\}^n\}$ is automatically an exact representation of $\mathbb{Z}_2^n$, and the same is true of $\{X(b) : b \in \{0,1\}^n\}$.

There is no guarantee that these nice properties hold if we consider (instead of the all-X and all-Z questions) the set of mixed-basis questions induced by the energy test for our constant-gap Hamiltonian. In particular, there is no guarantee that the binary observables which can be constructed from Bob's set of mixed basis measurements will generate the whole Pauli group, in the way that $\{\sigma_Z(a) : a \in \{0,1\}^n\}$ and $\{\sigma_X(b) : b \in \{0,1\}^n\}$ generate the whole Pauli group. It becomes even more important to use the all-X and all-Z questions if we want to eventually make the Pauli braiding test question-succinct: we give some intuition as to why this is the case in the section 'Succinct Pauli braiding'.

The easiest solution seems to be to introduce a consistency test between Bob's mixed basis measurements (that we would

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

like Bob to use when he plays the energy test) and Bob's pure basis measurements (that we would like Bob to use when he plays the Pauli braiding test). More specifically—following the standard template for designing tests of this form—we will introduce two new questions into Alice's question set that are identical to Bob's pure basis questions (i.e. 'measure all in X' and 'measure all in Z'); we will ask Bob to play his pure basis operators against Alice's pure basis operators, in order to check that Bob's pure basis operators are consistent with Alice's pure basis operators; and then we will ask Bob to play his mixed basis operators against Alice's pure basis operators, and check that they agree whenever the bases align, which (since we checked that Bob's and Alice's pure basis operators agree) is essentially equivalent to checking that Bob's mixed basis operators are consistent with Bob's pure basis operators. We might hope that this test, combined with the usual analysis, will be sufficient to allow us to 'round' Bob's mixed operators in the same way that we can round Bob's all-X and all-Z measurements by using the usual Gowers-Hatami analysis.

Difficulties in the analysis. Unfortunately, instantiating this intuition proves to be nontrivial in the compiled setting, even though the analysis is fairly routine in the nonlocal setting. The tensor product structure of the provers' Hilbert space in the nonlocal setting is useful because it supports a large range of convenient operations that are loosely grouped together under the name of 'prover-switching'. The ordinary nonlocal analysis of a consistency test like this one would proceed primarily through prover-switching calculations. While we did find it necessary to prove some lemmas which capture certain applications of prover-switching in the compiled setting, we found that these lemmas were insufficient in order to analyse the mixed-vs-pure basis test.

More specifically, the main difficulty we encountered was the following. The statement we would like to show, in order to make the energy test work in the presence of mixed terms, is of the following form. Let $w \in \{\mathrm{id}, X, Z\}^n$ be a string indicating which Pauli bases to measure n qubits in. We want to show that, if Alice and Bob win in our protocol with high probability, then there exists an isometry V such that, for the distribution D on Pauli basis choices induced by the constant-gap Hamiltonian,

$$\mathop{\mathbb{E}}_{w\sim D}\ \mathop{\mathbb{E}}_{a\in\{0,1\}^n} \|O^w(a) - V^\dagger(\sigma_w(a)\otimes \mathrm{id}_{\mathrm{aux}})V\|^2 \leq \text{small},$$

where $\sigma_w(a)$ is the honest Pauli observable that corresponds to the tensor product

$$\sigma_w(a) = \bigotimes_i \sigma_{w_i}^{a_i}\,.$$

and $O^w(a)$ is Bob's potentially cheating version of $\sigma_w(a)$.

Normal pure-basis Gowers-Hatami tells us that, if Alice and Bob win with high probability in Pauli braiding, then for any $W \in \{X, Z\}$ it is the case that

$$\mathop{\mathbb{E}}_{a\in\{0,1\}^n} \|W(a) - V^\dagger(\sigma_W(a)\otimes \mathrm{id}_{\mathrm{aux}})V\|^2 \leq \text{small}, \qquad \text{(II.4)}$$

for some fixed isometry V. One idea for proceeding with the analysis might be to show that $O^w(a) \approx Z(c)X(d)$ using the mixed-vs-pure basis test (with c being the string such that $c_i = 1$ iff $a_i = 1$ and $w_i = W$, and similarly for d), and then to 'round' $Z(c)$ and $X(d)$ separately using Equation (II.4). Unfortunately, rounding something of the form $Z(c)X(d)$ naïvely using Equation (II.4) produces something of the form

$$V^\dagger(\sigma_Z(c)\otimes \mathrm{id}_{\mathrm{aux}})VV^\dagger(\sigma_X(d)\otimes \mathrm{id}_{\mathrm{aux}})V.$$

Since V is an isometry and not a unitary, $VV^\dagger$ is not necessarily id, and it is unclear how to get rid of it: we call this the '$VV^\dagger$ problem'. There are ways to bypass this problem in the nonlocal setting using tensor product structure, but we were not able to replicate these techniques in the compiled setting.

Instead, we bypass the problem by 'directly' proving a form of Gowers-Hatami that, perhaps surprisingly, allows us to round in expectation over *any* distribution over the Pauli group, even though the Pauli braiding test is only played with the uniform distribution. More specifically, we prove our version of Gowers-Hatami (Theorem II.1, see [1] for the formal version), which can be used to round *arbitrary* distributions over the underlying group, provided with the right hypothesis; and then we prove, using a 'distribution-switching' trick (see [1]), that the hypothesis of Theorem II.1 can be obtained for *any* distribution μ even if we only start with commutation relations that hold on uniform average over pure-basis elements (and a few other conditions, such as exact linearity), which is what we have access to through the pure-basis Pauli braiding test.

d) Succinct Pauli braiding: Finally, armed with the mixed-vs-pure basis test, we can focus on making the Pauli braiding test succinct (where, by 'Pauli braiding test', we mean the version in which Bob always gets asked either the all-X or the all-Z question). Our starting point for this mission is de la Salle's elegant simplification [15] of 'question reduction' from [20], in which he introduces a version of Pauli braiding where Alice's questions are sampled from ϵ-biased sets. The normal Pauli braiding game proceeds as follows:

- The verifier chooses two strings $a, b \in \{0,1\}^n$ uniformly at random.
- The verifier decides what to do next based on the parity of $a \cdot b$:
  - If $a \cdot b = 0$, the verifier referees a *commutation game* (in which honest Alice plays with $\sigma_Z(a)$ and $\sigma_X(b)$).
  - If $a \cdot b = 1$, the verifier referees an *anticommutation game* (in which, again, honest Alice plays by embedding $\sigma_Z(a)$ and $\sigma_X(b)$ into her strategy).

The commutation game is designed to test that two particular operators commute, and the anticommutation game (based on CHSH or Magic Square) is designed to test that two particular operators anticommute.

Note that the verifier has to send a, b to Alice for this protocol to work. The protocol was made succinct by de la Salle simply by choosing a, b from ϵ-biased sets instead of

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

from all of $\{0,1\}^n$. This is a natural idea, but it is at first surprising that it works at all: after all, the sets of Paulis $\{\sigma_Z(a) : a \in S\}$ and $\{\sigma_X(b) : b \in S\}$ for some ϵ-biased S, where $|S| = \mathrm{poly}(n)$, only cover an exponentially small fraction of the Pauli group! All that the protocol directly certifies is commutation and anticommutation relations among pairs of operators in these sets. Naïvely, to deduce relations about representations of *general* group elements, one would need to write these elements as $\mathrm{poly}(n)$-length words in the group elements from the ϵ-biased sets, and apply the relations on the ϵ-biased sets $\mathrm{poly}(n)$-many times. This would seemingly rule out a test with constant soundness.

Miraculously, however, everything still works as before, and the reason is that we do probe the entire group through Bob, who still measures the all-X and all-Z mesaurements. In particular, we have 'for free' (or by construction) that Bob's $X(b)$ operators, taken as a set, form an *exact* representation of $\mathbb{Z}_2^n$, and the same for his $Z(a)$ operators. Meanwhile, all elements of the Pauli group can be written as words of constant length in the operators $\{\sigma_Z(a) : a \in \{0,1\}^n\}$ and $\{\sigma_X(b) : b \in \{0,1\}^n\}$. In some sense, de la Salle's test works because probing the commutation relations between two exact representations of $\mathbb{Z}_2^n$ on only an ϵ-biased set is sufficient to establish the commutation relations everywhere, because the function of ϵ-biased sets is precisely to 'fool' exactly linear functions. In fact, de la Salle's test and its analysis are analogous to the "derandomized BLR test" for linear functions and the Fourier-based analysis of it presented in Section 6.4 of [33].

In order to use the succinct version of Pauli braiding in our protocol, we have to come up with a version of the analysis that works in the compiled setting. Unfortunately, de la Salle's original proof in the nonlocal case is written in the 'synchronous' setting, in which the provers (even malicious provers) are assumed to start out by sharing EPR pairs. This assumption simplifies the calculations because it allows us to move ('prover switch') measurements freely from one prover to the other. The synchronicity assumption is without loss of generality in the nonlocal setting by [34], but no compiled version of this result exists. Therefore, we have to redo the proof in our setting using the state-dependent distance, and come up with ways to use the cryptography to simulate the "prover switching" steps in de la Salle's analysis. (At the time of [17] it was not known whether the cryptography could in fact simulate these properties.) In the process, we pare down de la Salle's proof to the parts that are essential for analysing succinct Pauli braiding and state it in more computer-science-like language, which may be useful for future readers with a computer science background.

e) Related work: Simultaneously, a succinct argument system for QMA based only on the post-quantum security of LWE (a standard assumption) was achieved by [35]. Both of these works use tools from [7], in particular the technique of "subsampling" a Hamiltonian using a PRG, and the technique of transforming a semi-succinct protocol into a fully succinct one by using succinct arguments of knowledge. However, the methods they use to solve the key technical challenge of succinctly delegating many-qubit Pauli measurements are essentially disjoint. In particular, for us, the "heavy lifting" to achieve question-succinctness is performed *information theoretically*, in our question-succinct two-prover self-test for EPR pairs, whereas for them, succinctness is achieved by using specific technical features of a cryptographic construction using LWE.

Acknowledgements

We are grateful to Fermi Ma for allowing us to rewind him until we could extract an understanding from the succinct arguments in Section 9 of [7]. We also thank Yael Kalai, Alex Lombardi, and Thomas Vidick for useful discussions.

References

[1] T. Metger, A. Natarajan, and T. Zhang, "Succinct arguments for qma from standard assumptions via compiled nonlocal games," *arXiv preprint arXiv:2404.19754*, 2024.

[2] S. Aaronson, "The Aaronson $25.00 prize," 2007, https://scottaaronson.blog/?p=284.

[3] U. Mahadev, "Classical verification of quantum computations," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS).* IEEE, 2018, pp. 259–267.

[4] Z. Brakerski, V. Koppula, U. Vazirani, and T. Vidick, "Simpler proofs of quantumness," 2020.

[5] A. Gheorghiu and T. Vidick, "Computationally-secure and composable remote state preparation," in *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS).* IEEE, 2019, pp. 1024–1033.

[6] J. Zhang, "Classical verification of quantum computations in linear time," in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS).* IEEE, 2022, pp. 46–57.

[7] J. Bartusek, Y. T. Kalai, A. Lombardi, F. Ma, G. Malavolta, V. Vaikuntanathan, T. Vidick, and L. Yang, "Succinct classical verification of quantum computation," in *Advances in Cryptology–CRYPTO 2022: 42nd Annual International Cryptology Conference, CRYPTO 2022, Santa Barbara, CA, USA, August 15–18, 2022, Proceedings, Part II.* Springer, 2022, pp. 195–211, https://eprint.iacr.org/2022/857.

[8] A. Jain, H. Lin, and A. Sahai, "Indistinguishability obfuscation from well-founded assumptions," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 60–73.

[9] M. Zhandry, "Quantum lightning never strikes the same state twice. or: quantum money from cryptographic assumptions," *Journal of Cryptology*, vol. 34, pp. 1–56, 2021.

[10] A. Chiesa, F. Ma, N. Spooner, and M. Zhandry, "Post-quantum succinct arguments: breaking the quantum rewinding barrier," in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS).* IEEE, 2022, pp. 49–58.

[11] A. Lombardi, F. Ma, and N. Spooner, "Post-quantum zero knowledge, revisited or: How to do quantum rewinding undetectably," in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS).* IEEE, 2022, pp. 851–859.

[12] A. Gupte and V. Vaikuntanathan, "How to construct QFHE, generically," 2024, to appear.

[13] Z. Brakerski, P. Christiano, U. Mahadev, U. Vazirani, and T. Vidick, "A cryptographic test of quantumness and certifiable randomness from a single quantum device," *Journal of the ACM (JACM)*, vol. 68, no. 5, pp. 1–47, 2021.

[14] U. Mahadev, "Classical homomorphic encryption for quantum circuits," *SIAM Journal on Computing*, no. 0, pp. FOCS18–189, 2017.

[15] M. de la Salle, "Spectral gap and stability for groups and non-local games," 2022.

[16] Y. Kalai, A. Lombardi, V. Vaikuntanathan, and L. Yang, "Quantum advantage from any non-local game," 2021.

[17] A. Natarajan and T. Zhang, "Bounding the quantum value of compiled nonlocal games: From CHSH to BQP verification," in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, 2023, pp. 1342–1348.

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.

[18] J. S. Bell, “On the Einstein Podolsky Rosen paradox,” *Physics Physique Fizika*, vol. 1, no. 3, p. 195, 1964.

[19] V. Scarani, “The device-independent outlook on quantum physics (lecture notes on the power of Bell’s theorem),” *Acta Physica Slovaca*, vol. 62, no. 4, pp. 347–409, 2013.

[20] Z. Ji, A. Natarajan, T. Vidick, J. Wright, and H. Yuen, “MIP* = RE,” 2020.

[21] A. B. Grilo, “A simple protocol for verifiable delegation of quantum computation in one round,” 2017.

[22] A. Natarajan and T. Zhang, “Quantum free games,” in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023, pp. 1603–1616.

[23] A. Natarajan and T. Vidick, “Low-degree testing for quantum states, and a quantum entangled games PCP for QMA,” in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2018, pp. 731–742.

[24] Z. Ji, A. Natarajan, T. Vidick, J. Wright, and H. Yuen, “Quantum soundness of testing tensor codes,” in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2022, pp. 586–597.

[25] A. Natarajan and C. Nirkhe, “The status of the quantum pcp conjecture (games version),” 2024.

[26] D. Cui, G. Malavolta, A. Mehta, A. Natarajan, C. Paddock, S. Schmidt, M. Walter, and T. Zhang, “A computational Tsirelson’s theorem for the value of compiled XOR games,” 2024.

[27] J. D. Biamonte and P. J. Love, “Realizable hamiltonians for universal adiabatic quantum computers,” *Physical Review A*, vol. 78, no. 1, p. 012352, 2008.

[28] A. Natarajan and T. Vidick, “A quantum linearity test for robustly verifying entanglement,” in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 2017, pp. 1003–1015.

[29] W. T. Gowers and O. Hatami, “Inverse and stability theorems for approximate representations of finite groups,” *Sbornik: Mathematics*, vol. 208, no. 12, p. 1784, 2015.

[30] T. Vidick, “Interactions with quantum devices (course),” 2020, http://users.cms.caltech.edu/~vidick/teaching/fsmp/fsmp.pdf.

[31] A. Y. Kitaev, A. Shen, and M. N. Vyalyi, *Classical and quantum computation*. American Mathematical Soc., 2002, no. 47.

[32] A. Natarajan and J. Wright, “$\mathsf{NEEXP} \subseteq \mathsf{MIP}^*$,” 2019.

[33] R. O’Donnell, *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[34] T. Vidick, “Almost synchronous quantum correlations,” *Journal of mathematical physics*, vol. 63, no. 2, 2022.

[35] S. Gunn, Y. Kalai, A. Natarajan, and A. Villányi, “Classical commitments to quantum states,” 2024, to appear.

Authorized licensed use limited to: MIT. Downloaded on April 26,2025 at 19:56:30 UTC from IEEE Xplore. Restrictions apply.