Classical proofs of quantum knowledge

Thomas Vidick[*] Tina Zhang[†]

Abstract

We define the notion of a *proof of knowledge* in the setting where the verifier is classical, but the prover is quantum, and where the witness that the prover holds is in general a quantum state. We establish simple properties of our definition, including that *nondestructive* classical proofs of quantum knowledge are impossible for nontrivial states, and that, under certain conditions on the parameters in our definition, a proof of knowledge protocol for a hard-to-clone state can be used as a (destructive) quantum money verification protocol. In addition, we provide two examples of protocols (both inspired by private-key classical verification protocols for quantum money schemes) which we can show to be proofs of quantum knowledge under our definition. In so doing, we introduce new techniques for the analysis of such protocols which build on results from the literature on nonlocal games. Finally, we show that, under our definition, the verification protocol introduced by Mahadev (FOCS 2018) is a classical *argument* of quantum knowledge for QMA relations.

1 Introduction

The notion of a *proof of knowledge* was first introduced in the classical setting [GMR89, BG92] as a useful strengthening of the idea of an *interactive proof*. Intuitively, a proof of knowledge protocol allows a prover to convince a verifier that it 'knows' or 'possesses' some piece of secret information (a 'witness', w) which satisfies a certain relation R relative to a publicly known problem instance x. (Symbolically, we might say that the prover wants to convince its verifier that, for a particular x, it knows w such that $R(x, w) = 1$.) For example, the witness w might be a private password corresponding to a particular public username x, and a proof of knowledge protocol in this setting could allow the prover to demonstrate that it possesses the credentials to access sensitive information or make monetary transactions.

While the idea of a proof of knowledge is very natural, defining it formally is somewhat delicate, because the question arises of what exactly it means for a program or a machine to 'know' a piece of information. An interactive proof allows a prover to convince a verifier that some witness w *exists* such that $R(x, w) = 1$; however, this is insufficient for the password application, because (for a valid username x) a password w always exists such that the authenticating party accepts. We would like, instead, for the verifier to be convinced that the prover *personally* knows the relevant information w. But what does it mean for a machine to 'know' w? If w is sufficiently short, and the machine is sufficiently complicated, we will with high probability find that w is written in the machine's programming in some form, even if only by chance. This definition of 'knowing' is clearly also unsatisfactory for the password application.

The now-canonical formal definition of a classical proof of knowledge was settled in a series of works (see [BG92] for a summary) in the 1990s. The resolution is as follows: the prover P is said to 'know' a witness w if there is an extractor E which, given black-box access to P (including the ability to rewind P and run it again on different inputs), can efficiently compute w. The applications of classical proofs of knowledge include identification protocols [FFS88], signature schemes [CL06], and encryption schemes secure against chosen-ciphertext attack [SJ00].

In this work, we consider a particular generalisation of the classical concept of a proof of knowledge to the quantum setting. We imagine a situation where the *verifier* remains classical, but the *prover* is quantum,

[*] Department of Computing and Mathematical Sciences, California Institute of Technology, USA. vidick@caltech.edu
[†] Division of Physics, Mathematics and Astronomy, California Institute of Technology, USA. tinazhang@caltech.edu

and where the witness w is in general a quantum state; and we ask the prover to 'convince' the verifier that it knows that state. We call this type of protocol a *classical proof of quantum knowledge.* Recently, there have been works which show how a fully classical verifier can, under cryptographic assumptions, delegate a quantum computation on encrypted data to a quantum server [Mah18a], verify that such a server performed the computation correctly [Mah18b], delegate the preparation of single-qubit states to the server in a composable fashion [GV19], and test classically that the server prepared an EPR pair in its own registers [MV20]. In short, as long as classical computational resources and classical communication channels remain considerably less expensive than their quantum counterparts, it will be natural to wish to use classical devices to test quantum functionality. Although we focus here on information-theoretic rather than computational security, the current paper can be considered a new stone in the preceding line of work.

Quantum proofs of quantum knowledge have recently been explored by [BG19] and [CVZ19]; these two papers give a joint definition for quantum proofs of quantum knowledge, and exhibit several examples which meaningfully instantiate the definition. However, if we are interested in testing quantum functionality with *classical* devices, as we are here, then we must approach the subject differently. The reason is that, if we allow the extractor only black-box access to the prover (as is done in [BG19] and [CVZ19], as well as in the classical literature) in the setting where the verifier is classical but the prover is quantum, the problem the extractor faces becomes one of reconstructing a witness ρ based entirely on classical measurement outcomes, which seems as if it may be as hard as quantum state tomography. To give an idea of the difficulty of the problem, information-theoretic bounds have been proven which show that reconstructing a full classical description of a quantum state ρ from measurement outcomes requires (in general) measuring exponentially many copies of ρ [HHJ$^+$17], even when the the extractor is free to choose the measurements which are performed. To allow this seems excessive, and yet it may be hard to prove anything resembling knowledge extraction if we demand that the prover and extractor be efficient in this black-box setting. It would be more reasonable to relax the black-box requirement in some way so that we can reasonably expect efficient extractors to be found.

Our first contribution is to provide a workable definition of a proof of quantum knowledge for the setting where the only communication between verifier and prover is classical. In order to circumvent the difficulty described in the preceding paragraph, we do not require that the extractor uses the prover as a black box, but permit it to make use of the prover's internal state. Specifically, we define a new abstract party that we call the *intermediary*, and for any protocol $\mathcal{P}$ between a classical verifier and a quantum prover, we define a *mediated* version of that protocol, $\mathcal{P}'$, in which the prover does not directly compute its own messages, but is instead required to provide to the (trusted) intermediary any quantum state σ that it might wish to use in the protocol $\mathcal{P}$, in addition to a black box C that implements unitaries which represent the actions the prover would have performed in the protocol $\mathcal{P}$ in response to the verifier's challenges.[1] The intermediary then interacts with the verifier according to the prover's instructions, and at the end the verifier either accepts or rejects. The purpose of the intermediary is to make explicit the resources which the extractor has access to; we stress that it is only a formalism, and does not restrict in any way the range of malicious actions which the prover may take, or indeed alter its behaviour at all from its behaviour in the real protocol.[2] We then say the protocol $\mathcal{P}$ is a *proof of knowledge* for a quantum state ρ if there exists an extractor which can, for any prover P which passes with high probability in the mediated protocol $\mathcal{P}'$, extract from the σ and C that P gave its intermediary a quantum state that is close in trace distance to ρ.

There are two elementary but potentially interesting properties of this definition that can be simply proved, and we provide proofs of these properties in section 4. The first property is that, loosely speaking, *nondestructive* classical proofs of quantum knowledge are impossible for nontrivial states: that is, if a classical proof of quantum knowledge leaves the witness state intact, then the witness state can be cloned.

[1] Forcing the black box C to be unitary ensures, for example, that the extractor need not worry about the prover making a destructive measurement of its internal state for no reason other than to prevent the extractor from looking at it, because the extractor is able to decide whether to implement that measurement or not. See Section 3 for a more precise description of the nature of C, and see Section 7 for an example of how its unitarity is used.

[2] Gheorghiu and Vidick consider a similar model to this intermediary-based model in [GV19]; their model served as part of the inspiration for ours.

Since arbitrary quantum states cannot be cloned, we conclude that only a restricted set of relatively simple states can have nondestructive classical proofs of knowledge (for example, classical strings). The second property is that, under certain conditions on the parameters in the definition, a proof of knowledge protocol for a hard-to-clone state can also be used as a quantum money verification protocol. Intuitively, we might expect this to be the case, and indeed quantum money verification was one of the motivations which shaped our definition of a proof of quantum knowledge.

Our second main contribution is to provide two examples of protocols which can be shown to be proofs of knowledge under our definition, and in so doing introduce new techniques that may be used in the analysis of such protocols. As we have mentioned, quantum money verification protocols are natural candidates for proof-of-quantum-knowledge protocols: in a quantum money protocol, there is a prover who holds a purported money state, and who wishes to demonstrate to the verifier (who might be the bank or an independent citizen) that it does indeed 'hold' or 'possess' the quantum money state. The first person to describe quantum money was Wiesner [Wie83], who proposed money states that are tensor products of n qubits, each qubit of which is chosen uniformly at random from the set $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$. Wiesner's states can be described classically by $2n$ classical bits, and in a quantum money scheme this classical description is kept secret by the bank; a typical classical description is the pair of strings (x, θ), where the money state can be described (denoting by H_i a Hadamard gate on the ith qubit of the state) as $|\$\rangle_{x,\theta} = \prod_i H_i^{\theta_i} |x\rangle$. We choose to analyse a *private-key*, destructive classical money verification protocol between a prover and the bank for Wiesner's quantum money states which has been described previously in [MVW12]. The protocol is simply as follows: the verifier issues a uniformly random challenge string c to the prover, which encodes the bases (standard or Hadamard) in which the prover should measure the money state; the prover measures the ith qubit of the state in the standard basis if $c_i = 0$, or in the Hadamard basis if $c_i = 1$, and sends all the measurement outcomes as a string m to the verifier; and the verifier checks that, whenever $c_i = \theta_i$, $m_i = x_i$. The property which makes this protocol and these states interesting, of course, is that no prover who is given only one copy of the money state can pass verification twice.

Perhaps surprisingly, showing even that this simple protocol is a proof of knowledge according to our definition turns out to be a non-trivial task. (Note that the proof-of-knowledge property is either stronger than or incomparable to the property of being a money verification protocol, depending on parameter choices. The protocol under consideration was already shown to be a money verification protocol by [MVW12].) We may examine the following illustration of the difficulty. Consider, firstly, the following naïve approach to designing an extractor for the protocol described in the preceding paragraph. The extractor could, having access to the prover's initial internal state σ and entire circuit (implemented as a black box), pick a challenge c at random, apply the unitary which the prover would have applied in response to challenge c up until (and not including) the point where the prover would have measured the state, and then apply some unitary to 'correct' for the challenge bases in order to recover the original money state. However, the prover (upon receipt of the challenge) may take its honest money state and decide to apply X gates to some arbitrary subset of the qubits of the state which it was told to measure in the Hadamard basis, and Z gates to a subset of the qubits which it was told to measure in the standard basis. If the prover now measures this modified money state in the bases encoded by c, it will pass with probability 1—but, since it is with overwhelming probability the case that $c \neq \theta$ (the correct basis choice for the money state), the state that it measures in this scenario is almost certainly not the same state as the original money state.

A little thought will show that this is a fairly general obstacle to finding the money state in the prover's registers immediately before it performs the measurement whose outcomes it will send to the verifier. Since we know very little about what the prover might be doing to the money state at any other stage in its circuit, meanwhile, it is difficult to reason about finding the money state in the prover's registers at other points in its operation. This simple argument shows that, in order to design an effective extractor, it is crucial to consider the prover's responses to all challenges c at once—the question, of course, is one of how.

Our way of overcoming these difficulties introduces a novel technique which builds on results from the literature on nonlocal games. While this connection may be surprising at face value, it is not hard to see why a rigidity result from the self-testing literature might end up being a useful tool for extracting quantum states from uncooperative provers who only output classical strings. Rigidity results guarantee that, up to local isometry, two or more non-communicating players exhibiting certain correlations in their classical measurement outcomes must be performing particular quantum operations on a particular entangled state. The key idea of our proof for the Wiesner money verification protocol is as follows. Let the party which chooses and prepares the money state $|\$\rangle_{x,\theta} = \prod_i H_i^{\theta_i} |x\rangle$ that the prover receives be known as Alice, and let the prover be known as Bob. Consider the following thought experiment: instead of preparing $|\$\rangle_{x,\theta}$, Alice could prepare n EPR pairs and send half of each one to Bob. Let $E(\theta) = \{|\$\rangle\langle\$|_{x,\theta} \mid x \in \{0,1\}^n\}$ be a POVM. Then, if Alice measures $E(\theta)$ on her side of the state, and obtains the outcome x, Alice's and Bob's joint state will collapse to two copies of $|\$\rangle_{x,\theta}$. Note that, from Bob's perspective, the protocol is the same regardless of whether Alice sent EPR pairs and then measured $E(\theta)$, or whether she chose x and θ uniformly at random and sent him $|\$\rangle_{x,\theta}$ to begin with. However, if Bob succeeds with high probability in the money verification protocol, then he also succeeds with high probability at recovering a subset of the string x which represents Alice's measurement outcomes after she measures the POVM $E(\theta)$, and which also forms part of the classical description of the money state $|\$\rangle_{x,\theta}$. This observation makes it possible to apply a theorem from [NV16], which states that, if two noncommunicating parties exhibit correlations like those which Alice and Bob exhibit in this thought experiment, then they must once have shared EPR pairs, up to local isometry. Since Alice is honest and did nothing to her shares of the EPR pairs, the local isometry on her side is the identity map. Then, in order to recover the original money state, the proof-of-knowledge extractor simply has to compute the correct isometry on Bob's side (which can be done efficiently) and apply it to the state that Bob submits to the intermediary. A more detailed justification of this last sentence is given in section 5.

Wiesner states were the earliest and are the best-known kind of quantum money states, but there are other kinds, and one sort which has received some recent attention is the class of *subspace states* introduced in a quantum money context by [AC12]. Subspace states are states of the form $\frac{1}{\sqrt{|A|}} \sum_{x \in A} |x\rangle$ for some $n/2$-dimensional subspace $A \in \mathbb{Z}_2^n$, and they have similar no-cloning properties to those of Wiesner states; they are also of additional interest because they have been used in several schemes which make steps toward the goal of public-key quantum money [AC12], [Zha19], and in constructions of other quantum-cryptographic primitives such as quantum signing tokens [BDS16]. We were not able to find a simple classical verification protocol for subspace states that we could show to be a proof of quantum knowledge. Nonetheless, in Section 6, we propose a classical verification protocol for what we call *one-time-padded subspace states* (that is, subspace states which have had random Pauli one-time-pads applied to them by the bank), and we are able to show under our new definition, using similar techniques to those which we applied to Wiesner states, that this simple verification protocol is a proof of knowledge for one-time-padded subspace states. This verification protocol is remarkable for having a challenge from the verifier that is only one bit long.

Our final contribution is to show that, under our definition, a classical *argument* of quantum knowledge exists for any relation in the class QMA.[3] The notion of a *QMA relation* was formalised jointly by [BG19] and [CVZ19], as a quantum analogue to the idea of an *NP relation* which was described in the first paragraphs of this introduction. [BG19] and [CVZ19] show that any QMA relation has a *quantum* proof of quantum knowledge. The protocol that we show to be a *classical* argument of quantum knowledge for QMA relations, meanwhile, is the celebrated verification protocol introduced recently by [Mah18b]. Mahadev [Mah18b] shows, under cryptographic assumptions, that quantum properties (in her case, any language in BQP) can be decided by a classical polynomial-time verifier through classical interaction alone with a quantum polynomial-time prover. We note in retrospect that the proofs of the main results in [Mah18b] include statements which can be used to make the verification protocol which [Mah18b] introduces into a classical proof of quantum knowledge, in the same sense in which we have defined the latter. (In comparison, our proofs that specific quantum money schemes satisfy our definition of a proof of quantum knowledge do not use any cryptographic assumptions, and the protocols which we consider are very simple compared with the [Mah18b] protocol.) The [Mah18b] verification protocol can be shown to be an argument of quantum

[3] Argument systems differ from proof systems only in that the honest prover must be efficient, and that soundness is required to hold only against efficient provers. In this case, 'efficient' means quantum polynomial-time.

knowledge for any QMA relation; the only caveat, which was also a caveat for the *quantum* proofs of quantum knowledge for QMA exhibited by [BG19] and [CVZ19], is that an honest prover in the protocol may require multiple copies of a witness in order that the extractor can succeed in extracting *one* copy. We refer the reader to Section 7 for details.

Related and further work. Unruh [Unr12] was one of the first to consider the notion of a proof of knowledge in the quantum setting. In his work, as in the classical literature, a 'proof of knowledge' is a classical protocol which aims to establish that a prover 'knows' a *classical* string w satisfying an NP relation $R(x, \cdot)$. Unruh's work was novel because it was the first to consider the possibility that an adversarial prover may have quantum capabilities. This makes the design of an extractor more difficult, but Unruh shows that, under specific conditions, a classical proof of knowledge is automatically also sound against quantum adversaries.

As we have already mentioned, the notion of a proof of knowledge for a *quantum* relation was introduced recently in [BG19] and [CVZ19]. In these papers, the authors give a natural extension of the classical definition of a proof of knowledge to proofs of knowledge for *quantum* states ρ that 'satisfy' a QMA relation Q. (The statement 'ρ satisfies Q' here means that $Q(\rho, x) \geq \alpha$ for some parameter α, where Q is a quantum circuit, x is a classical problem instance, and $Q(\rho, x)$ is the probability that Q accepts witness ρ on input x.) These authors also show, building on earlier work by Broadbent et al. [BJSW16], that every QMA relation has a (quantum) proof of quantum knowledge. The main protocol from either paper can be made non-interactive (assuming the appropriate trusted setup), but all the protocols considered by these papers necessarily involve the exchange of quantum information between the verifier and the prover.

Other constructions for quantum money may yield to similar analyses to those by which we have approached the two examples we considered in sections 5 and 6; the construction based on the hidden matching problem by Gavinsky [Gav12] is one such candidate. As we noted in section 4, the connection that we make between proofs of quantum knowledge and quantum money protocols is somewhat unsatisfying in terms of parameters. We expect that Claim 4.4 can be tightened by considering sequential or parallel repetition, and we leave this question to further work. It would be interesting to find other quantum states (or collections of quantum states) in the space between quantum money states and witnesses for arbitrary QMA relations which admit simple proofs of quantum knowledge, with or without cryptographic assumptions. Looking toward applications, meanwhile, a natural candidate application for (non-interactive) proofs of quantum knowledge would be turning CPA-secure encryption schemes for quantum data into CCA-secure schemes. (Quantum CCA-secure schemes have already been constructed directly in [ABF$^+$16].)

It is also natural to consider *zero-knowledge* proofs of quantum knowledge. In our two examples in sections 5 and 6, the verifier is provided with secret classical information which completely specifies the state that the prover holds, so the notion of a zero-knowledge proof in this context is meaningless. However, for general QMA relations, and some applications, the idea of safeguarding the state against the verifier becomes more relevant. In prior work [VZ19], we showed that the protocol introduced in [Mah18b] can be made zero-knowledge. Since we show in the current work that [Mah18b] is an argument of quantum knowledge for any QMA relation, we believe zero-knowledge classical arguments of quantum knowledge can also be constructed for any QMA relation.

Organisation. The organisation of this paper is as follows. In section 2, we introduce some preliminary concepts. In section 3, we give our definitions of proofs of quantum knowledge, along with some intuition for our choices. In section 4, we prove some elementary properties of the definitions in section 3. In sections 5 and 6, we give proofs that a classical private-key quantum money verification protocol for Wiesner money states and a classical private-key verification protocol for one-time-padded subspace states, respectively, are proofs of quantum knowledge. Finally, in section 7, we show that any QMA relation has a classical argument of quantum knowledge.

Acknowledgements. We thank Andru Gheorghiu for useful feedback. Thomas Vidick is supported by NSF CAREER Grant CCF-1553477, AFOSR YIP award number FA9550-16-1-0495, a CIFAR Azrieli Global

Scholar award, MURI Grant FA9550-18-1-0161 and the IQIM, an NSF Physics Frontiers Center (NSF Grant PHY-1125565) with support of the Gordon and Betty Moore Foundation (GBMF-12500028).

2 Preliminaries

2.1 Terminology and notation

For definitions related to quantum circuits and basic quantum complexity classes such as BQP and QMA we refer to [Wat09]. We use 'QPT' as a shorthand for 'quantum polynomial time'. A quantum polynomial-time procedure is a polynomial-time uniformly generated family of quantum circuits.

We rely implicitly on Kitaev's circuit-to-Hamiltonian construction [KSVV02, KR03], which associates with any language $L \in \text{QMA}$ and $x \in \{0,1\}^*$ an instance of the *local Hamiltonian problem*. An instance of the local Hamiltonian problem is specified by a local Hamiltonian operator H and two real numbers α, β such that $\beta - \alpha \geq 1/\text{poly}(|x|)$ and H has smallest eigenvalue $\leq \alpha$ whenever $x \in L$, and $\geq \beta$ whenever $x \notin L$. We give the name 'ground state' to those states ρ such that $\text{Tr}(H\rho)$ takes on its minimum value, and we may refer to the space of all ground states of H (if there is more than one) as the 'ground space' of H. We call the minimum eigenvalue of H its 'ground energy'. We say H is 'gapped' when H is such that the gap between its ground energy and the next lowest energy level is a fraction of $\|H\|_{op}$ which is at least inverse-polynomial in $|x|$, where $\|H\|_{op}$ refers to the operator norm of H.

$|u|_H$ denotes the Hamming weight of a string $u \in \{0,1\}^n$, $\|u\|$ denotes the Euclidean norm of a vector $u \in \mathbb{C}^n$, and $\|u\|_1$ denotes its 1-norm. $\frac{1}{2}\|\rho - \rho'\|_1$ is the trace distance between two density matrices ρ and ρ'. We use the notation $\mathrm{d}_{|\psi\rangle}(A, B)$ to denote the distance (pseudo)metric $\|(A - B)\,|\psi\rangle\,\|^2$ between two operators A and B with respect to a specific state $|\psi\rangle$.

We use $\circ$ to denote composition: for example, if F and G are two circuits, $F \circ G(x)$ denotes firstly running G on x, and then running F on the output of G on x.

We use the notation $\mathsf{negl}(\lambda)$ to denote any negligible function of $\lambda \in \mathbb{N}$, i.e. a function f such that for any polynomial p, $p(\lambda)f(\lambda) \to_{\lambda\to\infty} 0$.

2.2 Quantum money

Public-key quantum money has received more recent attention than its private-key sibling, owing perhaps to the fact that the question of how to construct it is, at the time of writing, considered an interesting open problem. In our paper, however, we will focus on private-key quantum money, and so we provide only the definition of private-key quantum money here. The exposition below is taken with some modification from [AC12].

Definition 2.1. A *private-key quantum money scheme* $\mathcal{S}$ consists of three polynomial-time quantum algorithms:

- KeyGen, which takes as input a security parameter λ, and probabilistically generates a key k_{private}.
- Bank, which takes as input k_{private}, and probabilistically generates a quantum state \$ called a *banknote*. (Usually \$ will be an ordered pair (s, ρ_s), consisting of a classical *serial number* s and a *quantum money state* ρ_s, but this is not strictly necessary.)
- Ver, which takes as input k_{private} and an alleged banknote ¢ , and either accepts or rejects.

We say $\mathcal{S}$ has *completeness error* ε if $\mathsf{Ver}\left(k_{\text{private}}, \$\right)$ accepts with probability at least $1 - \varepsilon$ for all keys k_{private} and valid banknotes \$. If $\varepsilon = 0$ then $\mathcal{S}$ has *perfect completeness*.

Let Count (the *money counter*) take as input k_{private} as well as a collection of (possibly-entangled) alleged banknotes $¢_1, \ldots, ¢_r$, and output the number of indices $i \in \{1, \ldots, r\}$ such that $\mathsf{Ver}\left(k_{\text{private}}, ¢_i\right)$ accepts. Then we say $\mathcal{S}$ has *soundness error* δ if, given any quantum circuit $C\left(k_{\text{private}}, \$_1, \ldots, \$_q\right)$ of size poly(λ) (called the *counterfeiter*), which maps $q = \text{poly}(\lambda)$ valid banknotes $\$_1, \ldots, \$_q$ to $r = \text{poly}(\lambda)$ (possibly-entangled) alleged banknotes $\not{c}_1, \ldots, \not{c}_r$,

$$\Pr\Big(\mathsf{Count}\Big(k_{\text{public}}, C\left(k_{\text{public}}, \$_1, \ldots, \$_q\right)\Big) > q\Big) \leq \delta\ .$$

Here the probability is over the key choice k_{private}, valid banknotes $\$_1, \ldots, \$_q$ generated by $\mathsf{Bank}\left(k_{\text{private}}\right)$, and the behavior of Count and C.

We call $\mathcal{S}$ *secure* if it has completeness error $\leq 1/3$ and negligible soundness error.

Remark 2.2. By Theorem 16 in [AC12], it is sufficient for the security of a quantum money scheme to require that no counterfeiter C which attempts to map *one* valid banknote into *two* can succeed in causing the money counter to output 2 with more than negligible probability.

Remark 2.3. By Theorem 41 in [AC12], the completeness error 1/3 can be amplified to $1 - \exp(n)$ without materially affecting the other properties of the scheme, except that the soundness error will increase by a small amount.

3 Proofs of quantum knowledge

In the traditional view of an interactive protocol between a prover P and a verifier V, the two parties interact sequentially through a classical communication channel, and any party's (for example, the prover's) ith message $m_{P,i}$ is a function of the messages $\{m_{V,1}, \ldots, m_{V,i-1}\}$ it has received so far from the other party as well as its internal state. When one of the parties is quantum (again, for example, the prover), the function that computes $m_{P,i}$ may involve a measurement on a private quantum state. As we stated in the introduction, in order to allow an efficient extractor E to extract a quantum state from a prover that only exchanges classical messages with the verifier, we intend to grant E non-black-box access to the prover, since black-box extraction in the setting where all communication is classical seems as if it may be as hard as quantum state tomography. To make precise what the extractor has access to, we introduce an additional abstract party which we call the 'intermediary'. The intermediary allows the extractor to manipulate the state and measurements used by a real prover in the protocol in order to extract a certain state that the prover can then be argued to 'know'.

Specifically, for any real protocol $\mathcal{P}$, we define a *mediated protocol* $\mathcal{P}'$ between classical verifier and prover in the following form. In such a protocol, the prover does not directly compute its own messages, but is instead required to provide to the intermediary the following at the start of the protocol:

- a register X containing any state σ it might wish to use,
- a quantum black-box unitary C that takes as input a tuple (i, m, f, τ), where i is interpreted as a round number (each round consists in general of two classical messages, one from the verifier and one from the prover, in that order), m as the last message that the verifier sent (m might be $\perp$ for $i = 1$ if the prover sends the first message), $f \in \{-1, 1\}$ is an "inverse" flag, and τ is a quantum state. On all inputs of the form $|i, m, f\rangle\,|\varphi\rangle$ (where $|\varphi\rangle$ is a pure state having the same length in qubits as the size of the register X), C implements the transformation

$$C:\ |i, m, f\rangle\,|\varphi\rangle \mapsto |i, m, f\rangle\,(U_{i,m})^f\,|\varphi\rangle\ ,$$

  where $U_{i,m}$ represents the unitary that the prover would have implemented on its private space in the ith round of the protocol $\mathcal{P}$ upon reception of message m. For each i, it is understood that the prover's response in the ith round of the protocol $\mathcal{P}'$ will be obtained as follows. Starting from σ the prover applies U_{1,m_1}, where m_1 is the verifier's first message, and then measures the first k_1 qubits to obtain its first message m'_1.[4] Then, the prover applies U_{2,m_2}, where m_2 is the verifier's second message, on the

[4] The number of qubits measured by the prover in each step is part of the specification of the protocol, and known to all parties and, for simplicity, assumed independent of the verifier's messages.

post-measurement state obtained after the first round, and measures the first k_2 qubits to obtain its second message m_2'. The same process is then continued until the ith round in the obvious manner. We may sometimes refer to the combined operation of applying $U_{i,m}$ to the state in X and then measuring the first $k_{i,m}$ qubits of X as the act of performing the *measurement operator* $M_{i,m}$.[5] (Note that this formulation does not prevent the prover's strategy from depending on the verifier's messages other than its last: if it wishes to, the prover can record the verifier's historical messages in X.)

For the case of a QPT prover, we require that the prover pass to the intermediary an explicit polynomial-size quantum circuit that implements the black box C.

The intermediary then executes $\mathcal{P}$ with the verifier according to the prover's instructions (i.e., for each round i and each message m from the verifier, it applies $M_{i,m}$ to those qubits which the prover wishes $M_{i,m}$ to act on, and sends the classical measurement results to the verifier). After all the rounds of interaction are over, the verifier outputs 0 (reject) or 1 (accept).

Remark 3.1. In the definitions which follow, we make the requirement that the extractor E runs in polynomial time, if executing the prover's unitary C on any input counts as a unit-time procedure. This requirement is analogous to a standard requirement in classical definitions of proofs of knowledge, where running the prover only counts as a single timestep in the accounting of the extractor's complexity so that the notion of an efficient extractor is still well-defined when the prover is unbounded. For technical clarity, we require that (when the prover is unbounded) C is implemented as a black box. When the prover is QPT, meanwhile (which is always the case when the protocol in question is an *argument* instead of a proof, meaning that the extractor is only required to succeed on QPT provers), we can allow the extractor to have an explicit description of the circuit C, since C will be efficient when the prover is QPT, although in practice this will very rarely make a difference.

Remark 3.2. We stress that the intermediary exists only as an abstract formalism which allows us to precisely define the resources to which the extractor in Definitions 3.3, 3.5 and 3.8 has access. It does not constrain the prover's behaviour in any way—in fact, from the verifier's viewpoint, the protocols $\mathcal{P}$ and $\mathcal{P}'$ are exactly the same. In [GV19], a similar intermediary is introduced so that a particular classical protocol between quantum parties can be shown to be universally composable. If one wanted to use our definition of a proof of quantum knowledge in a composable framework, then it would likely prove necessary to make use of the intermediary in a similar way. We do not study composability questions in this paper.

We now proceed to introduce our definitions of classical proofs (and arguments) of quantum knowledge. We give multiple definitions that account for different settings in which the notion may prove useful. First, in Section 3.1 we give two definitions, Definition 3.3 and Definition 3.5, that introduce proofs of quantum knowledge for a single state and a collection of states respectively.[6] These definitions do not refer to any complexity classes, and ask that the extractor has the ability to recover a state that is close in trace distance to the target state. This choice of definitions is motivated by the applications to quantum money given in Section 5 and Section 6.

Second, in Section 3.2 we give a definition, Definition 3.8, that applies to *QMA relations*. Informally, a QMA relation is a quantum analogue for an NP relation; see Section 3.2 for the definition (which is due to [BG19] and [CVZ19]). In Definition 3.8 the extractor is not required to recover a specific state, but rather any valid witness according to the relation. In Section 7, we give an application of Definition 3.8 by showing that the measurement protocol introduced by [Mah18b] for classically verifying BQP statements can be considered an argument of quantum knowledge for any QMA relation.

[5] Formally, $M_{i,m}$ is specified by the collection of Kraus operators $\{K_{i,m,z}\}_{z\in\{0,1\}^{k_i}}$ where $K_{i,m,z} = (|z\rangle\langle z| \otimes \mathrm{Id})U_{i,m}$.

[6] In fact, Definition 3.3 is a special case of Definition 3.5; however, as we find that the notation and terminology that come along with Definition 3.5 can be somewhat cumbersome, we introduce Definition 3.3 as a simpler formulation about which useful properties (such as Claim 4.2) can still be proved. (Note that Claim 4.2 generalises easily to Definition 3.5 and Definition 3.8.)

3.1 Classical proofs of quantum knowledge for individual quantum states

In the following, words in square brackets should be *excluded* for a definition of a *proof* of quantum knowledge, and *included* for a definition of an *argument* of quantum knowledge.

Definition 3.3 (Classical proof [argument] of quantum knowledge for a state ρ)**.** Let $\{\rho_\lambda\}$ be a family of quantum states indexed by a security parameter λ. Let c, δ and κ be functions of λ. Let $\mathcal{P}$ be a protocol between a verifier V and a prover P (where both parties are given λ as input). We say that the protocol $\mathcal{P}$ is a (c, δ)*–proof [argument] of knowledge* with knowledge error κ for the quantum states $\{\rho_\lambda\}$ if:

- There exists a [QPT] prover P given ρ_λ as input such that the verifier accepts P with probability at least c, and
- There exists an extractor E satisfying the following: for any [QPT] prover P^* such that the verifier outputs 1 in the mediated version $\mathcal{P}'$ of $\mathcal{P}$ with probability $1 - \epsilon > \kappa$, E is able, given the C and σ that define P^* in $\mathcal{P}'$, to output a state ρ'_λ such that

$$\frac{1}{2}\|\rho_\lambda - \rho'_\lambda\|_1 \leq \delta,$$

  for a function δ of ϵ and λ such that $\delta < 1$ if $1 - \epsilon > \kappa$. Moreover, E runs in time polynomial in $|x|/(1 - \epsilon - \kappa)$, if executing the prover's unitary C on any input counts as a unit-time procedure.

Remark 3.4. The reader may wonder why, since Definition 3.3 refers to a single fixed ρ (for every value of the security parameter), the extractor E cannot simply look at the description of ρ and output ρ without reference to the prover's state σ or circuit C. In fact, it *could*, if—for example—ρ were specified by a classical description of a circuit which builds ρ; however, this is not necessarily the case. For example, ρ might be specified only by an obfuscated verification circuit (alternatively, a Hamiltonian) that accepts only states similar to ρ and rejects most others. Another useful intuitive formulation is to think of a protocol satisfying Definition 3.3 as a proof that the prover 'held onto' ρ: the prover may have received ρ from some trusted source some time ago, without knowing its description or anything about it except what it might be able to find out through minimally perturbative measurements. In performing the protocol with the verifier, the prover tries to demonstrate that it has not damaged ρ since it received the state or given it away.

Definition 3.5 (Classical proof [argument] of quantum knowledge for a finite collection $\mathcal{Q}$ of quantum states with respect to distribution D)**.** Let $\mathcal{Q}_\lambda = \{\rho_1, \rho_2, \ldots, \rho_{n(\lambda)}\}_\lambda$ be a family of collections of quantum states (one collection $\mathcal{Q}_\lambda$ for each value of the security parameter λ). From now on, we assume that λ is given as input to all parties, and omit the security parameter in the notation whenever its presence can be inferred from context.

Let D be a probability distribution over the set $\mathcal{Q}$ with density f, so that $\Pr_{\sigma \leftarrow D}(\sigma = \rho_i) = f(i)$. In addition, let $\mathcal{P}$ be a protocol between a verifier and a prover. Consider the following game involving prover, verifier, and intermediary.

1. The prover receives a state ρ_j sampled according to D.
2. After having received ρ_j, the prover executes $\mathcal{P}'$, the mediated version of $\mathcal{P}$, with the intermediary, who in turn interacts with the verifier. Let the state that the prover sends to the intermediary in $\mathcal{P}'$ after having received ρ_j be denoted by σ_j, and let the (black-box) circuit that it sends be denoted by C. [When the prover is computationally bounded, we require that, for all j, there is an efficient quantum operation T, also given to the extractor as a quantum circuit, which maps ρ_j to σ_j.] Let the verifier's eventual output in the mediated protocol $\mathcal{P}'$ be denoted by b_j.

The protocol $\mathcal{P}$ involving a verifier and a prover is a (c, δ)*–proof [argument] of knowledge* with knowledge error κ for the family of states $\mathcal{Q}$ with respect to the distribution D if:

- There exists a [QPT] prover P such that $\sum_{j=1}^{n} f(j)b_j \geq c$ in the mediated interaction between P and the honest verifier, and

- There exists an extractor E satisfying the following: for any [QPT] prover P^* such that

$$\sum_{j=1}^{n} f(j)\Pr[b_j = 1] = 1 - \epsilon > \kappa,$$

  E is able, for any state $\rho_j \in \mathcal{Q}$, and given the C, $[T]$ and σ_j that define the behaviour of P^* on input ρ_j, to output a state ρ'_j such that

$$\sum_{j=1}^{n} f(j)\frac{1}{2}\|\rho_j - \rho'_j\|_1 \leq \delta,$$

  for a function δ of ϵ and λ such that $\delta < 1$ if $1 - \epsilon > \kappa$. Moreover, E runs in time polynomial in $|x|/(1 - \epsilon - \kappa)$, if executing the prover's unitary C on any input counts as a unit-time procedure.

Finally, we introduce a definition which is related to Definitions 3.3 and 3.5, and for which we have not yet considered concrete applications, but which we believe may prove useful in time. Definition 3.6 defines an object which we call a *proof of possession.* The motivation for this object is as follows. Intuitively, Definition 3.3 guarantees that any prover who passes with high probability in a protocol satisfying the definition for a particular state ρ must *still hold*, in its entirety, some state close to the state ρ. This is formalised by the existence of an extractor which can construct a state close to ρ by making use of that prover's (black-box) circuit and internal state. However, as we show in the proof of Claim 4.2, classical proofs of quantum knowledge must be destructive. For some applications, it may therefore be useful to consider a protocol which does not guarantee that the prover still holds ρ, but only that it must *once* have come into contact with ρ—that is, that it can produce sensitive information which it is impossible or intractable to obtain without having measured ρ at some point before. To give an example, one protocol which would satisfy the definition of a proof of possession would be a version of Protocol 5.1 in which the challenge is chosen by the *prover* instead of uniformly at random by the verifier. Intuitively, we can hardly expect to prove that the prover has the entire money state $|\$\rangle_{x,\theta}$ if all it can do is pass in one challenge. However, we also expect that it is intractable to pass on *any* challenge c in Protocol 5.1 without having measured the state $|\$\rangle_{x,\theta}$ before. Definition 3.6 attempts to formalise this idea.

Definition 3.6 (Proof [argument] of possession)**.** Let λ be a security parameter given as input to all parties. Suppose that there exist a family of states $\{\rho_\lambda\}$, a quantum circuit $V_{\rho,\lambda}$, and a family of states $\{\rho'_\lambda\}$ such that:

- $V_{\rho,\lambda}(\rho'_\lambda) = 1$,
- for any [QPT] adversary $\mathcal{A}$ granted black-box access to $V_{\rho,\lambda}$, and any state σ_λ output by $\mathcal{A}$,

$$\Pr[V_{\rho,\lambda}(\sigma_\lambda \leftarrow \mathcal{A}) = 1] = \mathsf{negl}(\lambda),$$

- there exists a [QPT] procedure $\mathcal{B}$ which receives ρ_λ as input such that

$$\Pr[V_{\rho,\lambda}(\tau_\lambda \leftarrow \mathcal{B}) = 1] = 1 - \mathsf{negl}(\lambda).$$

Given $(\{\rho_\lambda\}, \{V_{\rho,\lambda}\}, \{\rho'_\lambda\})$ which satisfy the above properties, we call a proof [argument] of knowledge for the family of states $\{\rho'_\lambda\}$ a *proof [argument] of possession* for the family $\{\rho_\lambda\}$.

3.2 Classical proofs of quantum knowledge for QMA relations

We start by recalling the definition of a QMA relation, following [CVZ19, BG19]. Fix a uniformly generated family of polynomial-size quantum circuits $Q = \{Q_n\}_{n\in\mathbb{N}}$ such that for every n, Q_n takes as input a string

$x \in \{0,1\}^n$ and a quantum state σ on $p(n)$ qubits (for some polynomial $p(n)$) and returns a single bit as output. For any $0 \leq \gamma \leq 1$ define

$$R_{Q,\gamma} = \bigcup_{n\in\mathbb{N}} \left\{(x,\sigma) \in \{0,1\}^n \times \mathrm{D}(\mathbb{C}^{p(n)}) \,\middle|\, \Pr(Q_n(x,\sigma) = 1) \geq \gamma\right\}$$

and

$$N_{Q,\gamma} = \bigcup_{n\in\mathbb{N}} \left\{x \in \{0,1\}^n \,\middle|\, \forall \sigma \in \mathrm{D}(\mathbb{C}^{p(n)})\,,\ \Pr(Q_n(x,\sigma) = 1) < \gamma\right\} .$$

Definition 3.7 (QMA relation)**.** A *QMA relation* is specified by a triple (Q,α,β) where $Q = \{Q_n\}_{n\in\mathbb{N}}$ is a uniformly generated family of quantum circuits such that for every n, Q_n takes as input a string $x \in \{0,1\}^n$ and a quantum state $|\psi\rangle$ on $p(n)$ qubits and returns a single bit, and $\alpha, \beta : \mathbb{N} \to [0,1]$ are such that $\alpha(n) - \beta(n) \geq 1/p(n)$ for some polynomial p and all $n \in \mathbb{N}$. The QMA relation associated with (Q,α,β) is the pair of sets $R_{Q,\alpha}$ and $N_{Q,\beta}$.

We say that a *language* $L = (L_{yes}, L_{no})$ *is specified by a* QMA *relation* (Q,α,β) if

$$L_{yes} \subseteq \bigcup_{n\in\mathbb{N}} \left\{x \in \{0,1\}^n | \exists \sigma \in \mathrm{D}(\mathbb{C}^{p(n)}) \text{ s.t. } (x,\sigma) \in R_{Q,\alpha}\right\} , \tag{1}$$

and $L_{no} \subseteq N_{Q,\beta}$.

Note that, whenever $\alpha - \beta > 1/\operatorname{poly}(n)$, a language L that is specified by (Q,α,β) lies in QMA. Conversely, any language in QMA is specified by some QMA relation (of course such a relation is not unique).

In the following, words in square brackets should be *excluded* for a definition of a *proof* of quantum knowledge, and *included* for a definition of an argument of quantum knowledge.

Definition 3.8 (Classical proof [argument] of quantum knowledge for a QMA relation)**.** Let (Q,α,β) be a QMA relation. Let c, δ and κ be functions of a security parameter λ. Let $\mathcal{P}$ be a protocol between a verifier V and a prover P (where both parties are given λ and an instance $x \in \{0,1\}^*$ as input). We say that the protocol $\mathcal{P}$ is a (c,δ)*–proof [argument] of knowledge* with knowledge error κ for the QMA relation (Q,α,β) if:

- Whenever $(x,\sigma) \in R_{Q,\alpha}$ there exists a [QPT] prover P that is given σ as input and such that on common input (λ, x) the verifier accepts P with probability at least c;
- There exists an extractor E satisfying the following: for any [QPT] prover P^* such that the verifier outputs 1 in the mediated version of $\mathcal{P}$ on common input (λ, x) with probability $1-\epsilon > \kappa$, E is able, given the C and σ^* that define P^* in the mediated version of $\mathcal{P}$, to output a state σ' such that

  $$\Pr\left(Q_{|x|}(x,\sigma') = 1\right) \geq 1 - \delta ,$$

  for a function δ of ϵ and λ such that $\delta < 1$ if $1 - \epsilon > \kappa$, and such that $1 - \delta > \beta$. Moreover, E runs in time polynomial in $|x|/(1-\epsilon-\kappa)$, if executing the prover's unitary C on any input counts as a unit-time procedure.

Remark 3.9. Definition 3.3 can be considered a special case of Definition 3.8: if $Q(x,\cdot)$ is designed such that it only accepts states that are close in trace distance to a particular state (which could happen, for example, if $Q(x,\cdot)$ measured a Hamiltonian which was gapped and had a nondegenerate ground state), then Definition 3.8 reduces to Definition 3.3. We preserve both definitions for clarity and ease of use. In particular, note that Definition 3.3 does not require any classical problem instance x to be provided to the prover; this is more convenient in an application to quantum money, because in a quantum money setting the natural problem instance associated with a quantum money state is usually a description of the Hamiltonian of which the money state is a ground state, and giving this to the prover would allow for easy copying.

4 Simple properties

4.1 Nondestructive proofs of quantum knowledge are impossible for nontrivial states

In this section, we state and prove two simple properties of our definitions in section 3. The first of these is that *nondestructive* classical proofs of knowledge do not exist for nontrivial states (for the sense of 'nondestructive' made more precise by Definition 4.1). This is a simple no-go theorem which precludes certain types of classical proofs of quantum knowledge, and in particular makes it certain that classical proofs of quantum knowledge cannot be repeated (for example, for amplification purposes) unless the prover holds multiple copies of the state of interest.

Definition 4.1. If a measurement operator M acting on a quantum state ρ is such that ρ is left unchanged after the measurement M has been performed, we say that the measurement M was 'nondestructive'. A *nondestructive protocol* $\mathcal{P}$ is a protocol in which all the measurements $M_{i,m}$ that the intermediary performs for the prover in the mediated version of the protocol, $\mathcal{P}'$, are nondestructive for all possible m.

Claim 4.2. *If there is a prover P which is able to succeed in a classical nondestructive (c, δ)–proof (or argument) of knowledge for a family of states $\{\rho_\lambda\}$ with probability $1 - \epsilon$, then there is a procedure A* [7] *which, given one copy of the state ρ_λ, can produce polynomially many copies of a state ρ'_λ such that $\frac{1}{2}\|\rho'_\lambda - \rho_\lambda\|_1 \leq \delta(\epsilon)$.*

Proof. Suppose there is an efficient prover P which is able to succeed in a classical nondestructive proof of knowledge for $\{\rho_\lambda\}$ with probability $1 - \epsilon$. In each round, the prover performs a general measurement M_{i,m_i} on its quantum state (where m_i is the verifier's ith message) and sends the classical measurement outcome to the verifier. Let the prover's initial quantum state be σ, and let the response it sends to the verifier in the ith round be r_i. Since the extractor E exists, there must be an efficient isometry Φ (which can be constructed from the prover's circuit C) that acts on σ and produces $\rho'_\lambda \in \mathcal{H}_{main} \otimes \mathcal{H}_{aux}$, where ρ'_λ is a state such that $\text{Tr}_{aux}(\rho'_\lambda)$ is at trace distance at most δ away from ρ_λ. From now on, we will use the notation $\rho_\lambda \approx \rho'_\lambda$ to indicate that ρ_λ and $\text{Tr}_{aux}(\rho'_\lambda)$ have trace distance at most δ.

We claim that, since the measurements M_{i,m_i} which produced $r_1, \ldots, r_n$ (where the protocol has n rounds) were nondestructive, A can construct a state τ from only the classical messages $r_1, \ldots, r_n$ such that $\Phi(\tau) \approx \Phi(\sigma)$. Firstly, note that, because all of the measurements M_{i,m_i} were nondestructive, the outcomes r_i are entirely determined by the verifier's messages m_i and the prover's initial state σ. For any sequence of messages $m_1, \ldots, m_n$, consider the sequence of subspaces $S_{r_1(m_1)}, \ldots, S_{r_n(m_n)}$, where $r_i(m_i)$ is the response the prover gives to message m_i, and $S_{r_i(m_i)}$ is the eigenspace of the measurement operator $M_{i,m_i} \cdots M_{1,m_1}$ with eigenvalue r_i. Note that, since the protocol is nondestructive for all m, the prover's initial state σ must lie in the intersection $T_{r(m)} = \bigcap_i S_{r_i(m_i)}$ for all possible sequences of messages $m = m_1, \ldots, m_n$. This means that σ must lie in the intersection $\bigcap_m T_{r(m)}$. Conversely, if A constructs a state τ which lies in the intersection $\bigcap_m T_{r(m)}$, and then gives τ to a prover P', along with the (black-box) circuit C associated with P, P' can succeed in the protocol with probability $1 - \epsilon$ by applying C to τ in the same way that the original prover P would have applied it to σ. The protocol is a proof (or argument) of knowledge for $\{\rho_\lambda\}$, and P' is efficient when P is efficient, so we conclude that $\Phi(\tau) \approx \rho_\lambda$.

We conclude that A can be used to clone the states $\{\rho_\lambda\}$: for any value of λ, it can copy $r_1, \ldots, r_n$ into a new register, construct some state τ from them, and apply an isometry to τ to get a state ρ'_λ such that $\rho'_\lambda \approx \rho_\lambda$. □

Remark 4.3. It is straightforward to see that, while Claim 4.2 is stated as a property of Definition 3.3, it generalises to Definition 3.5 and Definition 3.8.

[7] A is in general not efficient. This is acceptable because the no-cloning theorem for general states holds information-theoretically.

4.2 Proofs of quantum knowledge are also quantum money verification protocols

The other simple property which we prove is that, under certain assumptions on the parameters in Definition 3.5, any protocol satisfying Definition 3.5 can be used as a quantum money verification protocol. Intuitively, one might expect this to be the case, and Claim 4.4 shows that it is. What may be less easy to see at first sight is that the requirements on the parameters of Definition 3.5 which Claim 4.4 makes are very demanding, considering that naive sequential repetition may not be an option for a quantum money verification protocol (owing to the fact that giving out multiple copies of a money bill could compromise no-cloning security). In fact, for example, the parameters we get from the proofs in sections 5 and 6 do *not* satisfy these requirements. We leave the consideration of how these requirements may be relaxed, or else how a version of sequential repetition might be developed for quantum money schemes, to the realm of future work.

Claim 4.4. *Let λ be a security parameter. Let $\mathcal{P}$ be a protocol between a prover and a verifier, and let the prover's success probability in this protocol be $1-\epsilon$. Suppose there is a negligible function $\nu(\cdot)$ and a function $\delta(\cdot,\cdot)$ such that*

1. *$\delta(\epsilon,\lambda) < \left(\delta_0 = \frac{2}{3+\sqrt{5}}\right)$ for all (λ,ϵ) such that*

   (a) $1-\epsilon = f(\lambda)$ is a non-negligible function of λ, and

   (b) $\lambda > M_f$, where $M_f \in \mathbb{N}$ is a bound that is allowed to depend on f;

2. *$\mathcal{P}$ is a $(c = 2/3, \delta)$–proof of knowledge protocol with knowledge error $\nu(\lambda)$ for a collection of states $\mathcal{Q}_\lambda = \{\rho_1,\ldots,\rho_{n(\lambda)}\}_\lambda$ such that, given one copy of a state ρ chosen from distribution D over $\mathcal{Q}_\lambda$, it is [computationally intractable / impossible] to produce two states σ_1 and σ_2 which both pass in a verification protocol Ver_ρ with non-negligible probability.*

Then $\mathcal{P}$ is also a quantum money verification protocol for a money scheme secure against [computationally bounded / all] counterfeiters.

Remark 4.5. Claim 4.4 can be used in its stated form when we already have a verification protocol Ver_ρ for the family of money states $\mathcal{Q}$ and we wish to prove that a different protocol is also a verification protocol for $\mathcal{Q}$. If we are worried that Claim 4.4 is not particularly useful when we already have a verification protocol Ver_ρ, we can alter the property $\mathcal{Q}$ must satisfy into the following: given one copy of a state ρ chosen from distribution D over $\mathcal{Q}_\lambda$, it is [computationally intractable / impossible] to produce two states σ_1 and σ_2 which are both [computationally / statistically] indistinguishable from ρ. Then, if we require $\delta = \mathsf{negl}(\lambda)$ as well as $\delta < \delta_0$, and note that we can replace Ver_ρ in the proof below with any [computationally bounded] adversary, we will get the modified claim.

Alternatively, we can also require that no [efficient] procedure can generate σ_1 and σ_2 which are both within a certain trace distance of ρ. This case is actually easier to prove (given the nature of Definition 3.5) than the case we have worked out in the proof below and the case of [computational] indistinguishability whose proof we have sketched in the previous paragraph: it comes directly out of the definition of the extractor's success. (The reason there is a difference between the other two cases and this one is that this case does not preclude entanglement between σ_1 and σ_2 which causes σ_2 to fail with higher probability when σ_1 succeeds, or vice versa.)

Proof. We prove Claim 4.4 by contradiction: we assume that there is a prover P who can pass twice in the proof-of-knowledge protocol $\mathcal{P}$ with sufficiently high probability, given only one copy of the state ρ drawn from D, and then we show a reduction to an adversary who can produce σ_1 and σ_2 which both pass with non-negligible probability in Ver_ρ. Let there be a prover P who receives one copy of ρ and who can pass twice in the protocol $\mathcal{P}$ with probability greater than $\nu(\lambda) = \kappa$ in each execution. Note that, although Definition 3.5 states that the prover receives ρ each time the protocol is run, we can model a prover P who only receives one copy of ρ and must pass in the protocol twice as two communicating provers, P_1 and P_2, who play the game in Definition 3.5 sequentially, and specify that P_2 discards its copy of ρ without using it, but that P_2 is allowed to receive communications in the form of a quantum state σ_m from P_1. Consider an adversary A for no-cloning which proceeds as follows.

1. A runs the extractor E guaranteed by Definition 3.5 on the state σ which P_1 provides to its intermediary. (We can assume, without loss of generality, that E is an isometry, i.e. it does not trace out any qubits.) Since $\mathcal{P}$ is a $(1-\mu(\lambda),\delta)$–proof of knowledge, the extractor is able to output a state $\tau_1 \in \left(\mathcal{H} = \mathcal{H}_\sigma \otimes \mathcal{H}_{work} = \mathcal{H}_{main} \otimes \mathcal{H}_{aux}\right)$ (where $\mathcal{H}_\sigma$ is the Hilbert space containing P_1's state σ, and $\mathcal{H}_{work}$ is the extractor's workspace, and $\mathcal{H}_{main}$ and $\mathcal{H}_{aux}$ repartition $\mathcal{H}$ into a space containing the extractor's useful output and auxiliary output) such that, on average over the choice of ρ, $\frac{1}{2}\|\rho - \mathrm{Tr}_{aux}(\tau_1)\|_1 \leq \delta$. (From now on, we omit the qualifier 'on average over the choice of ρ' when its presence can be inferred from context.)

2. A submits $\mathrm{Tr}_{aux}(\tau_1)$ to Ver_ρ for quantum money verification. Let the state inside $\mathcal{H}_{main} \otimes \mathcal{H}_{aux}$ after verification occurs be τ_2. Ver_ρ must accept with probability at least $1-\delta$; then, by the gentle measurement lemma, $\frac{1}{2}\|\tau_1 - \tau_2\|_1 \leq \sqrt{\delta}$ if Ver_ρ accepts.

3. A runs the extractor E backwards on the state τ_2. Let $E^{-1}(\tau_2) = \tau_3$.

4. A gives $\mathrm{Tr}_{work}(\tau_3)$ back to P_1, allows P_1 to finish executing and pass its quantum message to P_2, and then runs the extractor on P_2. Let σ_m denote the message that P_2 would have received from P_1 if step 2 in this list had never happened, and let σ'_m denote the message that P_2 does receive. By Definition 3.5, $E \circ P_2(\sigma_m)$ is at trace distance at most δ from ρ. Then, by the contractivity of the trace distance, $\frac{1}{2}\|E \circ P_2(\sigma_m) - E \circ P_2(\sigma'_m)\|_1 \leq \frac{1}{2}\|\tau_1 - \tau_2\|_1 \leq \sqrt{\delta}$, and, applying the triangle inequality, $\frac{1}{2}\|\rho - E \circ P_2(\sigma'_m)\|_1 \leq \sqrt{\delta} + \delta$. A then submits $E \circ P_2(\sigma'_m)$ for quantum money verification.

As long as $\delta < \delta_0$, $\sqrt{\delta} + \delta < 1$, and the probability that $E \circ P_2(\sigma'_m)$ passes verification is $1 - (\sqrt{\delta} + \delta)$, which can be lower-bounded by the nonzero constant $1 - (\sqrt{\delta_0} + \delta_0)$. The adversary A will then pass twice in quantum money verification with probability at least $(1-\delta)(1-(\sqrt{\delta}+\delta))$ on average over the choice of ρ, which completes the reduction.

□

5 PoQK for Wiesner money states

Our first concrete example of a proof of quantum knowledge protocol is a verification protocol for Wiesner's quantum money states. As we recalled in the introduction, quantum money states in Wiesner's scheme are n-qubit states such that each qubit is chosen uniformly at random from the set $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$. Any such state can be described classically by $2n$ classical bits; a typical classical description is the pair of strings (x, θ), where the money state can be described (denoting by H_i a Hadamard gate on the ith qubit of the state) as $|\$\rangle_{x,\theta} = \prod_i H_i^{\theta_i} |x\rangle$. Here, n is equated with the security parameter λ in Definition 3.5, and the distribution $\mathcal{D}$ is taken to be uniform over all $|\$\rangle\langle\$|_{x,\theta}$ for $x, \theta \in \{0,1\}^n$.

We now describe our proof of knowledge protocol.

Protocol 5.1. The following is a destructive verification protocol for Wiesner money states.

1. The verifier sends an n-bit uniformly random string c to the prover, where n is the length of the money state $\rho = |\$\rangle\langle\$|_{x,\theta}$ in qubits.

2. If the ith bit of c is 0, the prover measures the ith qubit of its money state in the standard basis; and if it is 1, the prover measures the ith qubit in the Hadamard basis. The prover sends all of the measurement outcomes it obtains in this way to the verifier.

3. Let s be an n-bit string representing the bases in which the money state has been measured nondestructively. That is, $s_i = 1$ if and only f $c_i = \theta_i$. The verifier checks that, whenever $s_i = 1$, the outcome is what it should be, i.e. x_i.

Lemma 5.2. *Fix some $n \in \mathbb{N}$. There is a constant $\kappa < 1$ such that Protocol 5.1 is a $(c = 1, \delta = O(\epsilon^{1/4}))$–proof of knowledge with knowledge error κ for the family of states $\mathcal{F} = \{|\$\rangle\langle\$|_{x,\theta} \mid x, \theta \in \{0,1\}^n\}$ with respect to the uniform distribution over $\mathcal{F}$.*

Proof. Suppose that we are given a prover who passes in Protocol 5.1 with probability $1-\mu$. Let the party which chooses and prepares the Wiesner money state that the prover receives in step 1 of Definition 3.5 be known as Alice, and let the prover be known as Bob. Consider the following thought experiment: instead of preparing $|\$\rangle_{x,\theta}$, Alice prepares n EPR pairs and sends half of each pair to Bob. For any $\theta \in \{0,1\}^n$ let $E(\theta) = \{|\$\rangle\langle\$|_{x,\theta} \,|\, x \in \{0,1\}^n\}$ be a POVM. Then, if Alice measures $E(\theta)$ on her remaining n half-EPR pairs and obtains the outcome x, Alice's and Bob's joint state collapses to two copies of $|\$\rangle_{x,\theta}$. Note that, from Bob's perspective, the protocol is the same regardless of whether Alice sent EPR pairs and then measured $E(\theta)$, or whether she sent him $|\$\rangle_{x,\theta}$ to begin with.

Suppose that we associate a register A with Alice's n qubits, and a register B with Bob's n qubits. Let $T : \mathcal{H}_B \to \mathcal{H}_{B'}$ be the map applied by Bob upon reception of his n qubits from Alice. Let the shorthand $\sigma_Z(a)$, for some string $a \in \{0,1\}^n$, denote the n-qubit observable that is a tensor product of single qubit observables, and which is σ_Z on those qubits i such that $a_i = 1$, and I otherwise. Let $\sigma_X(a)$ be defined similarly. Define $Z^A(b) \equiv \sigma_Z(b)$, $X^A(a) \equiv \sigma_X(a)$.

For Bob, meanwhile, we assume WLOG that the data he submits to the intermediary consists of the first $n+m$ qubits of $\mathcal{H}_{B'}$, together with a circuit that, upon receipt of challenge $c \in \{0,1\}^n$, applies (again, without loss of generality) an arbitrary unitary U_c on those $n+m$ qubits and then measures the first n qubits of $\mathcal{H}_{B'}$ in the basis indicated by c (computational basis if $c_i = 0$ and Hadamard basis if $c_i = 1$). For any $c \in \{0,1\}^n$, therefore, we define $Z^B(c) \equiv U_{\bar{c}}^*(\sigma_Z(c) \otimes \mathrm{Id})U_{\bar{c}}$, and $X^B(c) \equiv U_c^*(\sigma_X(c) \otimes \mathrm{Id})U_c$, where $\bar{c}$ denotes the bitwise complement of c, and where the identities act on the last m qubits (so that the σ operators act on the first n only).

Suppose that Alice measures $E(\bar{b})$ on her side of the state, for some $b \in \{0,1\}^n$, and obtains the outcome x. (This is equivalent to creating $|\$\rangle_{x,\bar{b}}$.) By hypothesis, we know that $U_{\bar{b}}$ applied to $\mathrm{Tr}_A(|\psi_{AB}\rangle)$ must produce a state that, when measured in the standard basis, will yield x_i whenever $b_i = 1$, except with probability μ over the choice of b. Let β be the outcome that Bob obtains from measuring his state in the standard basis after applying $U_{\bar{b}}$. Then, except with probability μ,

$$\begin{aligned} &\forall i, b_i x_i = b_i \beta_i \\ \Longrightarrow\ &\bigoplus_i b_i x_i = \bigoplus_i b_i \beta_i . \end{aligned}$$

Since $E(\bar{b})$ commutes with $Z^A(b) \equiv \sigma_Z(b)$, and Bob's standard basis measurement commutes with $Z^B(b) \equiv U_{\bar{b}}^* \sigma_Z(b) U_{\bar{b}}$, it must be the case that

$$\begin{aligned} \mathop{\mathrm{E}}_b \mathrm{Tr}\big(\langle\psi|\, Z^A(b) \otimes Z^B(b)\, |\psi\rangle \big) &= \mathop{\mathrm{E}}_b \mathrm{Tr}\big(\langle\psi|\, \sigma_Z(b) \otimes U_{\bar{b}}^* \sigma_Z(b) U_{\bar{b}}\, |\psi\rangle \big) \\ &\leq (1)(1-\mu) + (-1)\mu = 1 - 2\mu , \end{aligned} \tag{2}$$

where the expectation is taken under the uniform distribution over $b \in \{0,1\}^n$.

Now suppose instead that Alice measures $E(a)$, for some $a \in \{0,1\}^n$, and obtains the outcome x. Then U_a applied to $\mathrm{Tr}_A(|\psi_{AB}\rangle)$ must produce a state that, when measured in the Hadamard basis, will yield x_i whenever $a_i = 1$, except with probability μ. Similarly, then, we have

$$\begin{aligned} \mathop{\mathrm{E}}_b \mathrm{Tr}\big(\langle\psi|\, X^A(a) \otimes X^B(a)\, |\psi\rangle \big) &= \mathop{\mathrm{E}}_b \mathrm{Tr}\big(\langle\psi|\, \sigma_X(a) \otimes U_{\bar{a}}^* \sigma_Z(a) U_{\bar{a}}\, |\psi\rangle \big) \\ &\leq 1 - 2\mu . \end{aligned} \tag{3}$$

At this point we are ready to apply the following lemma, whose proof is given at the end of the section. (The lemma is an adaptation of results that appeared in [NV16].) In the lemma, we let $|\mathrm{EPR}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ denote an EPR pair.

Lemma 5.3. *Let $|\psi\rangle_{AB'} = (\mathrm{Id} \otimes \Gamma_B)(|\mathrm{EPR}\rangle_{AB}^{\otimes n})$, for an arbitrary CPTP map $\Gamma_B : \mathcal{H}_B \to \mathcal{H}_{B'}$. Suppose that for every $a, b \in \{0,1\}^n$ there exist observables $\{X^B(a)\}$, $\{Z^B(b)\}$ on $\mathcal{H}_{B'}$ such that*

$$\forall W \in \{X, Z\} , \qquad \mathop{\mathrm{E}}_a \big\| \big(\sigma_W^A(a) - W^B(a)\big)\, |\psi\rangle \big\|^2 \leq \varepsilon , \tag{4}$$

for some $0 \leq \varepsilon \leq 1$. Then there exists an isometry $\Phi_B : \mathcal{H}_{B'} \to ((\mathbb{C}^2)^{\otimes n})_B \otimes \mathcal{H}_{\hat{B}} \oplus \mathcal{H}_{\hat{B}'}$ and a state $|aux\rangle_{\hat{B}}$ on $\mathcal{H}_{\hat{B}}$ such that

$$\mathrm{Tr}\Big(\big(\langle \mathrm{EPR}|_{AB}^{\otimes n} \otimes \langle aux|_{\hat{B}}\big)\big(\mathrm{Id} \otimes \Phi^B(|\psi\rangle\langle\psi|_{AB'})\big)\big(|\mathrm{EPR}\rangle_{AB}^{\otimes n} \otimes |aux\rangle_{\hat{B}}\big)\Big) = 1 - O\big(\varepsilon^{1/2}\big). \tag{5}$$

Moreover, given black-box implementations for unitaries C_1 and C_2 that apply $X^B(a)$ and $Z^B(b)$ to a state τ, given as input τ as well as strings a and b respectively, it is possible to construct an explicit quantum circuit that implements Φ_B by making a constant number of calls to C_1 and C_2 and uses $O(n)$ additional gates. Finally, the lemma also holds, with the same conclusion (but weaker implied constants), if (4) holds when the expectation is restricted to those strings a such that $|a|_H = \frac{n}{2}$.

Setting $\varepsilon = 4\mu$, it follows from (2) and (3) that equation (4) of Lemma 5.3 is satisfied. Suppose now that Alice measures $E(\theta)$ after preparing the EPR pairs, and obtains the outcome x. Eq (5) implies (by the contractivity of the trace distance, and the equality $\|\rho - \sigma\|_1 = \sqrt{1 - \mathrm{Tr}(\rho\sigma)}$ that holds when ρ, σ are pure) that

$$\|\Phi_B \mathrm{Tr}_A\big(|\$\rangle\langle\$|_{x,\theta} \otimes T(|\$\rangle\langle\$|_{x,\theta})\big) - |\$\rangle\langle\$|_{x,\theta} \otimes |aux\rangle\langle aux|_{\hat{B}}\|_1 = O(\varepsilon^{1/4})\ .$$

It follows that the extractor can apply Φ_B (which, according to Lemma 5.3, has an explicit circuit description as a function of the circuits that the prover gave to the intermediary) to σ and trace out $\hat{B}$ in order to recover a state $O(\varepsilon^{1/4})$ close to $|\$\rangle_{x,\theta}$. This completes the proof.

□

We conclude the section by giving the proof of Lemma 5.3. The proof uses the following general lemma, that is based on [GH17]. See e.g. [CS17, Lemma 4.7] for a proof. In the lemma, $\mathrm{U}(\mathcal{H})$ denotes the group of unitaries acting on Hilbert space $\mathcal{H}$.

Theorem 5.4. *Let G be a finite group. Let $f : G \mapsto \mathrm{U}(\mathcal{H}_{B'})$ and $|\psi\rangle_{AB'} \in \mathcal{H}_A \otimes \mathcal{H}_{B'}$ be such that*

$$\mathop{\mathrm{E}}_{x,y\in G} \langle\psi|\, \mathrm{Id}_A \otimes f(x)f(yx)^\dagger f(y)\, |\psi\rangle \geq 1 - \delta\ , \tag{6}$$

for some $\delta > 0$. Then there exists an isometry $V : \mathcal{H}_{B'} \to \mathcal{H}_{B''}$ and a representation $g : G \mapsto \mathrm{U}(\mathcal{H}_{B''})$ such that

$$\mathop{\mathrm{E}}_{x\in G} \big\|\, \mathrm{Id}_A \otimes \big(f(x) - V^\dagger g(x) V\big)\, |\psi\rangle\, \big\|^2 \leq 2\delta\ . \tag{7}$$

In this paper we are particularly concerned with the Pauli group $G = \mathcal{P}_n$, which can be defined as the $2 \cdot 4^n$-element matrix group generated by the n-qubit Pauli matrices σ_X and σ_Z, i.e.

$$\mathcal{P}_n = \big\{ \pm\sigma_X(u)\sigma_Z(v) \big|\, u, v \in \{0,1\}^n \big\}\ . \tag{8}$$

We represent each element of $\mathcal{P}^n$ by a triple $(\varepsilon, u, v) \in \{\pm 1\} \times \{0,1\}^n \times \{0,1\}^n$ in the natural way. In particular, the identity element is $e = (1, 0^n, 0^n)$. For $x = (\varepsilon, u, v) \in \mathcal{P}_n$, let $f(x) = \varepsilon \hat{X}^B(u)\hat{Z}^B(v)$.

In this case and under the additional assumption that $f(-x) = -f(x)$ for all $x \in \mathcal{P}_n$ the isometry V promised in Theorem 5.4 takes a particularly simple form: for $|\varphi\rangle \in \mathcal{H}_{B'}$, we have $\mathcal{H}_{B''} = \mathbb{C}^{2^n} \otimes \mathbb{C}^{2^n} \otimes \mathcal{H}_B$ and (as can be verified from e.g. the proof given in [CS17])

$$V|\varphi\rangle = \frac{1}{\sqrt{2 \cdot 2^{2n}}} \sum_{x=(\varepsilon,u,v)\in\mathcal{P}_n} \big(\big(\varepsilon\sigma_X(u)\sigma_Z(v)\big) \otimes \mathrm{Id}\, |\mathrm{EPR}\rangle^{\otimes n}\big) \otimes f(x)\, |\varphi\rangle\ .$$

In particular, an efficient circuit for V is easily constructed from an efficient circuit for applying f controlled on x.

Proof of Lemma 5.3. Since the "finally" part of the lemma, where (4) is only promised to hold on average over strings of Hamming weight $\frac{n}{2}$, implies the initial statement of the lemma, we directly show that part.

For a string $u \in \{0,1\}^n$ define two random strings (a,b) of Hamming weight $n/2$ as follows. First assume that the Hamming weight of u is even. Let $S \subseteq \{1,\dots,n\}$ denote the location of the '1' entries in u. Let S_1 be a uniformly random subset of S of size $|S|/2$, and S_2 a uniformly random subset of $\overline{S}$ of size $|\overline{S}|/2$. Set all entries of a in $S_1 \cup S_2$ to '1', and all other entries to '0'. Set $b = u \oplus a$. Then it is clear that a and b both have Hamming weight $\frac{n}{2}$. Furthermore, if u is uniformly distributed then a and b are each uniformly distributed over all strings of Hamming weight $\frac{n}{2}$. If the Hamming weight of u is odd, a similar construction, flipping an additional coin to decide if $|S_1| = (|S| \pm 1)/2$, applies.

For $u \in \{0,1\}^n$ define $\hat{X}^B(u) = X^B(a)X^B(b)$, where a and b are generated according to the process described above, independently for each u. Similarly, for $v \in \{0,1\}^n$ define $\hat{Z}^B(v) = Z^B(a)X^B(b)$, with a, b generated independently for each v. Then $\{\hat{X}^B(u)\}$ and $\{\hat{Z}^B(v)\}$ are unitaries on $\mathcal{H}_{B'}$.

For $x = (\varepsilon, u, v) \in \mathcal{P}_n$, let $f(x) = \varepsilon \hat{X}^B(u)\hat{Z}^B(v)$.

Claim 5.5. *The following holds:*

$$\mathop{\mathrm{E}}_{x,y,z\in\mathcal{P}^n} \langle\psi|\, \mathrm{Id}_A \otimes f(x)f(yx)^\dagger f(y)\,|\psi\rangle \geq 1 - O(\varepsilon)\;, \tag{9}$$

where the expectation is over three uniformly random group elements.

Proof. First we observe that the claim holds, with no error, in case f is replaced by $f^A(x) = \varepsilon\sigma_Z^A(v)\sigma_X^A(u)$, for $x = (\varepsilon, u, v)$. Next we note that for all $W \in \{X, Z\}$ it holds that

$$\mathop{\mathrm{E}}_{u\in\{0,1\}^n} \big\| \big(\sigma_W^A(u) - \hat{W}^B(u)\big)\,|\psi\rangle\,\big\|^2 = O(\varepsilon)\;. \tag{10}$$

Indeed, this follows by two applications of (4), the triangle inequality, and the fact that the marginal distributions of a and b for u chosen uniformly at random are both uniform. Applying (10) twice, once for $W = Z$ and once for $W = X$, we get

$$\mathop{\mathrm{E}}_{\epsilon\in\{\pm1\},u,v\in\{0,1\}^n} \big\| \big(f^A(\varepsilon,u,v) - f(\varepsilon,u,v)\big)\,|\psi\rangle\,\big\|^2 = O(\varepsilon)\;. \tag{11}$$

Given that f^A satisfies (9), this concludes the proof. □

Claim 5.5 shows that the function f satisfies the assumption of Theorem 5.4, for some $\delta = O(\varepsilon)$. Let g be the representation promised by the theorem. Let $g = g_+ \oplus g_-$ where $g_+(-e) = \mathrm{Id}$ and $g_-(-e) = -\mathrm{Id}$ (recall that e denotes the neutral element of $\mathcal{P}_n$). The only irreducible representation of $\mathcal{P}_n$ that does not send $-e$ to the identity is the Pauli matrix representation $\tau_{\mathcal{P}}$ that we used to define the group in (8). Thus $g_- = \tau_{\mathcal{P}} \otimes \mathrm{Id}_d$, for some integer d. Let d' be the dimension of g_+ and Π_+ the projection on its range. Using that by definition $f(-x) = -f(x)$ for all $x \in G$, we get from (7) using the triangle inequality that

$$\begin{aligned}\mathop{\mathrm{E}}_{x\in\mathcal{P}_n} \big\| \mathrm{Id}_A \otimes V^\dagger 2\Pi_+ V\,|\psi\rangle\,\big\|^2 &= \mathop{\mathrm{E}}_{x\in\mathcal{P}_n} \big\| \mathrm{Id}_A \otimes V^\dagger\big(g(x) + g(-x)\big)V\,|\psi\rangle\,\big\|^2 \\ &= O(\varepsilon)\;. \end{aligned} \tag{12}$$

In particular, it then follows from (7) that

$$\mathop{\mathrm{E}}_{x\in\mathcal{P}_n} \big\| \mathrm{Id}_A \otimes \big(f(x) - V^\dagger\big((\tau_{\mathcal{P}}(x)\otimes \mathrm{Id}_d)\oplus 0_{d'}\big)V\,|\psi\rangle\,\big\|^2 = O(\varepsilon)\;. \tag{13}$$

Let $|\psi'\rangle = (\mathrm{Id} \otimes V)\,|\psi\rangle$. Using the assumption (4) twice and (13) we get that for $W \in \{X, Z\}$,

$$\mathop{\mathrm{E}}_{u\in\{0,1\}^n} \langle\psi'|\,\sigma_W^A(u) \otimes \big((\sigma_W^B(u)\otimes \mathrm{Id}_d)\oplus 0_{d'}\big)\,|\psi'\rangle \;=\; 1 - O\big(\sqrt{\varepsilon}\big)\;. \tag{14}$$

It is easy to verify by direct calculation that

$$\mathop{\mathrm{E}}_{u,v}\, \sigma_X^A(u)\sigma_Z^A(v) \otimes \sigma_X^B(u)\sigma_Z^B(v) \;=\; |\mathrm{EPR}\rangle\langle\mathrm{EPR}|^{\otimes n}\;.$$

It then follows from (14) that $|\psi'\rangle = |\mathrm{EPR}\rangle^{\otimes n}\,|aux\rangle + |\psi''\rangle$ for some $|\psi''\rangle$ such that $\|\,|\psi''\rangle\,\|^2 = O(\varepsilon)$. The lemma follows, setting $\Phi_B(X) = VXV^\dagger$ for all X. □

6 PoQK for subspace money states

Our second example of a proof of quantum knowledge protocol is a verification protocol for a modification of Aaronson's and Christiano's *subspace states* [AC12]. Aaronson and Christiano present a quantum money scheme in which an n-qubit money state is specified by a (secret) $(n/2)$-dimensional subspace $A \in \mathbb{Z}_2^n$, and defined as $|A\rangle = \frac{1}{\sqrt{|A|}} \sum_{x \in A} |x\rangle$. (We identify n with the security parameter λ in Definition 3.5.) Aaronson and Christiano give a proof of no-cloning security for a scheme that uses these subspace states as money bills, in a black-box model where the prover can only access A through a pair of measurement operators that respectively implement projections on A and $A^\perp$. Their objective in defining such a scheme is to make progress towards public-key quantum money.

As we mentioned in the introduction, we do not know if it is possible to devise a natural proof of quantum knowledge for the Aaronson-Christiano subspace states as they have thus far been described. What makes finding such a protocol challenging is that, in contrast with Wiesner's money scheme, there is no obvious classical verification protocol for subspace states in which there can be a single 'right answer' to a challenge. We may consider, for example, a classical verification protocol for these states similar to a protocol which was considered in [BDS16], where the prover is asked to measure all the qubits of a subspace state in either the standard or the Hadamard basis, and any vector $x \in A$ (resp. $y \in A^\perp$) is a valid outcome for a measurement in the standard (resp. Hadamard) basis. It is difficult to argue that such a protocol is a PoQK for subspace states using similar techniques to those which we used for Wiesner states, because the large number of possible 'right answers' which the verifier would accept means that the correlations that the prover and the verifier must share if the prover passes are much weaker than those for which we can argue in Protocol 5.1. Nonetheless, we are able to give a proof of knowledge for a version of the subspace scheme in which a (secret) quantum one-time pad has been applied to every subspace state. This protocol has the noteworthy property that the challenge issued by the verifier is only a single bit long.

Protocol 6.1. Let A be a uniformly random $(n/2)$-dimensional subspace of $\mathbb{Z}_2^n$, and let $A^\perp$ be the orthogonal complement of A. Let d, e be strings in $\{0,1\}^n$, which we identify with elements of $\mathbb{Z}_2^n$. Let the shorthand $X(a)$, for some string $a \in \{0,1\}^n$, denote the n-qubit unitary that is a tensor product of single qubit gates, and which is Pauli X on those qubits i such that $a_i = 1$, and I otherwise. Let $Z(a)$ be defined similarly. The following is a destructive verification protocol for one-time-padded subspace states of the form $X(d)Z(e)|A\rangle = \frac{1}{\sqrt{|A|}} X(d)Z(e) \sum_{x \in A} |x\rangle$.

- The verifier sends a single-bit challenge $c \in \{0,1\}$ to the prover.
- If the challenge is 0, the prover measures the entire subspace state in the standard basis, obtaining an n-bit string of outcomes $m \in \{0,1\}^n$, and sends m to the verifier. If the challenge is 1, the prover measures the subspace state in the Hadamard basis and likewise sends the outcomes m to the verifier.
- If $c = 0$, the verifier checks that $m \oplus d$ is in A. If $c = 1$, the verifier checks that $m \oplus e$ is in $A^\perp$.

Lemma 6.2. *For any $n \in \mathbb{N}$ and any subspace A of $\mathbb{Z}_2^n$ define $|A\rangle = \frac{1}{\sqrt{|A|}} \sum_{x \in A} |x\rangle$. There exists a constant $\kappa < 1$ such that Protocol 6.1 is a $(c = 1,\ \delta = O(\epsilon^{1/5}))$–proof of knowledge with knowledge error κ for the set $\{X(d)Z(e)|A\rangle \,:\, d, e \in \{0,1\}^n, \dim(A) = n/2\}$, with respect to the uniform distribution over this set.*

Proof. Suppose that the probability that a given prover passes in Protocol 6.1 is $1 - \mu$. Let the party which chooses and prepares the subspace state that the prover receives in step 1 of Definition 3.5 be known as Alice, and let the prover be known as Bob. Instead of preparing $X(d)Z(e)|A\rangle$, imagine that Alice prepares n EPR pairs and sends half of each pair to Bob, keeping the other half. For convenience, we introduce some notation:

- Let $|\$\rangle_{x,\theta}$ be a Wiesner money state representing the string x encoded in bases θ.
- Let $\{s_i : i \in \{1, \ldots, n\}\} = \{100...0, 010...0, 001...0, \ldots, 000...1\}$ be the standard basis for $\mathbb{Z}_2^n$.

- Let $\mathcal{Z} = \{z_i : i \in \{1, \ldots, n\}\}$ be a uniformly random basis for $\mathbb{Z}_2^n$ chosen by Alice.

- Let V be the unitary defined as follows:

$$\begin{aligned} V : V\left|x\right\rangle &= V\left|x_1 s_1 + \cdots + x_n s_n\right\rangle \\ &= \left|x_1 z_1 + \cdots + x_n z_n\right\rangle \ . \end{aligned} \tag{15}$$

- Let L_θ for a string $\theta \in \{0,1\}^n$ be the subspace of $\mathbb{Z}_2^n$ whose elements are always 0 in the positions where $\theta_i = 0$, and can be either 0 or 1 in the positions where $\theta_i = 1$.

- Let $X(z)$ for some vector $z = (z_1, \ldots, z_n) \in \mathbb{Z}_2^n$ denote the tensor product of n single-qubit gates which is Pauli X in those positions i where $z_i = 1$, and I otherwise. Define $Z(z)$ similarly. Let $X_{\mathcal{Z}}(a)$, for a basis $\mathcal{Z} = \{z_j\}$, denote the operator

$$\prod_j \left(X(z_j)\right)^{a_j} ,$$

  where z_j denotes a particular vector from the basis set $\mathcal{Z}$, and a_j denotes the jth bit of a. Define $Z_{\mathcal{Z}}(a)$ similarly.

For $\theta \in \{0,1\}^n$ and a basis $\mathcal{Z}$ of $\mathbb{Z}_2^n$, let $E(\theta, \mathcal{Z}) = \{V\left|\$\right\rangle\!\left\langle\$\right|_{x,\theta} V^* \mid x \in \{0,1\}^n\}$ be a POVM, where V is defined as a function of $\mathcal{Z}$ in (15). Then, if Alice measures $E(\theta, \mathcal{Z})$ on her side of the state, and obtains the outcome x, Alice's and Bob's joint state collapses to two copies of

$$\left|\$'\right\rangle_{x,\theta,\mathcal{Z}} = \frac{1}{\sqrt{|L_\theta|}} \sum_{\lambda \in L_\theta} X_{\mathcal{Z}}(d) Z_{\mathcal{Z}}(e) \left|\lambda_1 z_1 + \cdots + \lambda_n z_n\right\rangle ,$$

with $d_i = x_i$ for i such that $\theta_i = 0$, and $e_i = x_i$ for i such that $\theta_i = 1$. Note that the distribution of $\left|\$'\right\rangle$ over uniform $x, \theta, \mathcal{Z}$ is identical (ignoring global phase) to that of a uniformly random one-time-padded subspace state. (The global phase enters because the constraint above on d_i and e_i in terms of x and θ only specifies half of the $2n$ coordinates of d and e together.)

Suppose that we associate the register A with Alice, and the register B with Bob. For simplicity, we assume that Alice's register contains n qubits, and that Bob's contains $n + m$ for some arbitrary $m \geq 0$ ancilla qubits. Alice prepares n EPR pairs and sends half of each one to Bob; Bob then behaves as the prover in Protocol 6.1 would, applying a CPTP map T to his side of the state. Denote the resulting bipartite state $\left|\psi\right\rangle_{AB'}$.

For $a, b \in \{0,1\}^n$ define $Z^A(b) \equiv V\sigma_Z(b)V^*$, $X^A(a) \equiv V\sigma_X(a)V^*$. For Bob, meanwhile, consider $Z^B(b) \equiv U_0^* V(\sigma_Z(b) \otimes \mathrm{Id})V^* U_0$ (where, for $c \in \{0,1\}$, U_c is the unitary that the prover in Protocol 6.1 applies to his side of the state in response to challenge c, and where $\sigma_Z(b)$ acts on the first n qubits and Id acts on the last m), and $X^B(a) \equiv U_1^* V(\sigma_X(a) \otimes \mathrm{Id})V^* U_1$.

Suppose that Alice measures $E(\bar{b}, \mathcal{Z})$ on her side of the state, and obtains the outcome x. (This is equivalent to creating $\left|\$'\right\rangle_{x,\bar{b}}$.) If the prover succeeds in Protocol 6.1 with probability $1 - \mu$ on average over the choice of $\mathcal{Z}$ then the probability that the prover succeeds conditioned on $c = 0$ is at least $1 - 2\mu$, on average over the choice of $\mathcal{Z}$. 'Success' in the $c = 0$ case means that the prover obtains, after measuring the state in the standard basis, an outcome m such that $m \oplus (x \cdot b) \in A$. Equivalently, 'success' means that, if the prover had measured in the basis defined by the POVM $\{V\left|x\right\rangle\!\left\langle x\right| V^* \ \mid x \in \{0,1\}^n\}$, it would have obtained a string β such that, $\forall i, b_i x_i = b_i \beta_i$, because elements of the subspace A expressed in the $\mathcal{Z}$ basis have zeroes in the positions i where $b_i = 1$. Then, except with probability 2μ over the choice of b and the choice of $\mathcal{Z}$,

$$\begin{aligned} &\forall i, b_i x_i = b_i \beta_i \\ \implies &\bigoplus_i b_i x_i = \bigoplus_i b_i \beta_i \ . \end{aligned}$$

$E(\bar{b}, \mathcal{Z})$ commutes with $Z^A(b)$, and the POVM measurement $\{V\,|x\rangle\langle x|\,V^* \ \mid x \in \{0,1\}^n\}$ commutes with $Z^B(b)$. Therefore, it must be the case that

$$\mathop{\mathrm{E}}_{\mathcal{Z}}\Big[\mathop{\mathrm{E}}_{b\,:\,|b|_H=n/2} \mathrm{Tr}\big(\langle\psi|\,Z^A(b)\otimes Z^B(b)\,|\psi\rangle\big)\Big] \leq (1)(1-2\mu) + (-1)2\mu = 1-4\mu\ ,$$

for a uniformly random choice of $\mathcal{Z}$. (From now on, we will omit the qualifier '$|b|_H = n/2$' in the expectation over b, although it remains implicit.) Using the notation $\mathrm{d}_{|\psi\rangle}(A,B) = \|(A-B)\,|\psi\rangle\,\|^2$ for any A, B, we get

$$\begin{aligned}\mathop{\mathrm{E}}_{\mathcal{Z}}\mathop{\mathrm{E}}_{b}\ \mathrm{d}_{|\psi\rangle}(Z^A(b), Z^B(b))^2 &= \mathop{\mathrm{E}}_{b}\ \big(2 - 2\mathrm{Tr}\big(\langle\psi|\,Z^A(b)\otimes Z^B(b)\,|\psi\rangle\big)\\ &\leq 2 - 2(1-4\mu)\\ &= 8\mu\ .\end{aligned}$$

Now suppose that Alice measures $E(a, \mathcal{Z})$, and obtains the outcome x. By similar reasoning to the above, we have

$$\begin{aligned}\mathop{\mathrm{E}}_{\mathcal{Z}}\mathop{\mathrm{E}}_{a}\ \mathrm{d}_{|\psi\rangle}(X^A(a), X^B(a))^2 &= \mathop{\mathrm{E}}_{a}\ \big(2 - 2\,\langle\psi|\,X^A(a)\otimes X^B(a)\,|\psi\rangle\big)\\ &\leq 2 - 2(1-4\mu)\\ &= 8\mu\ .\end{aligned}$$

Let $d_{\mathcal{Z},X} \equiv \mathrm{E}_a\,\mathrm{d}_{|\psi\rangle}(X^A(a), X^B(a))^2$, and let $d_{\mathcal{Z},Z} \equiv \mathrm{E}_b\,\mathrm{d}_{|\psi\rangle}(Z^A(a), Z^B(a))^2$. Using Markov's inequality, for any $k > 1$ we have that

$$\begin{aligned}\Pr_{\mathcal{Z}}\big(d_{\mathcal{Z},X} \geq k\mu\big) &\leq \frac{8}{k}\ ,\\ \Pr_{\mathcal{Z}}\big(d_{\mathcal{Z},Z} \geq k\mu\big) &\leq \frac{8}{k}\ .\end{aligned}$$

It is clear that the assumption (4) of Lemma 5.3 is satisfied (with $\epsilon = k\mu$) when $d_{\mathcal{Z},X} \leq k\mu$ and $d_{\mathcal{Z},Z} \leq k\mu$, which occurs with probability at least $1 - \frac{16}{k}$ over the choice of $\mathcal{Z}$. (Note that here we use the "finally" part of the lemma, which allows us to restrict the condition to uniform expectation over strings of Hamming weight $\frac{n}{2}$. Note further that the lemma requires Alice's operators to be exact Pauli operators, while here they are conjugated by the unitary V. Since this unitary can be shifted to Bob's system using the relation $(W \otimes \mathrm{Id})\,|\mathrm{EPR}\rangle = (\mathrm{Id} \otimes W^T)\,|\mathrm{EPR}\rangle$ for any W, the lemma still applies.)

The conclusion of the lemma implies that, if a basis $\mathcal{Z}$ is chosen uniformly at random, then with probability at least $1 - \frac{16}{k}$, there exists an isometry $\Phi^B_{\mathcal{Z}}$ such that

$$\mathrm{Tr}\Big(\big(\langle\mathrm{EPR}|^{\otimes n}_{AB} \otimes \langle aux|_{\hat{B}}\big)\,\big(\mathrm{Id}_A \otimes \Phi^B_{\mathcal{Z}}(|\psi\rangle\,\langle\psi|_{AB'})\big)\big(\,|\mathrm{EPR}\rangle^{\otimes n}_{AB} \otimes |aux\rangle_{\hat{B}}\big)\Big) = 1 - O\big((k\epsilon)^{1/2}\big)\ . \qquad (16)$$

For the moment, let $\mathcal{Z}$ be chosen so that such $\Phi^B_{\mathcal{Z}}$ exists. Suppose then that Alice measures $E(\theta, \mathcal{Z}')$ after preparing the EPR pairs, and obtains the outcome x. Equation (16) implies (by the contractivity of the trace distance, and the definition $\|\rho - \sigma\|_1 = \sqrt{1 - \mathrm{Tr}(\rho\sigma)}$ that holds when ρ, σ are pure) that

$$\|\Phi^B_{\mathcal{Z}}\mathrm{Tr}_A\big(\,|\$'\rangle_{x,\theta,\mathcal{Z}'} \otimes T(|\$'\rangle_{x,\theta,\mathcal{Z}'})\big) - |\$'\rangle_{x,\theta,\mathcal{Z}'}\,|aux\rangle_{\hat{B}}\,\|_1 = O((k\epsilon)^{1/4}). \qquad (17)$$

Then, for the remaining $\frac{16}{k}$ fraction of $\mathcal{Z}$s where an isometry $\Phi^B_{\mathcal{Z}}$ satisfying equation (16) may not exist, the trace distance between any state the extractor outputs and $|\$'\rangle_{x,\theta,\mathcal{Z}'}\,|aux\rangle_{\hat{B}}$ is still upper bounded by 1. By the convexity of the trace distance, the trace distance between the density matrix which the extractor outputs and the state $|\$'\rangle_{x,\theta,\mathcal{Z}'}\,|aux\rangle_{\hat{B}}$ is at most

$$\begin{aligned}&\frac{16}{k} + \Big(1 - \frac{16}{k}\Big)O((k\epsilon)^{1/4})\\ &\leq \frac{16}{k} + O((k\epsilon)^{1/4})\ .\end{aligned}$$

Choosing $k = (1/\varepsilon)^{1/5}$, this expression is $O(\varepsilon^{1/5})$. It follows that the extractor can apply Φ_B to σ and trace out $\hat{B}$ in order to recover a state $O(\epsilon^{1/5})$ close to $|\$'\rangle_{x,\theta}$. □

7 Arguments of Quantum Knowledge for QMA relations

The main result of this section is Theorem 7.4, which gives a classical argument of quantum knowledge for any QMA relation. (If we want the completeness property of this argument of knowledge to hold for a prover that is given a single copy of a witness for the relation, we need to assume a certain structure for the QMA relation; see the statement of Theorem 7.4 below.) The construction is based on the classical verification protocol for QMA introduced in [Mah18b], so we start by reviewing that protocol.

7.1 The verification protocol

We recall the high-level structure of the verification protocol from [Mah18b]. In this protocol, which we will refer to as the *verification protocol*, the input to the verifier is an n-qubit Hamiltonian H that is expressed as a linear combination of tensor products of σ_X and σ_Z Pauli operators. The input to the prover is a ground state of H. Both parties also receive a security parameter λ. At a high level, the verification protocol has the following structure:

1. The verifier selects a *basis string* $h \in \{0,1\}^n$ according to a distribution that depends on H. The verifier then randomly samples a pair of keys, (pk, sk), consisting of a public key pk and secret key sk. (The distribution according to which (pk, sk) is sampled depends on h.) The choice of keys specifies an integer $w = \text{poly}(n, \lambda)$. The verifier sends pk to the prover.

2. The prover returns an n-tuple of *commitment strings* $y = (y_1, \ldots, y_n)$, where each y_i lies in some alphabet $\mathcal{Y}$.

3. The verifier selects a *challenge bit* $c \in \{0,1\}$ and sends c to the prover.

4. If $c = 0$ ("test round"), the prover returns a string $b \in \{0,1\}^n$ and $x_1, \ldots, x_n \in \{0,1\}^w$. If $c = 1$ ("Hadamard round"), the prover returns a string $b \in \{0,1\}^n$ and $d_1, \ldots, d_n \in \{0,1\}^w$.

5. In case $c = 0$ the verifier uses pk, y, b and $x_1, \ldots, x_n$ to make a decision to accept or reject. (In a test round the verifier never checks any properties of the prover's state; it only checks that the prover is, loosely speaking, doing the correct operations.) In case $c = 1$ the verifier uses sk to decode y, b and $d_1, \ldots, d_n$ into *decoded measurement outcomes* $(m_1, \ldots, m_n) \in \{0,1\}^n$. (For the case of a honest prover, the decoded outcomes m correspond to the outcomes of measuring a ground state of H in the bases indicated by h, with $h_i = 0$ indicating that the ith qubit should be measured in the computational basis and $h_i = 1$ that the ith qubit should be measured in the Hadamard basis. The prover remains ignorant throughout the entire protocol of the verifier's choice of h.)

6. In case $c = 1$ the verifier makes a decision based on the decoded measurement outcomes and the instance x, as described in [Mah18c, Protocol 8.1].

As explained in [Mah18c, Section 2.3.1] the behavior of an arbitrary prover in this protocol can be characterized by a pair of unitaries (U_0, U) such that:

1. Upon reception of pk, the prover applies U_0 to its initial state (to which $|pk\rangle$ has been appended), measures the first $n \log |Y|$ qubits in the computational basis and returns the outcome;

2. Upon reception of $c = 0$, the prover directly measures the first (remaining) $n + nw$ qubits in the computational basis and returns the outcome;

3. Upon reception of $c = 1$, the prover applies the unitary U, measures the first (remaining) $n + nw$ qubits in the Hadamard basis and returns the outcome.

When executing this protocol as a mediated protocol (see Section 3) we may therefore assume without loss of generality that the information passed by the prover to the intermediary consists of (i) an initial state σ, and (ii) an explicit circuit implementation of the unitaries U_0 and U.

7.2 Construction of an extractor

To show that the protocol described in the previous section can be made into an argument of quantum knowledge for a QMA relation we first explain, in general terms, how to construct the extractor E. We first quote a claim from [Mah18c]. To make the claim comprehensible we need the following definition.

Definition 7.1. A prover in the verification protocol, represented by an initial state σ and a pair of unitaries (U_0, U), is called "trivial" (implicitly, for a given input Hamiltonian to the protocol) if two conditions hold: (i) its probability of being accepted in a test round (case $c = 0$) is negligibly close to 1, and (ii) the prover's unitary U commutes with a computational basis measurement of the first n qubits (i.e. the committed qubits).

Claim 7.2. *For all trivial provers P, there exists an n-qubit state ρ [which can be created from the prover's initial state using a polynomial-size quantum circuit] such that for all $h \in \{0,1\}^n$, the distribution over measurement results produced in the protocol with respect to P for basis choice h is computationally indistinguishable from the distribution which results from measuring ρ in the basis determined by h.*

As will soon be made clear, the "polynomial-size circuit" referred to in the statement of the claim is a circuit that is obtained by small modifications of the provers' actions in the protocol (i.e. the unitaries U_0 and U described in the previous section). The claim of "computational indistinguishability" is based on a cryptographic assumption, that underlies soundness of the verification protocol from [Mah18b]: informally, the distribution of outcomes is "indistinguishable" from measurements on ρ *to any (classical or quantum) computationally bounded adversary*, assuming that the Learning with Errors (LWE) problem is intractable for quantum polynomial-time procedures. We refer to [Mah18b] for more details on this computational assumption.

For our purposes we need a slightly modified version of Claim 7.2, stated below.

Claim 7.3. *There exist constants $c_1, C_1 > 0$ such that the following holds. Let H be a Hamiltonian. Let P be a prover that is accepted in the verification protocol associated with H, conditioned on a test round ($c = 0$), with probability $1 - \varepsilon$, for some $\varepsilon \geq 0$. Then there exists an n-qubit state ρ (which can be created from the prover's initial state using a polynomial-size quantum circuit) such that on average over $h \in \{0,1\}^n$ sampled by the verifier at step 1 of the protocol, the distribution over decoded measurement outcomes obtained by the verifier at step 5 of the protocol (in case $c = 1$) is computationally indistinguishable from some distribution on $\{0,1\}^n$ that is within statistical distance at most $C_1\varepsilon^{c_1}$ from the distribution which results from measuring ρ in the basis determined by h.*

We sketch a proof of this claim by relying on the proof of Claim 7.2 given in [Mah18c]. In the course of the proof we explain the terminology used in the statement of Claim 7.2.

Proof of Claim 7.3. As shown in [Mah18c], it is straightforward to construct a trivial prover from an arbitrary prover that succeeds with probability sufficiently close to 1 in the "test" part of the protocol without affecting the prover's answers in the "Hadamard" part by too much (assuming the prover has a high overall probability of success in the first place). First, we can obtain (i) because the prover itself can check if the answer it would provide in a test round would be accepted,[8] so a modified prover, if it so desires, can repeatedly simulate the original prover until it obtains a commitment string y such that the resulting post-measurement state would lead to acceptance in a test round. The argument to obtain (ii) is slightly more complicated, and appears as the proof of [Mah18c, Claim 7.3]. There it is argued that, for the case of a prover that satisfies (i), replacing U by a twirling (conjugation) of it by random Pauli σ_Z operators on the first n qubits does not affect the decoded outcomes obtained by the verifier in a way that would be efficiently noticeable.

In the context of a proof (or argument) of quantum knowledge, we would like the extractor to be able to turn the arbitrary prover from which it must extract a witness into a trivial prover, so that we can then apply

[8] This is not the case for answers in the Hadamard round, that require the secret key sk to be verified.

Claim 7.2. Unfortunately, the modification required for (i) that we have described requires polynomially many copies of a ground state of H. The intermediary, from which the extractor must be built, only receives a single copy of the prover's initial state, so that the modification we have sketched for turning an arbitrary prover into a trivial prover cannot be applied by the extractor under Definition 3.8.

Fortunately we note that the requirement that the prover succeeds with probability negligibly close to 1 in the test round is not necessary; instead, it is sufficient to require that the prover succeeds with probability sufficiently close to 1, where "sufficiently" can be an inverse polynomial in the instance size. To see why this is the case we observe that performing the modification for (ii) to a prover that only succeeds with probability $1-\varepsilon$ in the test round, as opposed to negligibly close to 1, will only affect the verifier's decoded measurement outcomes up to statistical distance $\mathrm{poly}(\varepsilon)$. This is because, as shown in the proof of [Mah18c, Claim 7.2] the transformation from a general prover to one that satisfies (i) only affects the internal state of the prover at the start of step 4 of the protocol to within trace distance at most $\mathrm{poly}(\varepsilon)$. Since the twirling operation has a negligible effect on the verifier's decoded measurement outcomes for the case of a prover satisfying (i), the same operation cannot lead to outcomes that are distinguishable from the original outcomes with advantage more than $\mathrm{poly}(\varepsilon)$ when applied to a prover that succeeds with probability only $1-\varepsilon$, instead of negligibly close to 1. Removing the assumption that the prover is trivial also requires us to weaken the statement "for all h" in Claim 7.2 to "on average over h" here; however, this is unimportant as the verifier can use a distribution on bases h of its choice at the first step of the protocol (and in particular the distribution can depend on the structure of the local Hamiltonian instance), and the "average" referred to in the claim is then taken with respect to the verifier's choice of distribution. □

To complete our explanation of Claim 7.2 we describe, following [Mah18c, Protocol 5.8], the construction of the state ρ. (We adapt the provedure so that it applies to a general prover, as opposed to a trivial prover in [Mah18c, Protocol 5.8]. The only modification is the twirling of U at step 3.)

1. Sample (pk, sk) as V would for the case when h is the all-1 string, $h = 1^n$;

2. Using pk, apply the prover's first unitary U_0 on the prover's initial state and measure the first $n \log |\mathcal{Y}|$ qubits to obtain commitment strings $(y_1, \ldots, y_n)$;

3. Apply the unitary U, twirled by a random σ_Z operator on each of the first n qubits. Do not measure the (remaining) first n qbuits. (These qubits are referred to as *committed qubits.*) Measure the following nw qubits in the Hadamard basis to obtain a string d.

4. For $i \in \{1, \ldots, n\}$, apply $\sigma_Z^{d_i \cdot (x_{0,i} \oplus x_{1,i})}$ to the i-th committed qubit. Here $x_{0,i}$ and $x_{1,i}$ are w-bit strings that can be inferred from y_i using sk, the inner product is taken modulo 2, and $\sigma_Z^0 = \mathrm{Id}$.

5. Return the state ρ of the n committed qubits.

Given efficient circuits for U_0 and U, the entire extractor construction is efficient.

7.3 Arguments of Quantum Knowledge for QMA relations

Combining the observations made in the previous subsections, we show the following.

Theorem 7.4. *Let (Q, α, β) be a* QMA *relation that has the following properties.*

1. *The completeness parameter α is negligibly close to* 1*, and the soundness parameter β is bounded away from* 1 *by an inverse polynomial.*

2. *For any $x \in R_{Q,\alpha}$ with $|x| = n$, there is a local Hamiltonian $H = H_x$, that is efficiently constructible from x, satisfying the following. First, we assume that H is expressed as a linear combination of tensor products of Pauli operators such that $-\mathrm{Id} \leq H \leq \mathrm{Id}$. Second, whenever there is σ such that $(x, \sigma) \in R_{Q,\alpha}$, then $\mathrm{Tr}(H\sigma)$ is negligibly close to -1 and moreover any σ such that $\mathrm{Tr}(H\sigma) \leq -1 + \delta$ satisfies $\Pr(Q_{|x|}(x, \sigma) = 1) \geq 1 - r(|x|)q(\delta)$ for some polynomials q, r. Third, whenever $x \in N_{Q,\beta}$ then the smallest eigenvalue of H is larger than -1 by some inverse polynomial in $|x|$.*

Then under the Learning with Errors assumption the verification protocol presented in Section 7.1 is a (c, s, δ)-argument of quantum knowledge for the family of states $\{(x, \sigma) : (x, \sigma) \in R_{Q,\alpha}\}$ with knowledge error κ, where: c is negligibly close to 1; κ is bounded away from 1 by an inverse polynomial; δ is a polynomial in ε (for any ε such that $1 - \varepsilon > \kappa$).

Before giving the proof of the theorem we comment on the assumptions made on the QMA relation. The first assumption is benign and follows from standard amplification techniques. The second assumption is somewhat more restrictive. For any QMA relation (Q, α, β), the existence of Hamiltonians $H = H_x$ satisfying all claimed properties follows from Kitaev's circuit-to-Hamiltonian construction, in its amplified form used in [Mah18c, Protocol 8.3]. However, this construction is not in general "witness-preserving" in the sense described in the second assumption above: to construct an eigenstate of the Hamiltonian H with small enough eigenvalue, one may need to use many copies of a witness σ for the QMA verification procedure $Q(x, \cdot)$. Hence depending on the structure of the original QMA relation one may or may not in general be able to obtain, using Theorem 7.4, an argument of quantum knowledge whose completeness property holds for a prover that is given a single witness for the QMA relation.

Proof. The completeness requirement follows immediately from completeness of the protocol from Section 7.1, as shown in [Mah18b]. The construction of the extractor is described at the start of this section. Fix an $(x, \sigma) \in R_{Q,\alpha}$ and let $n = |x|$. Suppose that P^* is a prover that succeeds with probability $1 - \varepsilon > \kappa$ in the associated protocol, where κ is as in the theorem statement. In particular, the prover is accepted in a test round with probability at least $1 - 2\varepsilon$. By Claim 7.3, measurement outcomes obtained on the extracted state ρ' are computationally indistinguishable from a distribution on n-bit strings that is within statistical distance at most $\mathrm{poly}(\varepsilon)$ from those obtained by the verifier at step 5 of the verification protocol. We claim that the extractor returns a state ρ' such that

$$\Pr(Q_n(x, \rho') = 1) \geq 1 - \mathrm{poly}(\varepsilon) \cdot \mathrm{poly}(|x|) \,. \tag{18}$$

The reason is that, as shown in [Mah18c, Section 8] (based on [FHM18]), the verifier's decision at step 6 of the verification protocol involves an efficient computation on the decoded measurement outcomes. Moreover, whenever the measurement outcomes are truly obtained from measurements on a state ρ', then the verifier's acceptance probability at step 6. is $\frac{1}{2}(1 - \mathrm{Tr}(H\rho'))$. Therefore, the aforementioned computational indistinguishability implies that the quantity $\mathrm{Tr}(H\rho')$ is within $\delta = \mathrm{poly}(\varepsilon)$ of the verifier's acceptance probability in the verification protocol. By the second property assumed in the theorem statement, this implies the claimed bound (18). Assuming κ chosen to be sufficiently close to 1, ε is small enough that the right-and side of (18) is positive. □

References

[ABF$^+$16] Gorjan Alagic, Anne Broadbent, Bill Fefferman, Tommaso Gagliardoni, Christian Schaffner, and Michael St Jules. Computational security of quantum encryption. In *International Conference on Information Theoretic Security*, pages 47–71. Springer, 2016.

[AC12] Scott Aaronson and Paul Christiano. Quantum money from hidden subspaces. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 2012.

[BDS16] Shalev Ben-David and Or Sattath. Quantum tokens for digital signatures. *arXiv preprint arXiv:1609.09047*, 2016.

[BG92] Mihir Bellare and Oded Goldreich. On defining proofs of knowledge. In *Annual International Cryptology Conference*, pages 390–420. Springer, 1992.

[BG19] Anne Broadbent and Alex B Grilo. Zero-knowledge for qma from locally simulatable proofs. *arXiv preprint arXiv:1911.07782*, 2019.

[BJSW16] Anne Broadbent, Zhengfeng Ji, Fang Song, and John Watrous. Zero-knowledge proof systems for QMA. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 31–40. IEEE, 2016.

[CL06] Melissa Chase and Anna Lysyanskaya. On signatures of knowledge. In Cynthia Dwork, editor, *Advances in Cryptology - CRYPTO 2006*, pages 78–96, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[CS17] Andrea Coladangelo and Jalex Stark. Robust self-testing for linear constraint system games. *arXiv preprint arXiv:1709.09267*, 2017.

[CVZ19] Andrea Coladangelo, Thomas Vidick, and Tina Zhang. Non-interactive zero-knowledge arguments for QMA, with preprocessing. *arXiv preprint arXiv:1911.07546*, 2019.

[FFS88] Uriel Feige, Amos Fiat, and Adi Shamir. Zero-knowledge proofs of identity. *Journal of Cryptology*, 1:77–94, 06 1988.

[FHM18] Joseph F Fitzsimons, Michal Hajdušek, and Tomoyuki Morimae. Post hoc verification of quantum computation. *Physical review letters*, 120(4):040501, 2018.

[Gav12] Dmitry Gavinsky. Quantum money with classical verification. In *2012 IEEE 27th Conference on Computational Complexity*, pages 42–52. IEEE, 2012.

[GH17] William Timothy Gowers and Omid Hatami. Inverse and stability theorems for approximate representations of finite groups. *Sbornik: Mathematics*, 208(12):1784, 2017.

[GMR89] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on computing*, 18(1):186–208, 1989.

[GV19] Alexandru Gheorghiu and Thomas Vidick. Computationally-secure and composable remote state preparation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1024–1033. IEEE, 2019.

[HHJ$^+$17] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017.

[KR03] Julia Kempe and Oded Regev. 3-local Hamiltonian is QMA-complete. *Quantum Information and Computation*, 3(3):258–264, 2003.

[KSVV02] Alexei Yu Kitaev, Alexander Shen, Mikhail N Vyalyi, and Mikhail N Vyalyi. *Classical and quantum computation*. Number 47. American Mathematical Soc., 2002.

[Mah18a] Urmila Mahadev. Classical homomorphic encryption for quantum circuits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 332–338. IEEE, 2018.

[Mah18b] Urmila Mahadev. Classical verification of quantum computations. In *Foundations of Computer Science (FOCS), 2018 IEEE 59th Annual Symposium on*, pages 259–267, Oct 2018.

[Mah18c] Urmila Mahadev. Classical verification of quantum computations. *arXiv preprint arXiv:1804.01082*, 2018.

[MV20] Tony Metger and Thomas Vidick. Self-testing of a single quantum device under computational assumptions. *arXiv preprint arXiv:2001.09161*, 2020.

[MVW12] Abel Molina, Thomas Vidick, and John Watrous. Optimal counterfeiting attacks and generalizations for Wiesner's quantum money. In *Conference on Quantum Computation, Communication, and Cryptography*. Springer, 2012.

[NV16] Anand Natarajan and Thomas Vidick. Robust self-testing of many-qubit states. *arXiv e-prints*, page arXiv:1610.03574, Oct 2016.

[SJ00] Claus Schnorr and Markus Jakobsson. Security of signed ElGamal encryption. volume 1976, pages 73–89, 12 2000.

[Unr12] Dominique Unruh. Quantum proofs of knowledge. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 135–152. Springer, 2012.

[VZ19] Thomas Vidick and Tina Zhang. Classical zero-knowledge arguments for quantum computations. *arXiv preprint arXiv:1902.05217*, 2019.

[Wat09] John Watrous. Encyclopedia of complexity and system science, chapter quantum computational complexity, 2009.

[Wie83] Stephen Wiesner. Conjugate coding. *ACM Sigact News*, 15(1):78–88, 1983.

[Zha19] Mark Zhandry. Quantum lightning never strikes the same state twice. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 408–438. Springer, 2019.