Electronic Colloquium on Computational Complexity, Report No. 145 (2025)

Unentanglement and Post-Measurement Branching in Quantum Interactive Proofs

Sabee Grewal* William Kretschmer*

Abstract

We investigate two resources whose effects on quantum interactive proofs remain poorly understood: the promise of *unentanglement*, and the verifier's ability to condition on an intermediate measurement, which we call *post-measurement branching*. We first show that unentanglement can dramatically increase computational power: three-round unentangled quantum interactive proofs equal NEXP, even if only the first message is quantum. By contrast, we prove that if the verifier uses no post-measurement branching, then the same type of unentangled proof system has at most the power of QAM. Finally, we investigate post-measurement branching in two-round quantum-classical proof systems. Unlike the equivalence between public-coin and private-coin classical interactive proofs, we give evidence of a separation in the quantum setting that arises from post-measurement branching.

1 Introduction

Quantum proof systems are one framework for studying quantum mechanical effects on computational complexity. They allow us to ask how uniquely quantum resources—such as superposition, entanglement, and measurement—affect our ability to efficiently verify that certain statements are true. A broad goal in this area is understanding which resources increase computational power, which restrict it, and which turn out to be irrelevant.

In this work, we focus on two such resources: the promise of *unentanglement*, and the verifier's ability to perform *post-measurement branching*. At a high level, we show that each of these resources can fundamentally alter the computational power of interactive proof systems.

1.1 Unentanglement

Whether entanglement increases or decreases the computational power of quantum proof systems is subtle. A striking example is the separation between MIP* and MIP, the classes of problems decidable by multiprover interactive proof systems with and without entanglement between provers, respectively. Whereas MIP = NEXP [BFL91], it took considerable additional effort to show that the entangled-prover variant MIP* contains NEXP [IV12], because entangled provers could potentially cheat in ways that unentangled provers cannot. On the contrary, we now know that entanglement adds vastly more power to multiprover interactive protocols: MIP* was recently shown to equal RE [JNVWY22], and thus can solve undecidable problems.

Conversely, it is widely believed that in certain proof systems, *unentanglement* affords additional computational power. This is perhaps best captured by the QMA vs. QMA(2) question [KMY01], which asks whether two unentangled quantum proofs are more powerful than one. Despite considerable

*UT Austin. {sabee,kretsch}@cs.utexas.edu

ISSN 1433-8092

effort (e.g., [ABDFS09; CD10; BT12; GNN12; CF13; CS12; HM13]), nothing beyond the trivial containments $\mathsf{QMA} \subseteq \mathsf{QMA}(2) \subseteq \mathsf{NEXP}$ has been established since $\mathsf{QMA}(2)$'s introduction over two decades ago [KMY01]. Moreover, unlike for questions such as P vs. BPP or even BQP vs. PH, there is little consensus in the community about what the answer should be.

Underlying the QMA vs. $\mathsf{QMA}(2)$ problem is a more basic question: can we exhibit examples where the promise of unentanglement grants quantum proof systems exponentially more computational power? More recent work has attempted to study this question by defining variants of QMA and showing them equal to NEXP [JW23; BFM24; JW24; AGIMR24; BM25; BFLMW24]. For example, Jeronimo and Wu [JW23] introduced the class $\mathsf{QMA}^+(2)$, where the verifier receives two unentangled messages that are additionally promised to consist of real nonnegative amplitudes. Their main theorem $\mathsf{QMA}^+(2) = \mathsf{NEXP}$ showed that the combination of unentanglement and nonnegative amplitudes can give exponentially more power. However, this evidence for the power of unentanglement was called into question shortly thereafter, when Bassirian, Fefferman, and Marwaha [BFM24] showed that nonnegative amplitudes alone, without any promise of unentanglement, give the same power: $\mathsf{QMA}^+ = \mathsf{NEXP}$.

In fact, the only one of these works that showed how unentanglement specifically grants computational power is by Bassirian, Fefferman, Leigh, Marwaha, and Wu [BFLMW24]. There, they define "internally separable" variants $\mathsf{QMA}_{\mathsf{IS}}$ and $\mathsf{QMA}_{\mathsf{IS}}(2)$ of QMA and $\mathsf{QMA}(2)$, respectively, in which the quantum witnesses satisfy a certain multipartite entanglement condition. They then show that $\mathsf{QMA}_{\mathsf{IS}} \subseteq \mathsf{EXP}$ while $\mathsf{QMA}_{\mathsf{IS}}(2) = \mathsf{NEXP}$. (Note that their $\mathsf{QMA}_{\mathsf{IS}}(2) = \mathsf{NEXP}$ is highly sensitive to the completeness and soundness parameters, which cannot (apparently) be generically amplified.) On the whole, these works identify nontraditional sources of computational power (e.g., non-negative amplitudes and non-collapsing measurements), which have little to do with unentanglement.

In this work, we take a different approach to studying the power of unentanglement. Rather than considering new computational resources on top of unentanglement, we take a well-studied quantum proof system and investigate how it changes with an additional promise of unentanglement. We start with $\mathsf{QIP}[3]$, the class of problems decidable by a three-round quantum interactive proof system. Here the [3] denotes the number of rounds, unlike the (2) in $\mathsf{QMA}(2)$, which refers to the number of unentangled proofs; hence the careful difference in notation. It is a celebrated result that $\mathsf{QIP}[3] = \mathsf{QIP} = \mathsf{IP} = \mathsf{PSPACE}$ [JJUW11], showing that three-round quantum interactive proofs characterize polynomial space. Because quantum and classical interactive proof systems are so well understood, $\mathsf{QIP}[3]$ serves as a natural baseline for exploring the effects of unentanglement.

We introduce an unentangled variant of $\mathsf{QIP}[3]$, denoted $\mathsf{QIP}_{\mathsf{unent}}[3]$, that differs from $\mathsf{QIP}[3]$ in exactly one respect: the prover is restricted to apply channels that generate no entanglement between their private workspace qubits and any messages sent to the verifier. We formalize $\mathsf{QIP}_{\mathsf{unent}}[3]$ using so-called *entanglement-breaking* channels [HSR03], which are precisely the set of channels whose action on one half of any bipartite state yields a separable state across the bipartition. These channels are also known as *measure-and-prepare* channels, because they can be equivalently described as first measuring the input state and then preparing a new quantum state conditioned on the classical measurement outcome. Our definition of $\mathsf{QIP}_{\mathsf{unent}}[3]$ trivially captures $\mathsf{QMA}(2)$ as a special case: the prover sends the first witness in the first round, the verifier does nothing in the second round, and the prover sends the second witness (unentangled with the first) in the third and final round.

Our first result establishes that the promise of unentanglement in $\mathsf{QIP}_{\mathsf{unent}}[3]$ yields dramatically greater power than $\mathsf{QIP}[3] = \mathsf{PSPACE}$:

Theorem 1.1. $\mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$.

Note that, unlike prior comparable work [JW23; BFM24; BFLMW24], our result is insensitive to the precise completeness-soundness gap. In particular, Theorem 1.1 holds for any constant gap less

than 1.

For the upper bound, we argue that an NEXP machine can nondeterministically guess an exponentially-large description of the prover's strategy and then verify whether it causes the verifier to accept with high probability or not. To prove the complementary lower bound, we make use of a certain quantum PCP for NEXP introduced by Raz some 20 years ago [Raz05]. Raz showed that any NEXP language admits a polynomial-time quantum verifier that receives two inputs: a polynomial-length quantum witness, and an exponentially-large classical proof (readable by query access). The verifier measures the quantum witness and, based on the measurement outcome, queries a *single* polynomial-size block of the proof. We argue that this PCP can be simulated by a $\mathsf{QIP}_{\mathsf{unent}}[3]$ protocol: the prover sends the quantum witness in the first round, the verifier measures it and sends the resulting classical query to the prover in the second round, and the prover responds with the answer to the classical query in the final round.

One can view Theorem 1.1 as popularizing and modernizing Raz's quantum PCP result in the context of QMA(2) and the power of unentanglement, as the proof is not particularly difficult with Raz's result in hand. Indeed, Raz's discussion briefly mentions something close to Theorem 1.1: that an interactive prover with quantum power in the first round and classical power thereafter can convince a verifier of the solution to an NEXP problem [Raz05, Section 1.5]. Nevertheless, we found that Theorem 1.1 surprised every expert in the field with whom we consulted.

It is striking that the containment of NEXP in $\mathsf{QIP}_{\mathsf{unent}}[3]$ makes only partial use of the proof system's quantum capabilities, as the second and third messages are purely classical. Naturally, one might wonder why we need the promise of unentanglement at all: if the verifier knows that the final message is classical, then doesn't that already guarantee zero entanglement between the prover's first and last messages? The key observation is that the source of power is not unentanglement between the messages, but rather from the unentanglement between the *prover's workspace qubits* and the messages.

The following simple example illuminates the situation: consider a QIP[3] protocol in which the verifier challenges the prover to a version of the CHSH game [CHTW04], with the verifier playing both the role of the referee and one of the players. In the first round, the prover sends the verifier a single qubit. The verifier then uniformly samples $x \sim \{0,1\}$ and sends it to the prover, who responds with a single classical bit a. Next, the verifier uniformly samples $y \sim \{0,1\}$. If $y = 0$, the verifier measures the qubit sent by the prover in the $\{|0\rangle, |1\rangle\}$ basis; else they measure in the $\{|+\rangle, |-\rangle\}$ basis. Calling the measurement outcome b, the verifier accepts if and only if $xy = a \oplus b$. An entangled prover can succeed with probability $\cos^2(\pi/8) \approx 0.85$ by sending one half of a Bell pair in the first round, and, in the final round, measuring the other half according to the optimal CHSH strategy. By contrast, an *unentangled* prover can make the verifier accept with probability at most 0.75. Hence, even with only a single quantum message in the first round, the promise of unentanglement places a nontrivial restriction on the set of valid QIP[3] prover strategies.

1.2 Post-Measurement Branching and Unentanglement

One key difference between $\mathsf{QIP}_{\mathsf{unent}}[3]$ and QMA(2) is adaptivity: a $\mathsf{QIP}_{\mathsf{unent}}[3]$ verifier can condition their round-2 challenge to the prover on the result of a measurement, possibly applied to the round-1 message. By contrast, a QMA(2) verifier receives a pair of unentangled witnesses simultaneously, without the ability for either witness to depend on a chosen challenge. In the interest of isolating the source of $\mathsf{QIP}_{\mathsf{unent}}[3]$'s exponential power, it is natural to ask whether the quantum-classical-classical unentangled proof system for NEXP necessitated this sort of adaptivity. For example, if instead the verifier simply sent the prover random coin tosses in round 2, could they still verify solutions to NEXP problems?

In the context of quantum proof systems, we refer to this type of adaptivity as *post-measurement branching*, which means the ability to condition on a partial measurement of a state while retaining the residual post-measurement state. Our proposal to study the same quantum proof system with and without post-measurement branching mirrors the difference between $\mathsf{AM}[k]$ and $\mathsf{IP}[k]$: in $\mathsf{AM}[k]$ the verifier's messages to the prover consist of public coin tosses, whereas in $\mathsf{IP}[k]$ the messages can be arbitrary polynomial-time randomized computations on the prior transcript of the protocol. Classically, we know that adaptivity cannot help much: $\mathsf{IP}[k] \subseteq \mathsf{AM}[k+2]$ [GS86], and for constant k, $\mathsf{AM}[k] \subseteq \mathsf{AM}[2]$ [Bab85; BM88]. But should we expect the analogous equivalence to hold for quantum protocols, unentangled or otherwise?

Concretely, consider the subclass of $\mathsf{QIP}_{\mathsf{unent}}[3]$ in which the round-2 message consists of public coin tosses and the round-3 message is classical. We call this subclass $\mathsf{QMACM}_{\mathsf{unent}}$ because it behaves like QMAM [MW05], except that the prover's private and message registers are always unentangled and the last message is classical. Then does $\mathsf{QMACM}_{\mathsf{unent}} = \mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$? Our second result gives strong evidence that the answer is no:

Theorem 1.2. $\mathsf{QMACM}_{\mathsf{unent}} \subseteq \mathsf{QAM}$.

Here, QAM is the set of problems verifiable by an interaction in which the verifier (Arthur) sends public coin tosses and the prover (Merlin) responds with a quantum message [MW05]. In contrast to $\mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$, where restricting to unentangled provers significantly *increased* computational power compared to $\mathsf{QIP}[3] = \mathsf{PSPACE}$, here the unentangled variant $\mathsf{QMACM}_{\mathsf{unent}}$ is quite plausibly *weaker* than its entangled variant QMACM. Indeed, whereas $\mathsf{QMACM}_{\mathsf{unent}} \subseteq \mathsf{QAM} = \mathsf{BP} \cdot \mathsf{QMA} \subseteq \mathsf{BPP}^{\mathsf{PP}}$ (where $\mathsf{BP} \cdot \mathsf{QMA}$ denotes problems that have a randomized many-one reduction to QMA), we know of no better upper bound on the corresponding entangled proof system than $\mathsf{QMACM} \subseteq \mathsf{QMAM} = \mathsf{QIP}[3] = \mathsf{PSPACE}$ [JJUW11]. $\mathsf{QMACM}_{\mathsf{unent}}$ is possibly equal to QMA, and in fact the classes coincide with polynomial-size advice: $\mathsf{QMA/poly} = \mathsf{QAM/poly} = \mathsf{QMACM}_{\mathsf{unent}}\mathsf{/poly}$, because $\mathsf{QAM} = \mathsf{BP} \cdot \mathsf{QMA}$ and the $\mathsf{BP}\cdot$ operator can be derandomized with advice (cf. [Aar06; AH23]). Thus, Theorem 1.2 illustrates both the necessity of post-measurement branching for making certain proof systems equal to NEXP, and the surprising fact that unentanglement may hinder the power of an interactive proof system.

The proof of Theorem 1.2 involves simulating the $\mathsf{QMACM}_{\mathsf{unent}}$ proof system in two rounds by combining Merlin's first and last messages into one. In the QAM protocol, Arthur first sends Merlin polynomially many independent challenges that he could have sent in round 2 of the $\mathsf{QMACM}_{\mathsf{unent}}$ protocol. Then Arthur asks for both Merlin's quantum proof that he would have sent in round 1, and Merlin's classical answers that he would have given in round 3 in response to each of the challenges. We argue that the soundness of the protocol is approximately preserved if Arthur runs his original $\mathsf{QMACM}_{\mathsf{unent}}$ check on a random one of the challenges. This strategy is somewhat analogous to the $\mathsf{AM}[k] \subseteq \mathsf{AM}[2]$ collapse theorem [BM88], but requires heavier tools to handle the quantum part of the message. For example, a crucial ingredient in our proof comes from one-way communication complexity: any n-qubit quantum state ρ can be "compressed" into a $\mathrm{poly}(n)$-bit message, from which the expectation of $\exp(n)$ different measurements on ρ may be later estimated [Aar05a]. We do not directly use this compression scheme in the QAM protocol, but it indirectly allows us to apply a union bound over the set of n-qubit states as if there were only $2^{\mathrm{poly}(n)}$ of them, instead of $2^{\exp(n)}$.

1.3 Post-Measurement Branching with Classical Messages

Finally, we turn to the simplest setting in which the effect of post-measurement branching can be studied: two-round interactive proofs with classical messages and a quantum verifier. Specifically, we study the complexity classes QCAM and $\mathsf{QCIP}[2]$. In QCAM, the verifier's sole message consists

of random coin tosses. In $\mathsf{QCIP}[2]$, by contrast, the verifier may send an arbitrary classical message generated through a partial measurement of a quantum state; both the classical outcome and the corresponding post-measurement state can then be used later in the verification procedure.

Recall that AM and $\mathsf{IP}[2]$ coincide, and more generally, constant-round public-coin (AM) and private-coin (IP) protocols have the same computational power [Bab85; GS86; BM88]. In contrast, we show that the quantum setting exhibits an apparent separation: $\mathsf{QCIP}[2]$ potentially has greater power than QCAM. Our findings are summarized in the following theorem.

Theorem 1.3. $\mathsf{QCAM} = \mathsf{BP} \cdot \mathsf{QCMA} \subseteq \mathsf{BQ} \cdot \mathsf{QCMA} \subseteq \mathsf{QCIP}[2] \subseteq \mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}}}$.

$\mathsf{BP} \cdot \mathsf{QCMA}$ (resp. $\mathsf{BQ} \cdot \mathsf{QCMA}$) is the class of promise problems that admit a randomized (resp. quantum) many-one reduction to a promise problem in QCMA. We note that $\mathsf{QCAM} = \mathsf{BP} \cdot \mathsf{QCMA}$ was originally proven by Marriott [Mar03], but we include a complete proof in Section 4 in more standard notation.

The containments involving $\mathsf{BP} \cdot \mathsf{QCMA}$ and $\mathsf{BQ} \cdot \mathsf{QCMA}$ are reasonably straightforward applications of the definitions. Placing an upper bound on $\mathsf{QCIP}[2]$, however, takes more effort. Intuitively, it works as follows: first, use the base BQP machine to generate the verifier's round-1 message. Then, we will use the $\mathsf{NP}^{\mathsf{PP}}$ machine to simulate the prover. The idea is to nondeterministically guess the prover's round-2 message and then verify using the PP oracle whether that message would be accepted by the verifier. PP suffices for this step because of Aaronson's $\mathsf{PP} = \mathsf{PostBQP}$ theorem [Aar05b], which shows that PP equals the set of problems decidable by an efficient quantum machine with postselection. Using postselection, one can condition on producing the same message that the verifier sampled in round 1, resulting in the same residual state that the verifier uses to decide acceptance or rejection at the end.

Unlike the classical equivalence $\mathsf{IP}[2] = \mathsf{AM}$, Theorem 1.3 hints that $\mathsf{QCIP}[2]$ is more powerful than QCAM, because the set of problems quantumly reducible to QCMA is plausibly larger than the set classically reducible to QCMA. However, this distinction alone does not make full use of $\mathsf{QCIP}[2]$'s extra power, as $\mathsf{BQ} \cdot \mathsf{QCMA}$ is a class that uses no post-measurement branching. Looking at the higher end of the containments, we found it considerably more challenging to place an upper bound on $\mathsf{QCIP}[2]$ than QCAM *precisely because* the former uses post-measurement branching, and thus simulating Merlin requires a handle on the post-measurement state.

1.4 Open Problems

We conclude with some directions for future work.

1. We proved that $\mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$. Are there quantum proof systems between $\mathsf{QMA}(2)$ and $\mathsf{QIP}_{\mathsf{unent}}[3]$ that capture NEXP? An unentangled version of QMAM is a natural candidate to study.

2. A key difference between QCAM and $\mathsf{QCIP}[2]$ (and even between $\mathsf{BQ} \cdot \mathsf{QCMA}$ and $\mathsf{QCIP}[2]$) is the power of post-measurement branching. What more can be said about this power? For instance, can one show stronger containments than $\mathsf{BQ} \cdot \mathsf{QCMA} \subseteq \mathsf{QCIP}[2]$?

3. It is known that $\mathsf{IP}[k] = \mathsf{IP}[2] = \mathsf{AM}$ for every constant $k \geq 2$ [Bab85; GS86; BM88]. Is an analogous collapse true for $\mathsf{QCIP}[k]$?

4. Are there oracles relative to which any of the containments in Theorem 1.3 are strict?

2 Unentangled Quantum Interactive Proofs

In this section, we introduce an unentangled three-round quantum interactive proof system, denoted $\mathsf{QIP}_{\mathsf{unent}}[3]$. Our definition mirrors $\mathsf{QIP}[3]$, except that Merlin's actions are restricted to entanglement-breaking channels. After defining our model, we prove that $\mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$. The result adapts a certain type of quantum PCP for NEXP introduced by Raz [Raz05] (quoted below in Lemma 2.7).

We begin by recalling the definition of $\mathsf{QIP}[3]$. A $\mathsf{QIP}[3]$ verification procedure is specified by a polynomial-time uniformly generated family of quantum circuits $V = \{V_1^x, V_2^x : x \in \{0,1\}^*\}$. On an input x of length n, these circuits determine the actions of the verifier across the three-message interaction. Each circuit acts on $\mathrm{poly}(n)$-sized registers partitioned into *message* qubits $\mathcal{M}$, exchanged with the prover, and *verifier workspace* qubits $\mathcal{V}$, retained by the verifier throughout.

The prover is an unrestricted family $P = \{P_1^x, P_2^x : x \in \{0,1\}^*\}$ of arbitrary quantum operations that likewise act on the same message qubits $\mathcal{M}$ and *prover workspace* qubits $\mathcal{P}$ (which need not be polynomially-bounded in size). The interaction proceeds as follows:

1. The three registers $\mathcal{V}$, $\mathcal{M}$, $\mathcal{P}$ are each initialized to the all-zeros state.
2. The prover applies P_1^x to $\mathcal{P}$ and $\mathcal{M}$.
3. The verifier applies V_1^x to $\mathcal{V}$ and $\mathcal{M}$.
4. The prover applies P_2^x to $\mathcal{P}$ and $\mathcal{M}$.
5. Finally, the verifier applies V_2^x to $\mathcal{V}$ and $\mathcal{M}$ and measures a designated output qubit to decide acceptance or rejection.

Sometimes the prover and verifier are called "Merlin" and "Arthur" respectively. We will typically only use these names in interactive protocols where the verifier's messages consist of public coin tosses, consistent with the distinction between the complexity classes $\mathsf{AM}[k]$ and $\mathsf{IP}[k]$.

We now formally define the class $\mathsf{QIP}[3]$.

Definition 2.1 ($\mathsf{QIP}[3]$)**.** A promise problem $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ is in $\mathsf{QIP}[3, c, s]$ for polynomial-time computable functions $c, s : \mathbb{N} \to [0,1]$ if there exists a $\mathsf{QIP}[3]$ verification procedure V with the following properties:

- *Completeness.* For all $x \in A_{\mathrm{yes}}$, there exists a quantum prover P that causes V to accept x with probability at least $c(|x|)$.
- *Soundness.* For all $x \in A_{\mathrm{no}}$, every quantum prover P causes V to accept x with probability at most $s(|x|)$.

We define $\mathsf{QIP}[3] \coloneqq \mathsf{QIP}[3, 2/3, 1/3]$.

It is known that any polynomial-round quantum interaction can be parallelized to three rounds [KW00] and that $\mathsf{QIP}[3, 2/3, 1/3] = \mathsf{QIP}[3, 1, 2^{-\mathrm{poly}}] = \mathsf{PSPACE}$ [JJUW11].

Remark 2.2 (Entanglement between registers)**.** A key feature of $\mathsf{QIP}[3]$—one that is central to this work—is that the private workspaces of both the prover and verifier may be entangled with the message registers exchanged during the interaction.

We turn to defining our unentangled variant of $\mathsf{QIP}[3]$, which adds an additional restriction on the prover involving *entanglement-breaking* channels [HSR03]. These channels are so-called because they are precisely the set of channels Φ with the property that for any input density matrix ρ,

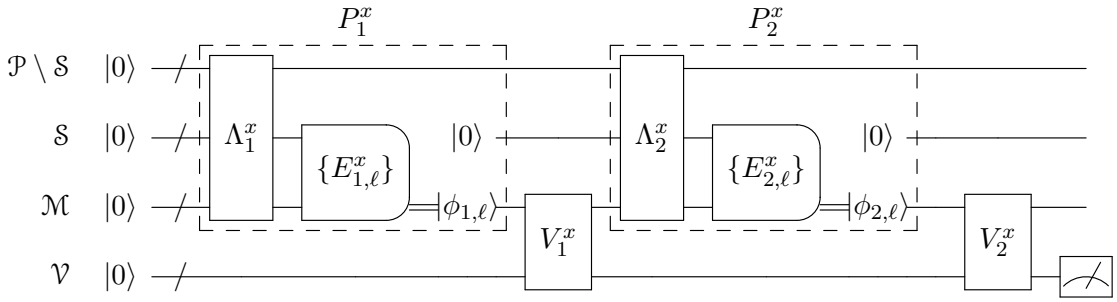

Figure 1: The general form of a $\mathsf{QIP}_{\mathsf{unent}}[3]$ interaction between an unentangled prover $P = \{P_1^x, P_2^x\}$ and verifier $V = \{V_1^x, V_2^x\}$.

$(I \otimes \Phi)(\rho)$ is separable (across the cut between the output of Φ and the tensored identity factor). An equivalent characterization of an entanglement-breaking channel Φ is one that takes the form

$$\Phi(\rho) \;=\; \sum_{\ell} \operatorname{tr}(E_\ell \rho)\; |\phi_\ell\rangle\langle\phi_\ell| \,.$$

for some POVM $\{E_\ell\}$ and set of states $|\phi_\ell\rangle$. Operationally, this means that an entanglement-breaking channel applies a measurement to the input state and prepares a new state conditioned on the classical outcome of the measurement. For this reason, entanglement-breaking channels are sometimes called *measure-and-prepare* channels.

An *unentangled* $\mathsf{QIP}[3]$ prover is a family $P = \{P_1^x, P_2^x : x \in \{0,1\}^*\}$ of quantum operations that likewise act on $\mathcal{P}$ and $\mathcal{M}$, subject to the restriction that each P_i^x is the sequential composition of:

1. Applying some arbitrary channel Λ_i^x to $\mathcal{P}$ and $\mathcal{M}$;
2. Applying an entanglement-breaking channel Φ_i^x whose input qubits are $\mathcal{S} \cup \mathcal{M}$ and output qubits are $\mathcal{M}$, for some $\mathcal{S} \subseteq \mathcal{P}$;
3. Reinitializing the qubits in $\mathcal{S}$ to $|0\rangle$. (This step is only needed to preserve the size of $\mathcal{P}$.)

Use of the entanglement-breaking channel Φ_i^x ensures that after completion of the prover's operation P_i^x, the message register $\mathcal{M}$ is unentangled from the prover's workspace qubits $\mathcal{P}$.

Figure 1 depicts the interaction between an unentangled prover and verifier. This type of interaction underlies $\mathsf{QIP}_{\mathsf{unent}}[3]$, whose formal definition replaces the prover in $\mathsf{QIP}[3]$ with an unentangled prover.

Definition 2.3 ($\mathsf{QIP}_{\mathsf{unent}}[3]$)**.** A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in $\mathsf{QIP}_{\mathsf{unent}}[3, c, s]$ for polynomial-time computable $c, s : \mathbb{N} \to [0,1]$ if there exists a $\mathsf{QIP}[3]$ verification procedure V with the following properties:

- *Completeness.* For all $x \in A_{\text{yes}}$, there exists an *unentangled* prover P that makes V accept with probability at least $c(|x|)$.
- *Soundness.* For all $x \in A_{\text{no}}$, every *unentangled* prover P makes V accept with probability at most $s(|x|)$.

We define $\mathsf{QIP}_{\mathsf{unent}}[3] \coloneqq \mathsf{QIP}_{\mathsf{unent}}[3, 2/3, 1/3]$.

When defining such complexity classes, one should always ask whether the the choices of completeness 2/3 and soundness 1/3 are arbitrary—in particular, can they be amplified, say by parallel repetition? It will follow from our results that the completeness and soundness parameters can be amplified to 1 and $o(1)$, respectively (Corollary 2.10).

Notice that $\mathsf{QIP}_{\mathsf{unent}}[3]$ places no restrictions on the verifier's ability to send entangled messages to the prover; the promise merely guarantees that the prover never maintains entanglement with their sent messages.

2.1 Upper Bounding $\mathsf{QIP}_{\mathsf{unent}}[3]$

To place an upper bound on $\mathsf{QIP}_{\mathsf{unent}}[3]$, we first make note of some useful properties of entanglement-breaking channels that allow us to reduce the complexity of the prover. This first lemma shows that one can straightforwardly upper bound the description complexity of an entanglement-breaking channel, which is *a priori* infinite.

Lemma 2.4. *Suppose Φ is an entanglement-breaking channel from m qubits to n qubits. Then Φ admits a decomposition in terms of a POVM $\{E_\ell\}$ and a set of pure states $\{|\phi_\ell\rangle\}$ with at most 4^{m+n} terms:*

$$\Phi(\rho) \;=\; \sum_{\ell=1}^{4^{m+n}} \operatorname{tr}(E_\ell \rho)\; |\phi_\ell\rangle\langle\phi_\ell| \,.$$

Proof. The proof follows [HSR03, Theorem 4] exactly, with the sole addition of some extra accounting. [HSR03, Theorem 4] shows that if Φ is entanglement-breaking, then its Choi state

$$(I \otimes \Phi)(|\beta\rangle\langle\beta|)$$

is separable (i.e., a mixture of product states), where $|\beta\rangle = \frac{1}{\sqrt{2^m}} \sum_{j \in \{0,1\}^m} |j\rangle\,|j\rangle$. By a result of Horodecki [Hor97], every separable state on $m+n$ qubits is a convex combination of at most 4^{m+n} pure product states, and therefore the Choi state admits a decomposition:

$$(I \otimes \Phi)(|\beta\rangle\langle\beta|) = \sum_{\ell=1}^{4^{m+n}} p_\ell\, |v_\ell\rangle\langle v_\ell| \otimes |w_\ell\rangle\langle w_\ell| \,,$$

where $\{p_\ell\}$ are probabilities summing to 1 and $\{|w_\ell\rangle\}$, $\{|v_\ell\rangle\}$ are lists of normalized pure states.

Now let Ω be the map

$$\Omega(\rho) \coloneqq \sum_{\ell=1}^{4^{m+n}} \operatorname{tr}(d p_\ell\, |v_\ell\rangle\langle v_\ell|\, \rho)\, |w_\ell\rangle\langle w_\ell| \,,$$

which has the form required by the lemma. Using $|v_\ell\rangle = \sum_j |j\rangle\,\langle j|v_\ell\rangle$, one easily verifies that

$$\begin{aligned}
(I \otimes \Omega)(|\beta\rangle\langle\beta|) &= \sum_{jk\ell} |j\rangle\langle k| \otimes |w_\ell\rangle\langle w_\ell|\, p_\ell \langle j|v_\ell\rangle\langle v_\ell|k\rangle \\
&= \sum_{\ell} p_\ell\, |v_\ell\rangle\langle v_\ell| \otimes |w_\ell\rangle\langle w_\ell| \\
&= (I \otimes \Phi)(|\beta\rangle\langle\beta|).
\end{aligned}$$

Two channels are equal if and only if their Choi states are the same, so $\Phi = \Omega$.

To complete the proof, one must verify that $\{dp_\ell \, |v_\ell\rangle\langle v_\ell|\}$ is a POVM. This follows by taking a partial trace of the Choi state:

$$\begin{aligned}\mathrm{tr}_2[(I \otimes \Phi)(|\beta\rangle\langle\beta|)] &= \frac{I}{d} \\ &= \sum_{\ell=1}^{4^{m+n}} p_\ell \, |v_\ell\rangle\langle v_\ell| \, .\end{aligned}$$

□

Next, we argue that the prover can further simplify their strategy by eliminating the use of private qubits:

Lemma 2.5. *Suppose there exists a* $\mathsf{QIP}_{\mathsf{unent}}[3]$ *prover* P *that makes* V *accept with probability at least* $p(|x|)$*. Then there exists a prover* $\overline{P}$ *that also makes* V *accept with probability at least* $p(|x|)$*, but for which*

1. *The output of* $\overline{P}_1^x$ *is a pure state on* $\mathcal{M}$*,*
2. $\overline{P}$ *uses no prover workspace qubits, and*
3. $\overline{P}_1^x$ *and* $\overline{P}_2^x$ *are themselves entanglement-breaking channels.*

Proof. After the prover applies P_1^x to $|0\rangle_{\mathcal{PM}}$ the state of registers $\mathcal{P}$ and $\mathcal{M}$ has the form

$$\sum_\ell p_\ell \, \sigma_{\ell,\mathcal{P}} \otimes |\psi_\ell\rangle\langle\psi_\ell|_{\mathcal{M}} \, ,$$

for some probabilities $\{p_\ell\}$, mixed states $\{\sigma_\ell\}$, and pure states $\{|\psi_\ell\rangle\}$ parameterized by the possible outcomes ℓ of the POVM underlying the entanglement-breaking channel Φ_1^x. Consider postselecting on a particular outcome ℓ. By convexity, there must exist a choice $\ell = \ell^*$ such that replacing P_1^x with direct preparation of $\sigma_{\ell^*,\mathcal{P}} \otimes |\psi_{\ell^*}\rangle\langle\psi_{\ell^*}|_{\mathcal{M}}$ causes the interaction between prover and verifier to accept with probability at least $p(|x|)$. Let P' be the prover derived from P by making this replacement, which causes it to satisfy Item 1. To avoid confusion with notation later in the proof, we drop the ℓ^* subscript and call the initial state simply $\sigma_\mathcal{P} \otimes |\psi\rangle\langle\psi|_{\mathcal{M}}$.

Next, we observe that one can eliminate the need to prepare $\sigma_\mathcal{P}$ on a second register, and thus remove the private workspace qubits. Let $\{E_\ell\}$ be the POVM and $\{|\phi_\ell\rangle\}$ be the set of states underlying Φ_2^x, for which

$$\Phi_2^x(\rho_{\mathcal{PM}}) \;=\; \sum_\ell \mathrm{tr}(E_\ell \rho_{\mathcal{PM}}) \; |\phi_\ell\rangle\langle\phi_\ell|_{\mathcal{M}} \, .$$

Now define $\overline{P}$ by

$$\overline{P}_1^x(\rho_\mathcal{M}) \;\coloneqq\; \mathrm{tr}(\rho_\mathcal{M}) \, |\psi\rangle\langle\psi|_{\mathcal{M}}$$

and

$$\overline{P}_2^x(\rho_\mathcal{M}) \;\coloneqq\; \sum_\ell \mathrm{tr}(E_\ell \Lambda_2^x(\sigma_\mathcal{P} \otimes \rho_\mathcal{M})) \, |\phi_\ell\rangle\langle\phi_\ell| \, .$$

Then clearly $\overline{P}$ satisfies Item 2, since $\overline{P}_1^x$ and $\overline{P}_2^x$ both map $\mathcal{M}$ to $\mathcal{M}$. Additionally, the interaction between $\overline{P}$ and V has the same acceptance probability as that between P' and V, because we essentially deferred initializing $\sigma_\mathcal{P}$ until $\overline{P}_2^x$, and then traced out $\mathcal{P}$ afterwards.

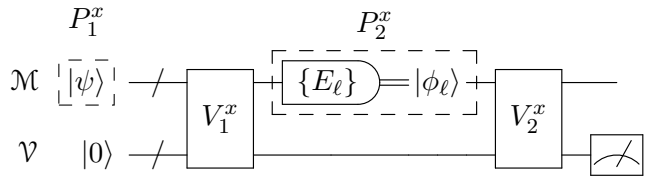

Figure 2: The simplified form of $\mathsf{QIP}_{\mathsf{unent}}[3]$ interaction between an unentangled prover $P = \{P_1^x, P_2^x\}$, in canonical form that derives from Lemma 2.5, and verifier $V = \{V_1^x, V_2^x\}$.

The last condition we must verify is that $\overline{P}_1^x$ and $\overline{P}_2^x$ are entanglement-breaking channels. This is immediate for $\overline{P}_1^x$. To see that $\overline{P}_2^x$ is entanglement-breaking, first let $\Psi_2^x(\rho_{\mathcal{M}}) \coloneqq \Lambda_2^x(\sigma_{\mathcal{P}} \otimes \rho_{\mathcal{M}})$. Then

$$\begin{aligned}\overline{P}_2^x(\rho_{\mathcal{M}}) &= \sum_\ell \operatorname{tr}(E_\ell \Psi_2^x(\rho_{\mathcal{M}}))\, |\phi_\ell\rangle\langle\phi_\ell| \\ &= \sum_\ell \operatorname{tr}(\Psi_2^{x*}(E_\ell)\rho_{\mathcal{M}})\, |\phi_\ell\rangle\langle\phi_\ell|\,,\end{aligned}$$

where Ψ_2^{x*} is the channel adjoint of Ψ_2^x. This satisfies the definition of entanglement breaking because $\Psi_2^{x*}(E_\ell)$ is a POVM, since the adjoint of a CPTP map is completely positive and unital. □

Figure 2 shows the canonical form of an unentangled prover that derives from Lemma 2.5 This simplification lets us simulate $\mathsf{QIP}_{\mathsf{unent}}[3]$ in nondeterministic exponential time.

Theorem 2.6. *For any $c(n) - s(n) \geq \frac{1}{2^{\mathrm{poly}(n)}}$, $\mathsf{QIP}_{\mathsf{unent}}[3, c, s] \subseteq \mathsf{NEXP}$.*

Proof. Given $A \in \mathsf{QIP}_{\mathsf{unent}}[3]$, let V be a corresponding $\mathsf{QIP}_{\mathsf{unent}}[3]$ verification procedure with completeness c and soundness s. On input $x \in \{0,1\}^n$, to decide whether $x \in A_{\mathrm{yes}}$ or $x \in A_{\mathrm{no}}$, the NEXP procedure is to nondeterministically guess classical descriptions of the entanglement-breaking channels P_1^n, P_2^n that act on $\mathcal{M}$ (up to some small error in diamond norm, say $\frac{c(|x|)-s(|x|)}{100}$, which requires $\mathrm{poly}(n)$ bits of precision), and then verify in exponential time whether the interaction between P_1^n, P_2^n and V_1^n, V_2^n causes the verifier to accept with probability at least $\frac{c(|x|)+s(|x|)}{2}$. Note crucially by Lemma 2.4 that the descriptions of P_1^n and P_2^n require at most $2^{\mathrm{poly}(n)}$ bits, as the NEXP machine can guess the lists of POVM elements and pure states that describe the channels. For any optimal prover, restricting attention to entanglement-breaking P_1^n, P_2^n with no prover workspace qubits is without loss of generality, by Lemma 2.5. As long as the total error incurred by the finite-precision approximation is less than $c(|x|) - s(|x|)$, the completeness/soundness guarantee of the $\mathsf{QIP}_{\mathsf{unent}}[3]$ verifier ensures that "yes" instances have an accepting witness to the NEXP verifier, while "no" instances do not. □

One could analogously define $\mathsf{QIP}_{\mathsf{unent}}[k]$ for any polynomially-bounded k, and hope to show the same containment in NEXP. Unfortunately, it appears that Lemma 2.5 does not directly generalize beyond three rounds in showing that workspace qubits are superfluous. For example, one could imagine a scenario in which the prover sends the verifier one half of an EPR pair in round 2, and then in rounds $k-1$ and k plays a sort of CHSH game with the prover. This would require the prover to retain their half of the EPR pair until round k.

2.2 Lower bounding $\mathsf{QIP}_{\mathsf{unent}}[3]$

In this section, we prove a complementary lower bound on $\mathsf{QIP}_{\mathsf{unent}}[3]$, implying that $\mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$. The proof relies on the following quantum PCP for NEXP, due to Raz [Raz05].

Lemma 2.7. *For any language $L \in \mathsf{NEXP}$, there exists a polynomial-time quantum oracle algorithm $Q^{(\cdot)}$ that makes a single* classical[1] *query to the oracle such that*

- Completeness. *For every $x \in L$, there exists a state $|\psi\rangle$ on $\mathrm{poly}(|x|)$ qubits and an oracle $f : \{0,1\}^{\mathrm{poly}(|x|)} \to \{0,1\}^{\mathrm{poly}(|x|)}$ for which $Q^f(x, |\psi\rangle)$ accepts with probability 1.*
- Soundness. *For every $x \notin L$, for every state $|\psi\rangle$ on $\mathrm{poly}(|x|)$ qubits and oracle $f : \{0,1\}^{\mathrm{poly}(|x|)} \to \{0,1\}^{\mathrm{poly}(|x|)}$, $Q^f(x, |\psi\rangle)$ accepts with probability $o(1)$ (as a function of $|x|$).*

The $\mathsf{QIP}_{\mathsf{unent}}[3]$ containment of NEXP amounts to a direct simulation of this PCP.

Theorem 2.8. $\mathsf{NEXP} \subseteq \mathsf{QIP}_{\mathsf{unent}}[3, 1, o(1)]$.

Proof. Consider a $\mathsf{QIP}_{\mathsf{unent}}[3]$ verifier in which V_1^x initially treats the register $\mathcal{M}$ as the input $|\psi\rangle$ to the algorithm $Q^{(\cdot)}$ from Lemma 2.7, then performs the intermediate measurement of $Q^{(\cdot)}(x, |\psi\rangle)$ to obtain the input $y \in \{0,1\}^{\mathrm{poly}(n)}$ to the classical query, and finally sends y over $\mathcal{M}$ to the prover. In the final round, V_2^x views $\mathcal{M}$ as the classical response to the oracle query $f(y)$, measures $\mathcal{M}$ in the computational basis, and then performs the final measurement of Q to decide acceptance or rejection.

This protocol has completeness 1, as witnessed by the unentangled prover P for which P_1^x prepares the state $|\psi\rangle$ and P_2^x evaluates the function f that causes $Q^f(x, |\psi\rangle)$ to accept with probability 1. For soundness, after putting the prover in the canonical form of Lemma 2.5, notice that we can further simplify the description of the prover by assuming without loss of generality that P_2^x evaluates some classical function, because the verifier measures $\mathcal{M}$ in the computational basis both at the end of P_1^x and at the start of P_2^x. Thus, such a prover strategy is fully specified by the state $|\psi\rangle$ sent at the start and the classical function $f : \{0,1\}^{\mathrm{poly}(n)} \to \{0,1\}^{\mathrm{poly}(n)}$ applied at the end. It follows that the $\mathsf{QIP}_{\mathsf{unent}}[3]$ soundness matches that of the underlying quantum PCP for NEXP, which Lemma 2.5 shows to be $o(1)$. □

Combining Theorems 2.6 and 2.8 gives:

Corollary 2.9. $\mathsf{QIP}_{\mathsf{unent}}[3] = \mathsf{NEXP}$

Together, Theorems 2.6 and 2.8 also show that the completeness/soundness gap of any $\mathsf{QIP}_{\mathsf{unent}}[3]$ protocol can be amplified from inverse-exponential to arbitrarily big:

Corollary 2.10. *For any $c(n) - s(n) \geq \frac{1}{2^{\mathrm{poly}(n)}}$, $\mathsf{QIP}_{\mathsf{unent}}[3, c, s] \subseteq \mathsf{QIP}_{\mathsf{unent}}[3, 1, o(1)]$.*

3 Post-Measurement Branching in Unentangled Proof Systems

The preceding section established that $\mathsf{NEXP} \subseteq \mathsf{QIP}_{\mathsf{unent}}[3]$, in sharp contrast to $\mathsf{QIP}[3] = \mathsf{PSPACE}$. This result highlights that the restriction to unentangled proofs can yield proof systems of significantly greater computational power. On the other hand, it is surprising that the containment of NEXP in $\mathsf{QIP}_{\mathsf{unent}}[3]$ made only partial use of the proof system's quantum capabilities, as the second and

[1] Meaning, the verifier measures a register to obtain $y \in \{0,1\}^{\mathrm{poly}(n)}$, and then queries the value of $f(y)$.

third messages were purely classical. To better understand the power of unentangled proof systems, in this section we consider placing further restrictions related to the adaptivity of the verifier.

Concretely, we define a subclass $\mathsf{QMACM}_{\mathsf{unent}}$ of $\mathsf{QIP}_{\mathsf{unent}}[3]$ in which the round-2 message consists of public coin tosses and the round-3 message is classical. A key distinction between $\mathsf{QIP}_{\mathsf{unent}}[3]$ and $\mathsf{QMACM}_{\mathsf{unent}}$ is that after the first round the verifier can measure part of the prover's first message $|\psi\rangle$, obtaining a classical outcome ℓ together with the corresponding post-measurement state $|\psi_\ell\rangle$ on the remaining qubits. The verifier's subsequent actions can then depend on ℓ and make use of $|\psi_\ell\rangle$. We refer to this capability—measuring part a state to extract classical information while retaining the residual post-measurement state—as *post-measurement branching*. Broadly, our goal in this section is to better understand the power of post-measurement branching in unentangled proof systems.

We formally define $\mathsf{QMACM}_{\mathsf{unent}}$ below:

Definition 3.1 ($\mathsf{QMACM}_{\mathsf{unent}}$)**.** A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in $\mathsf{QMACM}_{\mathsf{unent}}[c, s]$ for polynomial-time computable functions $c, s : \mathbb{N} \to [0, 1]$ if there exists a $\mathsf{QIP}[3]$ verification procedure V with the following properties:

- *Public coin tosses.* V_1^x samples a random $x \sim \{0,1\}^{\text{poly}(n)}$, records the result in $\mathcal{V}$, and copies it into $\mathcal{M}$.
- *Classical final message.* P_2^x sends a classical message. Equivalently, V_2^x measures $\mathcal{M}$ in the computational basis before any other processing.
- *Completeness.* For all $x \in A_{\text{yes}}$, there exists an *unentangled* prover P that makes V accept with probability at least $c(|x|)$.
- *Soundness.* For all $x \in A_{\text{no}}$, every *unentangled* prover P makes V accept with probability at most $s(|x|)$.

We define $\mathsf{QMACM}_{\mathsf{unent}} := \mathsf{QMACM}_{\mathsf{unent}}[2/3, 1/3]$.

It is not immediately clear whether $\mathsf{QMACM}_{\mathsf{unent}}$ admits amplification of the completeness and soundness parameters, although we will show its containment in a class that does (Theorem 3.5).

3.1 Upper Bounding $\mathsf{QMACM}_{\mathsf{unent}}$

We now turn to the main result of this section: $\mathsf{QMACM}_{\mathsf{unent}} \subseteq \mathsf{QAM}$. For completeness, we also include the definition of QAM, following Marriott and Watrous [MW05].

Definition 3.2 (QAM)**.** A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in $\mathsf{QAM}[c, s]$ if there exists a polynomial-time quantum algorithm $Q(x, y, |\psi\rangle)$ such that:

- *Completeness.* For all $x \in A_{\text{yes}}$, there exists a collection of $\text{poly}(n)$-qubit states $\{|\psi_y\rangle : y \in \{0,1\}^{\text{poly}(|x|)}\}$ such that $\Pr_y[Q(x, y, |\psi_y\rangle) = 1] \geq c(|x|)$.
- *Soundness.* For all $x \in A_{\text{no}}$, for every collection of $\text{poly}(n)$-qubit states $\{|\psi_y\rangle : y \in \{0,1\}^{\text{poly}(|x|)}\}$, $\Pr_y[Q(x, y, |\psi_y\rangle) = 1] \leq s(|x|)$.

We define $\mathsf{QAM} := \mathsf{QAM}[2/3, 1/3]$.

The definition differs stylistically from our definition of $\mathsf{QIP}[3]$ (Definition 2.1), which involved the prover and verifier applying alternating quantum channels, only because it is easier to abstract the prover's strategy into a single mapping from strings y to states $|\psi_y\rangle$. In particular, the string y represents Arthur's random coin tosses sent to Merlin, and $|\psi_y\rangle$ is Merlin's response. We also remark that QAM admits amplification by parallel repetition: $\mathsf{QAM}[c,s] = \mathsf{QAM}[1-2^{-\mathrm{poly}}, 2^{-\mathrm{poly}}]$ as long as $c-s$ is inverse-polynomially bounded [MW05, Theorem 4.2].

Our proof relies on the existence of an algorithm that compresses a quantum state into a polynomial-sized classical description from which the expectation values of many observables can later be estimated. This algorithm derives from a theorem of Aaronson about simulating bounded-error one-way quantum communication with classical communication, which was in turn used to show that $\mathsf{BQP/qpoly} \subseteq \mathsf{PP/poly}$ [Aar05a]. In fact, the lemma below is essentially *equivalent* to the non-existence of a superpolynomial quantum advantage in one-way communication complexity for a decision problem.

Lemma 3.3. *Let $M_1, \ldots, M_K$ be m-qubit measurement operators (i.e., PSD matrices with $0 \preceq M_i \preceq 1$ for each i), and fix an error parameter ε. There exist functions $\mathsf{Stat}(\rho, \varepsilon)$ and $\mathsf{Est}(s, i)$ with the following properties:*

1. *$\mathsf{Stat}(\rho, \varepsilon)$ takes as input a classical description of an m-qubit state ρ and an error parameter ε, and outputs a classical string of length $O(\log K \frac{m}{\varepsilon^2} \log \frac{m}{\varepsilon})$.*
2. *For any m-qubit ρ and any $i \in [K]$, $\mathsf{Est}(\mathsf{Stat}(\rho, \varepsilon), i)$ outputs a number that is ε-close to $\mathrm{tr}(\rho M_i)$.*

Proof. Consider a one-way communication problem between two parties, Alice and Bob, in which Alice receives a description of an m-qubit state ρ, and Bob receives an index $i \in [K]$ and a parameter $t \in [0,1]$ that is described to $O(\log \frac{1}{\varepsilon})$ bits of precision. Their goal is to decide whether $\mathrm{tr}(M_i \rho) \geq t + \varepsilon/10$ or $\mathrm{tr}(M_i \rho) \leq t - \varepsilon/10$, promised that one of these is the case.

Observe that this problem admits a bounded-error one-way communication protocol with complexity $O(\frac{m}{\varepsilon^2})$: Alice sends Bob $O(\frac{1}{\varepsilon^2})$ copies of ρ, Bob measures M_i on each of the copies, and accepts if and only if the sample mean is greater than t. By the simulation theorem for quantum one-way communication with classical communication [Aar05a, Theorem 3.4], this same problem admits a deterministic *classical* communication protocol in which Alice sends $O(\log K \frac{m}{\varepsilon^2} \log \frac{m}{\varepsilon})$ bits to Bob.

The two functions Stat and Est derive directly from this classical communication protocol. The encoding function Stat is simply the function that Alice uses to map ρ to a classical message. The decoding function Est runs Bob's half of the computation for $O(\frac{1}{\varepsilon})$ different values of t to find one that is within ε of $\mathrm{tr}(\rho M_i)$. The correctness of these two functions follows from the correctness of the classical communication protocol. □

We use the compression lemma above to argue that Arthur can "subsample" from the second-round messages while approximately maintaining the $\mathsf{QMACM_{unent}}$ acceptance probability. In this next lemma, the $\max_\rho$, E_y, and $\max_z$ operators correspond respectively to the first-, second-, and third-round messages of the $\mathsf{QMACM_{unent}}$ protocol. In plain words, the lemma says that if Arthur samples his message y from a random polynomial-size subset of $\{0,1\}^m$ instead of uniformly over $\{0,1\}^m$, then with high probability over the chosen subset, the completeness probability of the protocol is approximately unchanged.

Lemma 3.4. *Let $M_{y,z}$ be an m-qubit quantum measurement for each $y, z \in \{0,1\}^m$. Then except with probability at most $\exp(O(\frac{m^2}{\varepsilon^2}\log\frac{m}{\varepsilon}) - 2\varepsilon^2 r/9))$ over $y_1, \ldots, y_r \sim \{0,1\}^m$,*

$$\left| \max_{|\psi\rangle \in \mathcal{D}(m)} \mathop{\mathrm{E}}_{y\sim\{0,1\}^m} \max_{z\in\{0,1\}^m} \langle\psi|M_{y,z}|\psi\rangle - \max_{|\psi\rangle \in \mathcal{D}(m)} \mathop{\mathrm{E}}_{i\sim[r]} \max_{z\in\{0,1\}^m} \langle\psi|M_{y_i,z}|\psi\rangle \right| \leq \varepsilon. \tag{1}$$

Proof. Define

$$f(|\psi\rangle) \coloneqq \mathop{\mathrm{E}}_{y\sim\{0,1\}^m} \max_{z\in\{0,1\}^m} \langle\psi|M_{y,z}|\psi\rangle$$

and

$$g(|\psi\rangle) \coloneqq \max_{|\psi\rangle \in \mathcal{D}(m)} \mathop{\mathrm{E}}_{i\sim[r]} \max_{z\in\{0,1\}^m} \langle\psi|M_{y_i,z}|\psi\rangle.$$

(The latter is a slight abuse of notation, because g additionally depends on $y_1, \ldots, y_r$.)

Let Stat and Est be the functions from Lemma 3.3. For any string s, we claim that except with probability at most $\exp\bigl(-2\varepsilon^2 r/9\bigr)$ over the choices of $y_1, \ldots, y_r$,

$$\mathop{\mathrm{E}}_{y\sim\{0,1\}^m} \max_{z\in\{0,1\}^m} \mathsf{Est}(s,(y,z)) \leq \mathop{\mathrm{E}}_{i\sim[r]} \max_{z\in\{0,1\}^m} \mathsf{Est}(s,(y_i,z)) + \frac{\varepsilon}{3}.$$

This is a straightforward consequence of Hoeffding's inequality. Thus, for any state $|\psi\rangle$ satisfying $\mathsf{Stat}(|\psi\rangle\langle\psi|, \varepsilon/3) = s$, we have

$$\begin{aligned} f(|\psi\rangle) &\leq \mathop{\mathrm{E}}_{y\sim\{0,1\}^m} \max_{z\in\{0,1\}^m} \mathsf{Est}(s,(y,z)) + \frac{\varepsilon}{3} \\ &\leq \mathop{\mathrm{E}}_{i\sim[r]} \max_{z\in\{0,1\}^m} \mathsf{Est}(s,(y_i,z)) + \frac{2\varepsilon}{3} \\ &\leq g(|\psi\rangle) + \varepsilon, \end{aligned}$$

except with probability at most $\exp\bigl(-2\varepsilon^2 r/9\bigr)$, because of the correctness guarantee of Stat and Est. We similarly obtain $g(|\psi\rangle) \leq f(|\psi\rangle) + \varepsilon$ with probability at most $\exp\bigl(-2\varepsilon^2 r/9\bigr)$, which implies $|f(|\psi\rangle) - g(|\psi\rangle)| \leq \varepsilon$ with probability at most $2\exp\bigl(-2\varepsilon^2 r/9\bigr)$. Now, apply a union bound over all possible strings s of length $O(\frac{m^2}{\varepsilon^2}\log\frac{m}{\varepsilon})$ that can be output by $\mathsf{Stat}(|\psi\rangle, \varepsilon/3)$ to conclude that, except with probability at most $\exp(O(\frac{m^2}{\varepsilon^2}\log\frac{m}{\varepsilon}) - 2\varepsilon^2 r/9))$, for *every* $|\psi\rangle$, $|f(|\psi\rangle) - g(|\psi\rangle)| \leq \varepsilon$. From this it follows that $|\max_{|\psi\rangle\in\mathcal{D}(m)} f(|\psi\rangle) - \max_{|\psi\rangle\in\mathcal{D}(m)} g(|\psi\rangle)| \leq \varepsilon$, which is equivalent to the statement of the lemma. □

Note that in the above lemma, the only use of Lemma 3.3 is in arguing that we can "union bound" over all m-qubit quantum states $|\psi\rangle$ as if there were only $2^{\mathrm{poly}(m)}$ possible choices of ρ, instead of $2^{2^{\mathrm{poly}(m)}}$. If $|\psi\rangle$ were an m-bit string rather than an m-qubit quantum state, then one could *directly* union bound over the choices of $|\psi\rangle$ instead of going through $\mathsf{Stat}(|\psi\rangle\langle\psi|, \varepsilon)$.

We now use Lemma 3.4 to place $\mathsf{QMACM}_{\mathsf{unent}}$ in QAM by combining Merlin's first and third messages into one. The idea is for Arthur to choose a $\mathrm{poly}(m)$-size subsample of second-round random challenges, to which Merlin responds with both his original first-round quantum message and third-round responses to each of Arthur's random challenges.

Theorem 3.5. *For any $c(n) - s(n) \geq \frac{1}{\mathrm{poly}(n)}$, $\mathsf{QMACM}_{\mathsf{unent}}[c,s] \subseteq \mathsf{QAM}[c,(c+s)/2] \subseteq \mathsf{QAM}$.*

Proof. Suppose a promise problem $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ admits a $\mathsf{QMACM}_{\mathsf{unent}}[c,s]$ protocol in which given an input x of length n, the message register $\mathcal{M}$ has exactly $m = \mathrm{poly}(n)$ qubits. As a consequence of Lemma 2.5 and the classical final message, it is without loss of generality that

Merlin initially sends Arthur an m-qubit state $|\psi\rangle$, Arthur responds with a challenge $y \sim \{0,1\}^m$, and Merlin lastly sends Arthur $z \in \{0,1\}^m$ that is a deterministic function of y. Let $M_{y,z}$ be the measurement that Arthur applies to $|\psi\rangle$ given y and z, so that his acceptance probability is exactly $\langle\psi|M_{y,z}|\psi\rangle$. Then, Merlin's best strategy achieves acceptance probability exactly

$$\max_{|\psi\rangle\in\mathcal{D}(m)} \mathop{\mathrm{E}}_{y\sim\{0,1\}^m} \max_{z\in\{0,1\}^m} \langle\psi|M_{y,z}|\psi\rangle. \tag{2}$$

Now consider the following QAM protocol for A. For some $r = \mathrm{poly}(n)$ to be chosen later, Arthur starts by sending Merlin $m \cdot r$ random bits, which we interpret as challenge strings $y_1, \ldots, y_r$ each of length m. Merlin's response consists of an m-qubit state $|\psi\rangle$ and strings $z_1, \ldots, z_r \in \{0,1\}^m$. Arthur picks $i \in [r]$ at random, then measures M_{y_i,z_i} on ρ. The maximum acceptance probability of this QAM protocol is

$$\mathop{\mathrm{E}}_{y_1,\ldots,y_r\sim\{0,1\}^m} \max_{\substack{|\psi\rangle\in\mathcal{D}(m)\\ z_1,\ldots,z_r\in\{0,1\}^m}} \mathop{\mathrm{E}}_{i\in[r]} \langle\psi|M_{y_i,z_i}|\psi\rangle.$$

This is evidently equal to

$$\mathop{\mathrm{E}}_{y_1,\ldots,y_r\sim\{0,1\}^m} \max_{|\psi\rangle\in\mathcal{D}(m)} \mathop{\mathrm{E}}_{i\in[r]} \max_{z\in\{0,1\}^m} \langle\psi|M_{y_i,z}|\psi\rangle \tag{3}$$

because for fixed ρ, the choice of z_i only affects the ith term in the inner expectation, so to maximize the expectation it is optimal to maximize each term separately.

We claim that the QAM protocol has completeness at least c. Given Merlin's optimal strategy for the $\mathsf{QMACM}_{\mathsf{unent}}$ protocol, Merlin's strategy for the QAM protocol is to choose the same $|\psi\rangle$ that he would in the $\mathsf{QMACM}_{\mathsf{unent}}$ protocol, and then for each $i \in [r]$ to let z_i be the string z that he would send upon receiving $y = y_i$ from Arthur. This strategy for the QAM protocol causes Arthur to accept with probability c given any fixed $i \in [r]$, so the acceptance probability is certainly still c when averaging over $i \in [r]$.

It remains to bound the soundness. We do so by upper bounding the acceptance probability of the QAM protocol (Equation (3)) in terms of that of the $\mathsf{QMACM}_{\mathsf{unent}}$ protocol (Equation (2)), via application of Lemma 3.4. Let $p \leq \exp(O(\frac{m^2}{\varepsilon^2}\log\frac{m}{\varepsilon}) - 2\varepsilon^2 r/9))$ be the probability that Equation (1) fails to hold in Lemma 3.4. Then the QAM acceptance probability is bounded by

$$\mathop{\mathrm{E}}_{y_1,\ldots,y_r\sim\{0,1\}^m} \max_{\rho\in\mathcal{D}(m)} \mathop{\mathrm{E}}_{i\in[r]} \max_{z\in\{0,1\}^m} \mathrm{tr}(\rho M_{y_i,z_i}) \leq p + \varepsilon + \max_{\rho\in\mathcal{D}(m)} \mathop{\mathrm{E}}_{y\sim\{0,1\}^m} \max_{z\in\{0,1\}^m} \mathrm{tr}(\rho M_{y,z}).$$

Choose $\varepsilon = (c-s)/4$ and $r = O(\frac{m^2}{\varepsilon^4}\log\frac{m}{\varepsilon}) \leq \mathrm{poly}(|x|)$ so that $p \leq \varepsilon$. Then assuming $x \in A_{\mathrm{no}}$, the right hand side is at most $(c-s)/4 + (c-s)/4 + s = (c+s)/2$.

To summarize, we have shown that A admits a $\mathsf{QAM}[c,(c+s)/2]$ protocol. The completeness/soundness can be amplified to $[2/3, 1/3]$, or even further to $[1-2^{-\mathrm{poly}}, 2^{-\mathrm{poly}}]$ as shown by Marriott and Watrous [MW05, Theorem 4.2]. □

Observe that nowhere in the proof was it crucial for Arthur's message to consist of uniformly random public coins. One could instead envision a class "$\mathsf{qcc\text{-}QIP}^{\mathsf{no\text{-}pmb}}_{\mathsf{unent}}[3]$" between $\mathsf{QMACM}_{\mathsf{unent}}$ and $\mathsf{QIP}_{\mathsf{unent}}[3]$ in which the verifier's initial classical message is produced by an efficient quantum procedure acting only on the register $\mathcal{V}$, but the final prover message remains classical. Formally, that would mean V_1^x performs a measurement on $\mathcal{V}$ independent of the received message on $\mathcal{M}$, then swaps the measurement with the prover's message in $\mathcal{M}$. Intuitively, $\mathsf{qcc\text{-}QIP}^{\mathsf{no\text{-}pmb}}_{\mathsf{unent}}[3]$ is like $\mathsf{QIP}[3]$ except that only the first message is quantum and the verifier is not allowed to use post-measurement branching on the first quantum message. Then the same proof above would show containment of $\mathsf{qcc\text{-}QIP}^{\mathsf{no\text{-}pmb}}_{\mathsf{unent}}[3]$ in $\mathsf{QIP}[2]$. We did not attempt to define such a complexity class formally because it seems impossible to give it a short but sufficiently descriptive name.

4 Constant-Round Quantum-Classical Interactive Proofs

Section 2 and Section 3 illustrate that post-measurement branching is an interesting feature of quantum proof systems, and in particular one of the key ingredients that enables three-round unentangled protocols to capture NEXP. In this section, we turn to the simplest setting where post-measurement branching can be isolated: the classes QCAM and QCIP[2].

Both QCAM and QCIP[2] are two-round interactions initiated by Arthur. In QCAM, Arthur's message consists of uniformly random classical bits, whereas in QCIP[2] Arthur may prepare a state $|\psi\rangle$, measure part of it to obtain a message ℓ and the post-measurement state $|\psi_\ell\rangle$, and then use both in the remainder of the verification procedure.

The main results of this section are summarized in the following theorem.

Theorem 4.1. $\mathsf{QCAM} = \mathsf{BP} \cdot \mathsf{QCMA} \subseteq \mathsf{BQ} \cdot \mathsf{QCMA} \subseteq \mathsf{QCIP}[2] \subseteq \mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}}}$.

We will record formal definitions shortly.

4.1 Quantum-Classical Arthur-Merlin Games

We begin by establishing that $\mathsf{QCAM} = \mathsf{BP} \cdot \mathsf{QCMA}$. This equivalence is not new: as far as we can tell, it first appeared in Marriott's Master's thesis [Mar03, Theorem 8]. The upper bound $\mathsf{QCAM} \subseteq \mathsf{BP} \cdot \mathsf{QCMA}$ was later reproved in [LG17, Proposition 33], and the equivalence was stated without proof or citation in [AH23]. We include the argument here to present a complete proof in more standard notation than that of Marriott's thesis.

We start by defining QCAM and the BP operator.

Definition 4.2 (QCAM)**.** A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in $\mathsf{QCAM}[c, s]$ if there exists a polynomial-time quantum algorithm $Q(x, y, z)$ such that:

- *Completeness.* For all $x \in A_{\text{yes}}$, there exists a collection of $\text{poly}(n)$-length bit strings $\{z_y : y \in \{0,1\}^{\text{poly}(|x|)}\}$ such that $\Pr_y[Q(x, y, z_y) = 1] \geq c(|x|)$.
- *Soundness.* For all $x \in A_{\text{no}}$, for every collection of $\text{poly}(n)$-length bit strings $\{z_y : y \in \{0,1\}^{\text{poly}(|x|)}$, $\Pr_y[Q(x, y, z_y) = 1] \leq s(|x|)$.

We define $\mathsf{QCAM} \coloneqq \mathsf{QCAM}[2/3, 1/3]$.

Similar to QAM (Definition 3.2), the string y represents Arthur's random coin tosses sent to Merlin, and z_y is Merlin's response. One can more generally define $\mathsf{QCAM}[k]$ with $k \geq 2$ rounds of interaction. However, this class collapses to QCAM [KGN19, Theorem 7(iv)], in analogy with the collapse of the AM hierarchy [Bab85; BM88]. QCAM also admits amplification by running many iid trials in parallel and taking the majority vote: $\mathsf{QCAM}[c, s] = \mathsf{QCAM}[1 - 2^{-\text{poly}}, 2^{-\text{poly}}]$ as long as $c - s$ is inverse-polynomially bounded. A proof of this amplification trick is straightforward, and is also a special case of [KGN19, Lemma 20].

We now define the BP operator, which has a few essentially equivalent definitions in the literature (see [BGW24, Section 3.1]). For our purposes, the following definition is most convenient.

Definition 4.3 (BP operator)**.** Let $\mathcal{C}$ be any class of promise problems. Then $\mathsf{BP} \cdot \mathcal{C}$ consists of all promise problems $A = (A_{\text{yes}}, A_{\text{no}})$ for which there exist a promise problem $B = (B_{\text{yes}}, B_{\text{no}}) \in \mathcal{C}$ and a polynomial p such that, for every input x of length n,

- Completeness: If $x \in A_{\text{yes}}$, then $\Pr_y[(x, y) \in B_{\text{yes}}] \geq 2/3$,

- Soundness: If $x \in A_{\text{no}}$, then $\Pr_y[(x,y) \in B_{\text{no}}] \geq 2/3$,

where $y \in \{0,1\}^{p(n)}$ is uniformly distributed.

Theorem 4.4. $\mathsf{QCAM} = \mathsf{BP} \cdot \mathsf{QCMA}$.

Proof. Let A be a promise problem in $\mathsf{BP} \cdot \mathsf{QCMA}$ and let B be the corresponding promise problem in QCMA to which A reduces under Definition 4.3. Consider a QCMA verifier $Q(x,y,z)$ for B where (x,y) is the input to B and z is the witness. Suppose without loss of generality that completeness and soundness parameters of Q are amplified to 0.9 and 0.1, respectively. We claim that Q is also a QCAM verifier for A with completeness 0.6 and soundness 0.4. In particular:

1. If $x \in A_{\text{yes}}$, then $\Pr_y[(x,y) \in B_{\text{yes}}] \geq 2/3$, and therefore $\mathrm{E}_y[\max_z \Pr[Q(x,y,z) = 1)]] \geq 2/3 \cdot 0.9 = 0.6$.
2. If $x \in A_{\text{no}}$, then $\Pr_y[(x,y) \in B_{\text{no}}] \geq 2/3$, and therefore $\mathrm{E}_y[\max_z \Pr[Q(x,y,z) = 1)]] \leq 1/3 \cdot 1 + 2/3 \cdot 0.1 = 0.4$.

Put another way, the QCAM protocol is for Arthur to send Merlin random coin tosses y for which (x,y) forms an instance of B, and then to run the QCMA verifier on (x,y) and Merlin's response z. Hence $A \in \mathsf{QCAM}[0.6, 0.4] \subseteq \mathsf{QCAM}$ because one can amplify $(0.6, 0.4)$ to $(2/3, 1/3)$ by parallel repetition. This establishes $\mathsf{BP} \cdot \mathsf{QCMA} \subseteq \mathsf{QCAM}$.

For the converse, let A be a promise problem in QCAM. Take a QCAM verifier $Q(x,y,z)$ with completeness 0.9 and soundness 0.1. Let B be the QCMA promise problem parameterized by Q, where (x,y) is interpreted as the input to B and z is the witness. That is, $(x,y) \in B_{\text{yes}}$ if there is a z that causes $Q(x,y,z)$ to accept with probability at least $2/3$, and $(x,y) \in B_{\text{no}}$ if every z causes $Q(x,y,z)$ to accept with probability at most $1/3$. We claim that B shows $A \in \mathsf{BP} \cdot \mathsf{QCMA}$. In particular:

- If $x \in A_{\text{yes}}$, then $\mathrm{E}_y[\max_z \Pr[Q(x,y,z) = 1]] \geq 0.9$. Because $\Pr[Q(x,y,z)] \in [0,1]$, it must be the case that $\Pr_y[\max_z \Pr[Q(x,y,z) = 1] \geq 2/3] \geq 0.7$, and therefore $\Pr_y[(x,y) \in B_{\text{yes}}] \geq 0.7$.
- If $x \in A_{\text{no}}$, then $\mathrm{E}_y[\max_z \Pr[Q,x,y,z) = 1]] \leq 0.1$. Because $\Pr[Q(x,y,z)] \in [0,1]$, it must be the case that $\Pr_y[\max_z \Pr[Q(x,y,z) = 1] \leq 1/3] \geq 0.7$, and therefore $\Pr_y[(x,y) \in B_{\text{no}}] \geq 0.7$.

This shows that B satisfies Definition 4.3 with respect to A. □

We note that a similar result, $\mathsf{QAM} = \mathsf{BP} \cdot \mathsf{QMA}$, was also shown in Marriott's thesis [Mar03, Theorem 12]. A proof follows from an identical argument to that of Theorem 4.4.

4.2 Quantum-Classical Interactive Proofs

Here we establish $\mathsf{BQ} \cdot \mathsf{QCMA} \subseteq \mathsf{QCIP}[2] \subseteq \mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}}}$. We start with some definitions.

A $\mathsf{QCIP}[2]$ verification procedure is specified by a polynomial-time uniformly generated family of quantum circuits $V = \{V_1^x, V_2^x : x \in \{0,1\}^*$. On input x of length n, these circuits determine the verifier's actions across the two-round interaction. Each circuit acts on registers of size $\mathrm{poly}(n)$, partitioned into a message register $\mathcal{M}$, exchanged with the prover, and a verifier workspace register $\mathcal{V}$, retained by the verifier. The prover is modeled by an unrestricted family of quantum operations $P = \{P^x : x \in \{0,1\}^*\}$ that act on the same message register $\mathcal{M}$ and a *prover workspace* register $\mathcal{P}$ (which need not be polynomially-bounded in size).

The defining restriction of $\mathsf{QCIP}[2]$ is that *the message register is measured in the computational basis before transmision.* Equivalently, the final operation of both V_1^x and P^x on $\mathcal{M}$ is a computational basis measurement. The interaction proceeds as follows:

1. The three registers $\mathcal{V}, \mathcal{M}, \mathcal{P}$ are initialized to the all-zeros state.
2. The verifier applies V_1^x to $\mathcal{P}$ and $\mathcal{M}$. The register $\mathcal{M}$ is then measured in the computational basis and sent as a classical message.
3. The prover applies P^x to $\mathcal{P}$ and $\mathcal{M}$. Again, $\mathcal{M}$ is measured in the computational basis before being sent back.
4. Finally, the verifier applies V_2^x to $(\mathcal{V}, \mathcal{M})$ and measures a designated output qubit to decide acceptance or rejection.

We now formally define the class $\mathsf{QCIP}[2]$.

Definition 4.5 ($\mathsf{QCIP}[2]$)**.** A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in $\mathsf{QCIP}[2, c, s]$ for polynomial-time computable functions $c, s : \mathbb{N} \to [0, 1]$ if there exists a $\mathsf{QCIP}[2]$ verification procedure V with the following properties:

- Completeness: For all $x \in A_{\text{yes}}$, there exists a quantum prover P that causes V to accept x with probability at least $c(|x|)$.
- Soundness: For all $x \in A_{\text{no}}$, every quantum prover P causes V to accept x with probability at most $s(|x|)$.

We define $\mathsf{QCIP}[2] \coloneqq \mathsf{QCIP}[2, 2/3, 1/3]$.

We note that error reduction in $\mathsf{QCIP}(2)$ can be achieved by parallel repetition: running multiple independent instances of the protocol in parallel and deciding acceptance by a majority vote.

Next, we define the BQ operator, which was recently formalized by Buhrman, Le Gall, and Weggemans [BGW24].Intuitively, $\mathsf{BQ} \cdot \mathcal{C}$ consists of those promise problems that admit polynomial-time *quantum* reductions to problems in $\mathcal{C}$.

Definition 4.6 (BQ operator)**.** Let $\mathcal{C}$ be any class of promise problems. Then $\mathsf{BQ} \cdot \mathcal{C}$ consists of all promise problems $A = (A_{\text{yes}}, A_{\text{no}})$ for which there exist a promise problem $B = (B_{\text{yes}}, B_{\text{no}}) \in \mathcal{C}$ and a polynomial-time quantum algorithm $\mathcal{A}$ such that, for every input x of length n,

- Completeness: If $x \in A_{\text{yes}}$, then $\Pr_{y \sim \mathcal{A}(x)}[(x, y) \in B_{\text{yes}}] \geq 2/3$,
- Soundness: If $x \in A_{\text{no}}$, then $\Pr_{y \sim \mathcal{A}(x)}[(x, y) \in B_{\text{no}}] \geq 2/3$.

Here the distribution $y \sim \mathcal{A}(x)$ is obtained by running $\mathcal{A}$ on input x and measuring a designated polynomial-size output register in the computational basis.

We will now prove lower and upper bounds on $\mathsf{QCIP}[2]$.

Theorem 4.7. $\mathsf{BQ} \cdot \mathsf{QCMA} \subseteq \mathsf{QCIP}[2]$.

Proof. Let A be a promise problem in $\mathsf{BQ} \cdot \mathsf{QCMA}$. Let B be the corresponding promise problem in QCMA to which A reduces via the polynomial-time quantum algorithm $\mathcal{A}$ under Definition 4.3. Consider a QCMA verifier $Q(x, y, z)$ for B where (x, y) is the input to B and z is the witness. Suppose without loss of generality that completeness and soundness parameters of Q are amplified to 0.9 and 0.1, respectively. We claim that Q is also a $\mathsf{QCIP}[2]$ verifier for A with completeness 0.6 and soundness 0.4. In particular:

1. If $x \in A_{\text{yes}}$, then $\Pr_{y \sim \mathcal{A}(x)}[(x, y) \in B_{\text{yes}}] \geq 2/3$, and therefore $\mathrm{E}_{y \sim \mathcal{A}(x)}[\max_z \Pr[Q(x, y, z) = 1)]] \geq 2/3 \cdot 0.9 = 0.6$.

2. If $x \in A_{\text{no}}$, then $\Pr_{y \sim \mathcal{A}(x)}[(x,y) \in B_{\text{no}}] \geq 2/3$, and therefore $\mathrm{E}_{y \sim \mathcal{A}(x)}[\max_z \Pr[Q(x,y,z) = 1)]] \leq 1/3 \cdot 1 + 2/3 \cdot 0.1 = 0.4$.

The protocol is for Arthur to send to Merlin $y = \mathcal{A}(x)$ for which (x,y) forms an instance of B, and then to run the QCMA verifier on (x,y) and Merlin's response z. Hence $A \in \mathsf{QCIP}[2, 0.6, 0.4] \subseteq \mathsf{QCIP}[2]$ because one can amplify $(0.6, 0.4)$ to $(2/3, 1/3)$ by parallel repetition. This establishes $\mathsf{BQ} \cdot \mathsf{QCMA} \subseteq \mathsf{QCIP}[2]$. □

Theorem 4.8. $\mathsf{QCIP}[2] \subseteq \mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}}}$.

Proof. Let $V = \{V_1^x, V_2^x\}$ be a verifier with completeness 0.9 and soundness 0.1 for some promise problem $A \in \mathsf{QCIP}[2]$. The PP language will be specified by a PostBQP promise problem, because $\mathsf{PP} = \mathsf{PostBQP}$ [Aar05b] and any PP promise problem can be extended to a PP language, as PP is a syntactic class. Given a tuple (x,y,z), consider a QCIP[2] interaction between verifier and prover in which we postselect on V_1^x sending the classical message y, and the prover responds with z. Then deciding whether this postselected interaction between prover and verifier causes V to accept with probability at least $2/3$ (yes) or at most $1/3$ (no) is clearly in $\mathsf{PostBQP} = \mathsf{PP}$. The $\mathsf{NP}^{\mathsf{PP}}$ language, then, will be: given (x,y), decide whether there exists a string z for which (x,y,z) is a yes instance of the PP language. Finally, consider the following $\mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}}}$ machine that we claim decides A: run V_1^x to obtain $y \in \{0,1\}^{\text{poly}(|x|)}$, query the $\mathsf{NP}^{\mathsf{PP}}$ language on (x,y), and accept if and only if (x,y) is a yes-instance. This machine works because:

- If $x \in A_{\text{yes}}$, then $\mathrm{E}_{y \sim V_1^x}[\max_z \Pr[V \text{ accepts } z \mid y]] \geq 0.9$, and therefore $\Pr_{y \sim V_1^x}[\max_z \Pr[V \text{ accepts } z \mid y] \geq 2/3] \geq 0.7$. Hence, with probability at least 0.7 over the y sampled by the BQP machine, there exists a z for which (x,y,z) is a yes-instance of the PP language, and thus the BQP machine accepts.
- If $x \in A_{\text{no}}$, then $\mathrm{E}_{y \sim V_1^x}[\max_z \Pr[V \text{ accepts } z \mid y]] \leq 0.1$, and therefore $\Pr_{y \sim V_1^x}[\max_z \Pr[V \text{ accepts } z \mid y] \leq 1/3] \geq 0.7$. Hence, with probability at least 0.7 over the y sampled by the BQP machine, for every z (x,y,z) is a no-instance of the PP language, and thus the BQP machine rejects.

So, the $\mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}}}$ machine decides A with error probability at most 0.4, which can of course be amplified to arbitrarily small error. □

We conclude with several remarks regarding Theorem 4.7. First, unlike $\mathsf{BP} \cdot \mathsf{QCMA} = \mathsf{QCAM}$, the class QCIP[2] is not characterized by $\mathsf{BQ} \cdot \mathsf{QCMA}$. The reason is that $\mathsf{BQ} \cdot \mathsf{QCMA}$ does not capture the additional power conferred by post-measurement branching. In fact, one can show that the subclass of QCIP[2] in which post-measurement branching is disallowed is precisely equal to $\mathsf{BQ} \cdot \mathsf{QCMA}$.

Second, our upper bound generalizes straightforwardly to k-round interactions. Specifically, one can show that $\mathsf{QCIP}[2k]$ is contained in a tower of classes of the form

$$\mathsf{BQP}^{\mathsf{NP}^{\mathsf{PP}^{\mathsf{NP}^{\mathsf{PP}^{\cdots}}}}},$$

which in particular implies that $\mathsf{QCIP}[k] \subseteq \mathsf{CH}$ for any constant k.

Finally, by the same reasoning as in Theorem 4.7, one can show that $\mathsf{BQ} \cdot \mathsf{QMA} \subseteq \mathsf{QIP}[2]$. However, obtaining an upper bound on QIP[2] stronger than PSPACE is unclear: our proof technique in Theorem 4.8 breaks when Merlin's message is quantum, as it is in QIP[2].

Acknowledgments

We thank Scott Aaronson for clarifications regarding Lemma 3.3.

S.G. is supported in part by an IBM Ph.D. Fellowship. W.K. is supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator, and by NSF Grant CCF-231173. This work was done in part while S.G. was visiting the Simons Institute for the Theory of Computing, supported by NSF Grant QLCI-2016245.

References

[Aar05a] Scott Aaronson. "Limitations of Quantum Advice and One-Way Communication". In: *Theory of Computing* (2005). DOI: 10.4086/toc.2005.v001a001 (pages 4, 13).

[Aar05b] Scott Aaronson. "Quantum computing, postselection, and probabilistic polynomial-time". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2005). DOI: 10.1098/rspa.2005.1546 (pages 5, 19).

[Aar06] Scott Aaronson. "QMA/qpoly ⊆ PSPACE/poly: De-Merlinizing Quantum Protocols". In: *Proceedings of the 21st Annual IEEE Conference on Computational Complexity*. 2006. DOI: 10.1109/CCC.2006.36 (page 4).

[ABDFS09] Scott Aaronson, Salman Beigi, Andrew Drucker, Bill Fefferman, and Peter Shor. "The Power of Unentanglement". In: *Theory of Computing* (2009). DOI: 10.4086/toc.2009.v005a001 (page 2).

[AGIMR24] Scott Aaronson, Sabee Grewal, Vishnu Iyer, Simon C. Marshall, and Ronak Ramachandran. PDQMA = DQMA = NEXP*:* QMA *With Hidden Variables and Non-collapsing Measurements*. 2024. arXiv: 2403.02543 [quant-ph] (page 2).

[AH23] Scott Aaronson and Shih-Han Hung. "Certified Randomness from Quantum Supremacy". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023. DOI: 10.1145/3564246.3585145 (pages 4, 16).

[Bab85] László Babai. "Trading group theory for randomness". In: *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*. 1985. DOI: 10.1145/22145.22192 (pages 4, 5, 16).

[BFL91] László Babai, Lance Fortnow, and Carsten Lund. "Nondeterministic exponential time has two-prover interactive protocols". In: *Computational Complexity* (1991). DOI: 10.1007/BF01200056 (page 1).

[BFLMW24] Roozbeh Bassirian, Bill Fefferman, Itai Leigh, Kunal Marwaha, and Pei Wu. *Quantum Merlin-Arthur with an internally separable proof*. 2024. arXiv: 2410.19152 [quant-ph] (page 2).

[BFM24] Roozbeh Bassirian, Bill Fefferman, and Kunal Marwaha. "Quantum Merlin-Arthur and Proofs Without Relative Phase". In: *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*. 2024. DOI: 10.4230/LIPIcs.ITCS.2024.9 (page 2).

[BGW24] Harry Buhrman, François Le Gall, and Jordi Weggemans. *Classical versus quantum queries in quantum PCPs with classical proofs*. 2024. arXiv: 2411.00946 [quant-ph] (pages 16, 18).

[BM25] Roozbeh Bassirian and Kunal Marwaha. "Superposition detection and QMA with non-collapsing measurements". In: *Quantum* (2025). DOI: 10.22331/q-2025-08-28-1839 (page 2).

[BM88] László Babai and Shlomo Moran. "Arthur-Merlin games: A randomized proof system, and a hierarchy of complexity classes". In: *Journal of Computer and System Sciences* (1988). DOI: 10.1016/0022-0000(88)90028-1 (pages 4, 5, 16).

[BT12] Hugue Blier and Alain Tapp. "A Quantum Characterization Of NP". In: *Computational Complexity* (2012). DOI: 10.1007/s00037-011-0016-2 (page 2).

[CD10] Jing Chen and Andrew Drucker. *Short Multi-Prover Quantum Proofs for SAT without Entangled Measurements*. 2010. arXiv: 1011.0716 [quant-ph] (page 2).

[CF13] Alessandro Chiesa and Michael A. Forbes. "Improved Soundness for QMA with Multiple Provers". In: *Chicago Journal of Theoretical Computer Science* (2013). DOI: 10.4086/cjtcs.2013.001 (page 2).

[CHTW04] Richard Cleve, Peter Høyer, Benjamin Toner, and John Watrous. "Consequences and Limits of Nonlocal Strategies". In: *Proceedings of the 19th IEEE Annual Conference on Computational Complexity*. 2004. DOI: 10.1109/CCC.2004.1313847 (page 3).

[CS12] Andre Chailloux and Or Sattath. "The Complexity of the Separable Hamiltonian Problem". In: *Proceedings of the 2012 IEEE Conference on Computational Complexity*. 2012. DOI: 10.1109/CCC.2012.42 (page 2).

[GNN12] François Le Gall, Shota Nakagawa, and Harumichi Nishimura. "On QMA protocols with two short quantum proofs". In: *Quantum Info. Comput.* (2012) (page 2).

[GS86] Shafi Goldwasser and Michael Sipser. "Private coins versus public coins in interactive proof systems". In: *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*. 1986. DOI: 10.1145/12130.12137 (pages 4, 5).

[HM13] Aram W. Harrow and Ashley Montanaro. "Testing Product States, Quantum Merlin-Arthur Games and Tensor Optimization". In: *Journal of the ACM* (2013). DOI: 10.1145/2432622.2432625 (page 2).

[Hor97] Pawel Horodecki. "Separability criterion and inseparable mixed states with positive partial transposition". In: *Physics Letters A* (1997). DOI: 10.1016/S0375-9601(97)00416-7 (page 8).

[HSR03] Michael Horodecki, Peter W. Shor, and Mary Beth Ruskai. "Entanglement Breaking Channels". In: *Reviews in Mathematical Physics* (2003). DOI: 10.1142/S0129055X03001709 (pages 2, 6, 8).

[IV12] Tsuyoshi Ito and Thomas Vidick. "A Multi-prover Interactive Proof for NEXP Sound against Entangled Provers". In: *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. 2012. DOI: 10.1109/FOCS.2012.11 (page 1).

[JJUW11] Rahul Jain, Zhengfeng Ji, Sarvagya Upadhyay, and John Watrous. "QIP = PSPACE". In: *Journal of the ACM (JACM)* (2011). DOI: 10.1145/2049697.2049704 (pages 2, 4, 6).

[JNVWY22] Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen. MIP* = RE. 2022. arXiv: 2001.04383 [quant-ph] (page 1).

[JW23] Fernando Granha Jeronimo and Pei Wu. "The Power of Unentangled Quantum Proofs with Non-negative Amplitudes". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023. DOI: 10.1145/3564246.3585248 (page 2).

[JW24] Fernando Granha Jeronimo and Pei Wu. "Dimension Independent Disentanglers from Unentanglement and Applications". In: *39th Computational Complexity Conference (CCC 2024)*. 2024. DOI: 10.4230/LIPIcs.CCC.2024.26 (page 2).

[KGN19] Hirotada Kobayashi, François Le Gall, and Harumichi Nishimura. "Generalized Quantum Arthur–Merlin Games". In: *SIAM Journal on Computing* (2019). DOI: 10.1137/17M1160173 (page 16).

[KMY01] Hirotada Kobayashi, Keiji Matsumoto, and Tomoyuki Yamakami. *Quantum Certificate Verification: Single versus Multiple Quantum Certificates*. 2001. arXiv: quant-ph/0110006 [quant-ph] (pages 1, 2).

[KW00] Alexei Kitaev and John Watrous. "Parallelization, amplification, and exponential time simulation of quantum interactive proof systems". In: *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*. 2000. DOI: 10.1145/335305.335387 (page 6).

[LG17] Joshua Lockhart and Carlos E. González Guillén. *Quantum State Isomorphism*. 2017. arXiv: 1709.09622 [quant-ph] (page 16).

[Mar03] Chris Marriott. "Non-determinism and quantum information". Master's thesis. University of Calgary, 2003. DOI: 10.11575/PRISM/14704 (pages 5, 16, 17).

[MW05] Chris Marriott and John Watrous. "Quantum Arthur–Merlin Games". In: *Computational Complexity* (2005). DOI: 10.1007/s00037-005-0194-x (pages 4, 12, 13, 15).

[Raz05] Ran Raz. "Quantum information and the PCP theorem". In: *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. 2005. DOI: 10.1109/SFCS.2005.62 (pages 3, 6, 11).

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il