Quantum Arthur–Merlin Games

Chris Marriott John Watrous

Department of Computer Science
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada T2N 1N4

June 15, 2005

Abstract

This paper studies quantum Arthur–Merlin games, which are Arthur–Merlin games in which Arthur and Merlin can perform quantum computations and Merlin can send Arthur quantum information. As in the classical case, messages from Arthur to Merlin are restricted to be strings of uniformly generated random bits. It is proved that for one-message quantum Arthur–Merlin games, which correspond to the complexity class QMA, completeness and soundness errors can be reduced exponentially without increasing the length of Merlin's message. Previous constructions for reducing error required a polynomial increase in the length of Merlin's message. Applications of this fact include a proof that logarithmic length quantum certificates yield no increase in power over BQP and a simple proof that QMA $\subseteq$ PP. Other facts that are proved include the equivalence of three (or more) message quantum Arthur–Merlin games with ordinary quantum interactive proof systems and some basic properties concerning two-message quantum Arthur–Merlin games.

1 Introduction

Interactive proof systems and Arthur–Merlin games were introduced by [GMR89] and [Bab85] (see also [BM88]) in order to model the notion of computationally efficient verification. In an interactive proof system, a polynomial-time verifier with a private source of uniformly generated random bits interacts with a computationally unbounded prover in an attempt to check the validity of the claim that a common input string is contained in some prespecified language. Arthur–Merlin games are similar in principle to interactive proof systems, but are somewhat more restricted—the verifier (called Arthur in this setting) no longer has a private source of randomness, but instead has only a public source of randomness that is visible to the prover (called Merlin). Because Arthur is deterministic aside from the bits produced by the random source, one may without loss of generality view that an Arthur–Merlin game is simply an interactive proof system in which the verifier's messages to the prover consist only of uniformly generated bits from the public random source.

Although Arthur–Merlin games are more restricted than interactive proof systems in the sense just described, the two models are known to be computationally equivalent. In particular, any language having an interactive proof system in which a constant number of messages is exchanged between the prover and verifier also has an Arthur–Merlin game in which precisely two messages are exchanged, the first from Arthur to Merlin and the second from Merlin back to Arthur [GS89, BM88]. The complexity class consisting of all such languages is AM. Also following from

[GS89] is the fact that any language having an unrestricted (polynomial-message) interactive proof system also has a polynomial-message Arthur–Merlin game. The complexity class consisting of all such languages was initially called IP, but is now known to be equal to PSPACE [LFKN92, Sha92].

A third complexity class arising from these models is MA, which is the class consisting of all languages having an interactive proof system in which a single message is sent, from the prover to the verifier. One may view the definition of this class as a slight variation on the "guess and check" definition of NP, where instead of being deterministic the checking procedure may use randomness. As the usual convention for Arthur–Merlin games is to disallow Arthur the use of the public random source except for the generation of messages, the class MA would typically be described as consisting of all languages having two-message Arthur–Merlin games in which the first message is sent from Merlin to Arthur and the second from Arthur to Merlin. However, given that the information transmitted to Merlin in the second message is irrelevant from the point of view of the game, and may instead be viewed as just a use of the random source and not as a message, it is natural to refer to such games as one-message Arthur–Merlin games.

Quantum computational variants of interactive proof systems have previously been considered in several papers, including the general multiple-message case [Wat03, KW00, KM03, RW04, GW05] as well as the single-message case [AR03, JWB03, KR03, KKR04, KMY03, RS04, Vya03, Wat00]. As for classical interactive proof systems, quantum interactive proof systems consist of two parties—a prover with unlimited computation power and a computationally bounded verifier. Now, however, the two parties may process and exchange quantum information. The complexity class consisting of all languages having quantum interactive proof systems is denoted QIP, and satisfies PSPACE $\subseteq$ QIP $\subseteq$ EXP [KW00]. Here, EXP denotes the class of languages decidable by a deterministic Turing machine running in time 2^q for some polynomial q.

There are both similarities and some apparent differences in the properties of quantum and classical interactive proof systems. Perhaps the most significant difference is that any language having an unrestricted (polynomial-message) quantum interactive proof system also has a three-message quantum interactive proof system [KW00]. This cannot happen classically unless AM $=$ PSPACE.

This paper investigates various aspects of quantum Arthur–Merlin games. In analogy to the classical case, we define quantum Arthur–Merlin games to be restricted forms of quantum interactive proof systems in which the verifier's (Arthur's) messages to the prover (Merlin) are uniformly generated random bits, as opposed to arbitrary messages. Consequently, Arthur is not capable of sending quantum information to Merlin at any point during a quantum Arthur–Merlin game. Similar to the classical case, quantum Arthur–Merlin games give rise to complexity classes depending on the number of messages exchanged between Arthur and Merlin. In particular, we obtain three primary complexity classes corresponding to Arthur–Merlin games with one message, two messages, and three or more messages.

In the one-message case, Merlin sends a single message to Arthur, who checks it and makes a decision to accept or reject the input. The corresponding complexity class is denoted QMA, and has been considered previously in the papers cited above. In this situation Merlin's message to Arthur may simply be viewed as a quantum witness or certificate that Arthur checks in polynomial time with a quantum computer. To our knowledge, the idea of a quantum state playing the role of a certificate in this sense was first proposed by [Kni96], and the idea was later studied in greater depth by [Kit99]. Kitaev proved various fundamental properties of QMA, which are described in [KSV02] and [AN02].

One of the facts that Kitaev proved was that the completeness and soundness errors in a QMA protocol may be efficiently reduced by parallel repetition. Because quantum information cannot be copied, however, and Arthur's verification procedure is potentially destructive to Merlin's message, Arthur requires multiple copies of Merlin's message for this method to work. This method therefore requires a polynomial increase in the length of Merlin's message to Arthur in order to achieve exponentially decreasing error. In this paper, we prove that this increase in the length of Merlin's message is not required after all—using a different error reduction method, an exponential reduction in error is possible with no increase whatsoever in the length of Merlin's message to Arthur.

It is known that QMA is contained in the class PP, which can be proved using the GapP-based method of [FR99] together with some simple facts from matrix analysis. This fact was noted without proof in [KW00]. A proof of this fact was, however, given by [Vya03], who in fact strengthened this result to show that QMA is contained in a subclass $\mathrm{A_0PP}$ of PP. (Definitions of the classes PP and $\mathrm{A_0PP}$ can be found in Section 2 of this paper.) Based on our new error reduction method, we give a simplified proof of this containment. We also use our error reduction method to prove that one-message quantum Arthur–Merlin games in which Merlin's message has logarithmic length give no increase in power over BQP.

In the two-message case, Arthur flips some number of fair coins, sends the results of those coin-flips to Merlin, and Merlin responds with some quantum state. Arthur performs a polynomial-time quantum computation on the random bits together with Merlin's response, which determines whether Arthur accepts or rejects. The corresponding complexity class will be denoted QAM. Two facts about QAM are proved in this paper. The first is the very basic fact that parallel repetition reduces error exactly as in the classical case. (This fact does not follow from known facts about quantum interactive proof systems, as parallel repetition is only known to reduce error for general quantum interactive proof systems having perfect completeness.) The second fact is that QAM is contained in $\mathrm{BP} \cdot \mathrm{PP}$, the class obtained by applying the BP operator to the class PP.

Finally, in the three-message case, Merlin sends Arthur a message consisting of some number of qubits, Arthur flips some number of fair coins and sends the results to Merlin, and then Merlin responds with a second collection of qubits. Arthur performs a polynomial-time quantum computation on all of the qubits sent by Merlin together with the values of his own coin-flips, and decides whether to accept or reject. The corresponding complexity class will be denoted QMAM. It is proved that any language having an ordinary quantum interactive proof system is contained in QMAM, implying $\mathrm{QMAM} = \mathrm{QIP}$.

In spirit, the equality $\mathrm{QMAM} = \mathrm{QIP}$ resembles the theorem of [GS89] establishing that classical Arthur–Merlin games and interactive proof systems are equivalent in power. However, there is no similarity in the proofs of these facts. Moreover, our result is stronger than what is likely to hold classically. Specifically, we prove that any language having a quantum interactive proof system also has a three-message quantum Arthur–Merlin game in which Arthur's only message to Merlin consists of just a single coin-flip (in order to achieve perfect completeness and soundness error exponentially close to $1/2$). This is impossible classically unless interaction is useless in classical interactive proof systems; for if Arthur flips only one coin, Merlin may as well send his first message and the two possible second messages to Arthur in a single message. The reason why this strategy fails in the quantum case is that Merlin's first and second messages may need to be entangled in order to be convincing to Arthur, but it may not possible for Merlin to simultaneously entangle his two possible second messages with the first in a way that convinces Arthur to accept. This is an example of the principle that Bennett refers to as the "monogamy of entanglement" (see, for example, [Ter04]); the more a given system is entangled with a second system, the less it can be entangled with a third.

Organization of the paper

The remainder of this paper is organized as follows. We begin with Section 2, which discusses background information needed elsewhere in the paper, including a summary of basic notation and conventions that are used, definitions of some relevant counting complexity classes, and background on quantum computation and quantum interactive proof systems. The next three sections correspond to the three complexity classes QMA, QAM, and QMAM, respectively; Section 3 discusses one-message quantum Arthur–Merlin games, Section 4 discusses the two-message case, and Section 5 discusses the case of three or more messages. The paper concludes with Section 6, which mentions some open problems relating to quantum Arthur–Merlin games.

2 Background Information

This section summarizes various background information that is needed for the remainder of the paper, including information on quantum computation, counting complexity, and quantum interactive proof systems.

We begin with some remarks about notation and other simple conventions that are followed throughout. All strings and languages in this paper will be over the alphabet $\Sigma = \{0,1\}$. We denote by *poly* the set of all functions $f : \mathbb{N} \rightarrow \mathbb{N}\backslash\{0\}$ (where $\mathbb{N} = \{0,1,2,\ldots\}$) for which there exists a polynomial-time deterministic Turing machine that outputs $1^{f(n)}$ on input 1^n. For every integer $k \geq 2$, we fix a polynomial-time computable function that, for every choice of $x_1,\ldots,x_k \in \Sigma^*$, encodes the k-tuple $(x_1,\ldots,x_k)$ as a single element of Σ^*. These functions are assumed to satisfy the usual properties of tuple-functions, namely that they are one-to-one and polynomial-time invertible in each argument. As is typical, reference to these functions is often implicit; for instance, we write $f(x_1,\ldots,x_k)$ as shorthand for $f((x_1,\ldots,x_k))$ when $x_1,\ldots,x_k \in \Sigma^*$ and the domain of the function f is understood to be Σ^*.

Quantum computation

We will assume that the reader has familiarity with the mathematics of quantum information, which is discussed in the books of [KSV02] and [NC00]. The quantum complexity classes discussed in this paper are based on the quantum circuit model, with which we also assume familiarity.

All quantum circuits considered in this paper will be assumed to be composed only of Toffoli gates, Hadamard gates, and i-shift gates (which induce the mapping $|0\rangle \mapsto |0\rangle$, $|1\rangle \mapsto i\,|1\rangle$). This is a universal set of gates [Kit97], so there is no loss of generality in restricting our attention to this set. We assume that a reasonable encoding scheme has been fixed that allows quantum circuits to be encoded as binary strings having length at least the size of the encoded circuit and at most some fixed polynomial in the circuit's size.

A collection $\{A_x : x \in \Sigma^*\}$ of quantum circuits is said to be *generated in polynomial-time* if there exists a polynomial-time deterministic Turing machine that, on input $x \in \Sigma^*$, outputs an encoding of the circuit A_x. When such a family is parameterized by tuples of strings, it is to be understood that we are implicitly referring to one of the tuple-functions discussed previously. For instance, we will consider families of the form $\{A_{x,y} : x, y \in \Sigma^*\}$ when two- and three-message quantum Arthur–Merlin games are discussed.

Polynomial-time generated families are similar to the usual notion of a polynomial-time uniform family of circuits, except that the procedure generating the circuits are permitted to have access to the input x, rather than just the length of x written in unary. In essence, the input x

may be "hard-coded" into a given circuit in a polynomial-time generated family, so that it is not necessary to assume that the input x is input to the circuit itself. This is simply done as a matter of convenience and simplicity—all of the polynomial-time generated families of quantum circuits in this paper could be replaced by polynomial-time uniform families where the string given to the generating procedure is instead input directly into the circuit.

Let us illustrate the use of polynomial-time generated families of quantum circuits by defining BQP, the class of languages recognizable in quantum polynomial time with bounded error. A language L is in BQP if and only if there exists a polynomial-time generated family $\{A_x\}$ of quantum circuits such that the following conditions hold. First, it is required that there exist a function $k \in \mathit{poly}$ such that each circuit A_x act on precisely $k(|x|)$ qubits. (This condition is not really necessary, but will simplify further discussions.) Let $\Pi_1 = |1\rangle\langle 1| \otimes I_{k-1}$, where k is shorthand for $k(|x|)$ and, in general, I_n denotes the identity operator acting on n qubits. Then it is required that

1. if $x \in L$ then $\left\|\Pi_1 A_x |0^k\rangle\right\|^2 \geq \frac{2}{3}$, and
2. if $x \notin L$ then $\left\|\Pi_1 A_x |0^k\rangle\right\|^2 \leq \frac{1}{3}$.

In words, if the input is x, then the circuit A_x is run on the all-zero input and the first qubit is measured in the standard basis. If the measurement result is 1, the computation is viewed as accepting, otherwise it is rejecting. The usual notion of bounded error is required.

It will sometimes be helpful when describing certain quantum Arthur–Merlin games to refer to *quantum registers*. These are simply collections of qubits to which we assign some name. When we refer to the *reduced state* of a given register, we mean the mixed state obtained by tracing out all other registers beside the one to which we are referring.

Counting classes

Some of the results in this paper involve relations between complexity classes based on quantum Arthur–Merlin games and classes based on the notion of counting complexity. Here we briefly discuss this notion and the classes relevant to this paper; for more information about counting complexity, see [For97].

A function $f : \Sigma^* \to \mathbb{N}$ is an element of the function class #P if and only if there exists a polynomial-time nondeterministic Turing machine that, on each input $x \in \Sigma^*$, has precisely $f(x)$ accepting computation paths. For any function $f \in$ #P there exists a function $q \in \mathit{poly}$ such that $f(x) \leq 2^{q(|x|)}$ for all $x \in \Sigma^*$.

A function $f : \Sigma^* \to \mathbb{Z}$ is an element of the function class FP if it is computable in polynomial time, with the understanding that the output of the function is the integer represented in binary notation by the output of the computation.

A function $f : \Sigma^* \to \mathbb{Z}$ is an element of the function class GapP if and only if there exist functions $g, h \in$ #P such that $f(x) = g(x) - h(x)$ for all $x \in \Sigma^*$. The function class GapP possesses remarkable closure properties, including closure under subtraction, exponential sums, and polynomial products. In particular, if $f \in$ GapP and $q \in \mathit{poly}$, then the functions g and h defined as

$$g(x) = \sum_{i=1}^{2^{q(|x|)}} f(x,i), \quad h(x) = \prod_{i=1}^{q(|x|)} f(x,i)$$

are elements of GapP. (Here the integer i is identified with the string having no leading zeroes that encodes it in binary notation.) It is not difficult to show that FP $\subseteq$ GapP.

The complexity class PP consists of all languages $L \subseteq \Sigma^*$ for which there exists a function $f \in \mathrm{GapP}$ such that $x \in L$ if and only if $f(x) > 0$ for all $x \in \Sigma^*$. The class $\mathrm{A_0PP}$ consists of all languages $L \subseteq \Sigma^*$ for which there exist functions $f \in \mathrm{GapP}$ and $g \in \mathrm{FP}$ satisfying

$$x \in L \;\Rightarrow\; f(x) \geq g(x), \quad x \notin L \;\Rightarrow\; 0 \leq f(x) \leq \frac{g(x)}{2},$$

for all $x \in \Sigma^*$. Finally, the complexity class $\mathrm{BP} \cdot \mathrm{PP}$ refers to the BP operator applied to the class PP; it contains all languages $L \subseteq \Sigma^*$ such that there exists a language $A \in \mathrm{PP}$ and a function $q \in \mathit{poly}$ such that

$$\left| \left\{ y \in \Sigma^{q(|x|)} \;:\; (x,y) \in A \;\Leftrightarrow\; x \in L \right\} \right| \geq \frac{2}{3} 2^{q(|x|)}.$$

Counting complexity and quantum complexity were related by [FR99], who gave a simple proof that $\mathrm{BQP} \subseteq \mathrm{PP}$ based on the closure properties of GapP functions discussed above. (The containment $\mathrm{BQP} \subseteq \mathrm{PP}$ had been proved earlier by [ADH97] using a different method.) In fact, Fortnow & Rogers proved the stronger containment $\mathrm{BQP} \subseteq \mathrm{AWPP}$, where AWPP is a subclass of PP that we will not define in this paper. As a couple of the facts we prove are based on the method of Fortnow & Rogers, it will be helpful for us to summarize this method. The quantum Turing machine model was used in the original proof, but our summary is instead based on polynomial-time generated families of quantum circuits.

Suppose that $L \in \mathrm{BQP}$, which implies the existence of a polynomial-time generated family $\{A_x\}$ of quantum circuits satisfying the conditions of the definition of BQP discussed previously. The goal is to construct a GapP function f and a polynomially bounded FP function g such that

$$\frac{f(x)}{2^{g(x)}} = \langle 0^k | \, A_x^\dagger \Pi_1 A_x \, |0^k\rangle = \left\| \Pi_1 A_x \, |0^k\rangle \right\|^2.$$

Once this is done, the GapP function $h(x) = 2f(x) - 2^{g(x)}$ satisfies the required property to establish $L \in \mathrm{PP}$; namely that $h(x) > 0$ if and only if $x \in L$.

The functions f and g are of course based on the circuit family $\{A_x\}$. For a given string x, assume that the circuit A_x consists of gates $G_1, \ldots, G_{q(|x|)}$ for some function $q \in \mathit{poly}$. Each of the gates G_j, when tensored with the identity operator on the qubits not affected by G_j, gives rise to a $2^k \times 2^k$ matrix whose individual entries, indexed by pairs of strings of length k, can be computed in polynomial time given x. These entries are elements of the set

$$\left\{ 0, 1, i, 1/\sqrt{2}, -1/\sqrt{2} \right\}$$

because we assume A_x is composed only of Toffoli, Hadamard, and i-shift gates. Similarly, Π_1 is a $2^k \times 2^k$ matrix whose entries (this time restricted to the set $\{0,1\}$) are also computable in polynomial time given x.

The value $\langle 0^k | \, A_x^\dagger \Pi_1 A_x \, |0^k\rangle$ therefore corresponds to the $(0^k, 0^k)$ entry of the matrix product

$$G_1^\dagger \cdots G_q^\dagger \, \Pi_1 \, G_q \cdots G_1,$$

which can be expressed as an exponential sum of a polynomial product of the entries of these matrices. By letting the function g represents the total number of Hadamard transforms in the circuit A_x, it is fairly straightforward to construct an appropriate GapP function f based on closure properties of the class GapP. Further details can be found in [FR99] as well as in [Vya03].

Quantum interactive proofs

Here we discuss background information on quantum interactive proof systems that will be used later in the paper when it is proved that quantum Arthur–Merlin games have the same power as arbitrary quantum interactive proof systems. It will only be necessary for us to discuss the particular case of three-message quantum interactive proof systems, as any polynomial-message quantum interactive proof system can be simulated by a three-message quantum interactive proof. Moreover, such a proof system may be taken to have perfect completeness and exponentially small soundness error. These facts are proved in [KW00], to which the reader is referred for a more complete discussion of quantum interactive proof systems.

For a fixed input x, a three-message quantum interactive proof system operates as follows. The verifier begins with a k-qubit register V and the prover begins with two registers: an m-qubit register M and an l-qubit register P. The register V corresponds to the verifier's work-space, the register M corresponds to the message qubits that are sent back and forth between the prover and verifier, and the register P corresponds to the prover's workspace. The register M begins in the prover's possession because the prover sends the first message. The verifier's work-space register V begins initialized to the state $|0^k\rangle$, while the prover initializes the pair (M,P) to some arbitrary quantum state $|\psi\rangle$.

In the first message, the prover sends M to the verifier. The verifier applies some unitary transformation V_1 to the pair (V,M) and returns M to the prover in the second message. The prover now applies some arbitrary unitary transformation U to the pair (M,P) and returns M to the verifier in the third and final message. Finally, the verifier applies a second unitary transformation V_2 to the pair (V,M) and measures the first qubit of the resulting collection of qubits in the standard basis. The outcome 1 is interpreted as "accept" and 0 is interpreted as "reject".

Let Π_0, Π_1, Δ_0, and Δ_1 be projections defined as

$$\Pi_1 = |1\rangle\langle 1| \otimes I_{k+m-1}, \quad \Delta_1 = |0^k\rangle\langle 0^k| \otimes I_m, \quad \Pi_0 = |0\rangle\langle 0| \otimes I_{k+m-1}, \quad \Delta_0 = I_{k+m} - \Delta_1.$$

In other words, these are $k+m$ qubit projections that act on the pair of registers (V,M); Π_1 and Π_0 are projections onto those states for which the first qubit of the register V is 1 or 0, respectively, and Δ_1 and Δ_0 are projections onto those states for which the register V contains the state $|0^k\rangle$ or contains a state orthogonal to $|0^k\rangle$, respectively.

The maximum probability with which a verifier specified by V_1 and V_2 can be made to accept is

$$\left\| (\Pi_1 V_2 \otimes I_l)(I_k \otimes U)(V_1 \otimes I_l)(|0^k\rangle\,|\psi\rangle) \right\|^2, \tag{1}$$

maximized over all choices of the state $|\psi\rangle$ and the unitary transformation U. The number l is determined by the prover's strategy, so one may maximize over this number as well. However, there is no loss of generality in assuming $l = m+k$; with this many work qubits, the prover may store a purification of the reduced state of the pair (V,M), which is sufficient for an optimal strategy.

There is another way to characterize the maximum acceptance probability for a given verifier based on the fidelity function

$$F(\rho,\xi) = \operatorname{tr}\sqrt{\sqrt{\rho}\,\xi\sqrt{\rho}}.$$

To describe this characterization we will need to define various sets of states of the pair of registers (V,M). For any projection Λ on $k+m$ qubits let $\mathcal{S}(\Lambda)$ denote the set of all mixed states ρ of (V,M) that satisfy $\rho = \Lambda\rho\Lambda$, i.e., the collection of states whose support is contained in the space onto which Λ projects. Also let $\mathcal{S}_{\mathsf{V}}(\Lambda)$ denote the set of all reduced states of V that result from some state $\rho \in \mathcal{S}(\Lambda)$, i.e.,

$$\mathcal{S}_{\mathsf{V}}(\Lambda) = \{\operatorname{tr}_{\mathsf{M}} \rho \,:\, \rho \in \mathcal{S}(\Lambda)\},$$

where $\operatorname{tr}_{\mathsf{M}}$ denotes the partial trace over the register M.

Proposition 1. *The maximum probability with which a verifier specified by V_1 and V_2 can be made to accept is*

$$\max\left\{F(\rho,\xi)^2 \,:\, \rho \in \mathcal{S}_{\mathsf{V}}(V_1\Delta_1 V_1^{\dagger}),\ \xi \in \mathcal{S}_{\mathsf{V}}(V_2^{\dagger}\Pi_1 V_2)\right\}.$$

This proposition is essentially a restatement based on Uhlmann's Theorem (see [NC00]), of the fact that the quantity 1 above represents the maximum acceptance probability of the verifier described by V_1 and V_2. This equivalence is discussed further in [KW00].

3 QMA

A QMA verification procedure A is a family of quantum circuits $\{A_x : x \in \Sigma^*\}$ that is generated in polynomial time, together with a function $m \in poly$. The function m specifies the length of Merlin's message to Arthur, and it is assumed that each circuit A_x acts on $m(|x|) + k(|x|)$ qubits for some function k specifying the number of work qubits used by the circuit. As we have done in the previous section, when the input x has been fixed or is implicit we will generally write m to mean $m(|x|)$, k to mean $k(|x|)$, and so forth, in order to simplify our notation. When we want to emphasize the length of Merlin's message, we will refer to A as an m-qubit QMA verification procedure.

Consider the following process for a string $x \in \Sigma^*$ and a quantum state $|\psi\rangle$ on m qubits:

1. Run the circuit A_x on the input state $|\psi\rangle\,|0^k\rangle$.
2. Measure the first qubit of the resulting state in the standard basis, interpreting the outcome 1 as *accept* and the outcome 0 as *reject*.

The probability associated with the two possible outcomes will be referred to as $\Pr[A_x \text{ accepts } |\psi\rangle]$ and $\Pr[A_x \text{ rejects } |\psi\rangle]$ accordingly.

Definition 2. The class $\mathrm{QMA}(a,b)$ consists of all languages $L \subseteq \Sigma^*$ for which there exists a QMA verification procedure A for which the following holds:

1. For all $x \in L$ there exists an m qubit quantum state $|\psi\rangle$ such that $\Pr[A_x \text{ accepts } |\psi\rangle] \geq a$.
2. For all $x \notin L$ and all m qubit quantum states $|\psi\rangle$, $\Pr[A_x \text{ accepts } |\psi\rangle] \leq b$.

For any $m \in poly$, the class $\mathrm{QMA}_m(a,b)$ consists of all languages $L \subseteq \Sigma^*$ for which there exists an m-qubit QMA verification procedure that satisfies the above properties.

One may consider the cases where a and b are constants or functions of the input length $n = |x|$ in this definition. If a and b are functions of the input length, it is assumed that $a(n)$ and $b(n)$ can be computed deterministically in time polynomial in n. When no reference is made to the probabilities a and b, it is assumed $a = 2/3$ and $b = 1/3$.

Strong Error Reduction

It is known that QMA is robust with respect to error bounds in the following sense.

Theorem 3 (Kitaev)**.** *Let $a, b : \mathbb{N} \rightarrow [0,1]$ and $q \in poly$ satisfy*

$$a(n) - b(n) \geq \frac{1}{q(n)}$$

for all $n \in \mathbb{N}$. Then $\mathrm{QMA}(a,b) \subseteq \mathrm{QMA}(1-2^{-r}, 2^{-r})$ for every $r \in poly$.

A proof of this theorem appears in Section 14.2 of [KSV02]. The idea of the proof is as follows. If we have a verification procedure A with completeness and soundness probabilities given by a and b, we construct a new verification procedure that independently runs A on some sufficiently large number of copies of the original certificate and accepts if the number of acceptances of A is larger than $(a+b)/2$. The only difficulty in proving that this construction works lies in the fact that the new certificate cannot be assumed to consist of several copies of the original certificate, but may be an arbitrary (possibly highly entangled) quantum state. Intuitively, however, entanglement cannot help Merlin to cheat; under the assumption that $x \notin L$, the probability of acceptance for any particular execution of A is bounded above by b, and this is true regardless of whether one conditions on the outcomes of any of the other executions of A. This construction requires an increase in the length of Merlin's message to Arthur in order to reduce error.

The main result of this section is the following theorem, which states that one may decrease error without any increase in the length of Merlin's message.

Theorem 4. *Let $a, b : \mathbb{N} \rightarrow [0,1]$ and $q \in poly$ satisfy*

$$a(n) - b(n) \geq \frac{1}{q(n)}$$

for all $n \in \mathbb{N}$. Then $\mathrm{QMA}_m(a,b) \subseteq \mathrm{QMA}_m(1-2^{-r}, 2^{-r})$ for every $m, r \in poly$.

Proof. Assume $L \in \mathrm{QMA}_m(a,b)$, and A is an m-qubit QMA verification procedure that witnesses this fact. We will describe a new m-qubit QMA verification procedure B with exponentially small completeness and soundness error for the language L, which will suffice to prove the theorem.

It will simplify matters to assume hereafter that the input x is fixed—it will be clear that the new verification procedure can be generated in polynomial-time. As the input x is fixed, we will write A and B to denote A_x and B_x, respectively.

It will be helpful to refer to the m message qubits along with the k work-space qubits of A as a single $m+k$ qubit quantum register R. Define projections acting on the vector space corresponding to R as follows:

$$\Pi_1 = |1\rangle\langle 1| \otimes I_{m+k-1}, \quad \Delta_1 = I_m \otimes |0^k\rangle\langle 0^k|, \quad \Pi_0 = |0\rangle\langle 0| \otimes I_{m+k-1}, \quad \Delta_0 = I_{m+k} - \Delta_1. \qquad (2)$$

The measurement described by $\{\Pi_0, \Pi_1\}$ is just a measurement of the first qubit of R in the computational basis; this measurement determines whether Arthur accepts or rejects after the circuit A is applied. The measurement described by $\{\Delta_0, \Delta_1\}$ gives outcome 1 if the last k qubits of R, which correspond to Arthur's work-space qubits, are set to their initial all-zero state, and gives outcome 0 otherwise. (These projections are similar to those in Section 2 except that the message qubits and Arthur's work qubits are reversed for notational convenience.)

The procedure B operates as follows. It assumes that initially the first m qubits of R contain Merlin's message $|\psi\rangle$ and the remaining k qubits are set to the state $|0^k\rangle$.

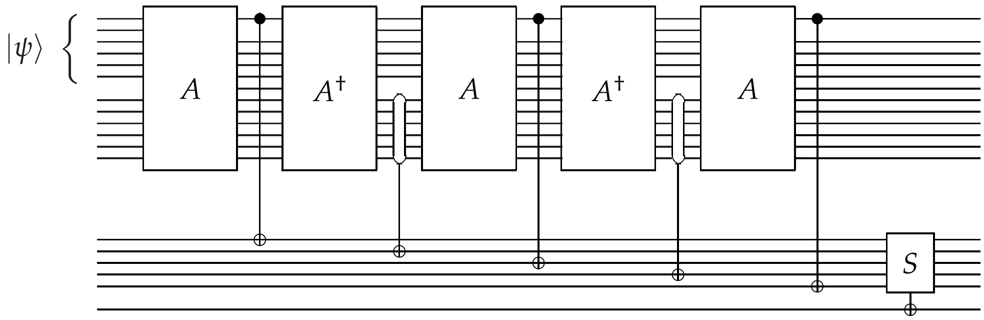

Figure 1: Example circuit diagram for verification procedure B.

1. Set $y_0 \leftarrow 1$ and $i \leftarrow 1$.
2. Repeat:

   a. Apply A to R and measure R with respect to the measurement described by $\{\Pi_0, \Pi_1\}$. Let y_i denote the outcome, and set $i \leftarrow i+1$.

   b. Apply $A^\dagger$ to R and measure R with respect to the measurement described by $\{\Delta_0, \Delta_1\}$. Let y_i denote the outcome, and set $i \leftarrow i+1$.

   Until $i \geq N$, where $N = 8\,q^2 r$.

3. For each $i = 1, \ldots, N$ set

$$z_i \leftarrow \begin{cases} 1 & \text{if } y_i = y_{i-1} \\ 0 & \text{if } y_i \neq y_{i-1}. \end{cases}$$

   Accept if $\sum_{i=1}^{N} z_i \geq N \cdot \frac{a+b}{2}$ and reject otherwise.

Although the description of this procedure refers to various measurements, it is possible to simulate these measurements with unitary gates in the standard way, which allows the entire procedure to be implemented by a unitary quantum circuit. 1 illustrates a quantum circuit implementing this procedure for the case $N = 5$. In this figure, S represents the computation described in the last step of B, and the last qubit rather than the first represents the output qubit to simplify the figure.

We first consider the behavior of the verification procedure B in the situation that the state $|\psi\rangle$ is an eigenvector of the operator

$$Q = (I_m \otimes \langle 0^k|) A^\dagger \Pi_1 A (I_m \otimes |0^k\rangle),$$

with corresponding eigenvalue p. We have

$$p = \langle \psi | \, Q \, |\psi\rangle = \left\| \Pi_1 A(|\psi\rangle \, |0^k\rangle) \right\|^2,$$

and thus p is the probability that the verification procedure A accepts $|\psi\rangle$. Let $|\phi\rangle = |\psi\rangle \, |0^k\rangle$, which implies that $|\phi\rangle$ is an eigenvector of $\Delta_1 A^\dagger \Pi_1 A \Delta_1$, also having corresponding eigenvalue p. We will show that the verification procedure B accepts $|\psi\rangle$ with probability

$$\sum_{N \cdot \frac{a+b}{2} \leq j \leq N} \binom{N}{j} p^j (1-p)^{N-j}. \tag{3}$$

Using standard Chernoff-type bounds, this probability can be shown to be greater than $1 - 2^{-r}$ when $p \geq a$ and less than 2^{-r} when $p \leq b$, given the choice of $N = 8q^2 r$.

The fact that $|\psi\rangle$ is accepted with the probability given in equation 3 will follow from the fact that the procedure B obtains each possible sequence $(z_1, \ldots, z_N)$ with probability

$$p^{w(z)}(1-p)^{N-w(z)}$$

for $w(z) = \sum_{i=1}^{N} z_i$. This is straightforward if $p = 0$ or $p = 1$, so assume $0 < p < 1$.

Define vectors $|\gamma_0\rangle$, $|\gamma_1\rangle$, $|\delta_0\rangle$, and $|\delta_1\rangle$ as follows:

$$|\gamma_0\rangle = \frac{\Pi_0 A \Delta_1 |\phi\rangle}{\sqrt{1-p}}, \quad |\gamma_1\rangle = \frac{\Pi_1 A \Delta_1 |\phi\rangle}{\sqrt{p}}, \quad |\delta_0\rangle = \frac{\Delta_0 A^\dagger \Pi_1 |\gamma_1\rangle}{\sqrt{1-p}}, \quad |\delta_1\rangle = \frac{\Delta_1 A^\dagger \Pi_1 |\gamma_1\rangle}{\sqrt{p}}.$$

As $\Delta_1 A^\dagger \Pi_1 A \Delta_1 |\phi\rangle = p\, |\phi\rangle$ and $|\phi\rangle$ is a unit vector we have

$$\begin{aligned} \langle\phi| \Delta_1 A^\dagger \Pi_1 A \Delta_1 |\phi\rangle &= p, \\ \langle\phi| \Delta_1 A^\dagger \Pi_0 A \Delta_1 |\phi\rangle &= \langle\phi| \Delta_1 A^\dagger (I - \Pi_1) A \Delta_1 |\phi\rangle = 1 - p, \end{aligned}$$

and thus $|\gamma_0\rangle$ and $|\gamma_1\rangle$ are unit vectors. Moreover, as

$$\Pi_1 A \Delta_1 A^\dagger \Pi_1 |\gamma_1\rangle = \frac{\Pi_1 A \Delta_1 \left(\Delta_1 A^\dagger \Pi_1 A \Delta_1\right) |\phi\rangle}{\sqrt{p}} = p\, |\gamma_1\rangle,$$

we have that $|\delta_0\rangle$ and $|\delta_1\rangle$ are unit vectors by similar reasoning. Note also that $|\delta_1\rangle = |\phi\rangle$, which follows immediately from the fact that $|\phi\rangle$ is an eigenvector of $\Delta_1 A^\dagger \Pi_1 A \Delta_1$ with eigenvalue p. Based on these observations we conclude that

$$\begin{aligned} A\, |\delta_0\rangle &= -\sqrt{p}\, |\gamma_0\rangle + \sqrt{1-p}\, |\gamma_1\rangle \\ A\, |\delta_1\rangle &= \sqrt{1-p}\, |\gamma_0\rangle + \sqrt{p}\, |\gamma_1\rangle. \end{aligned} \tag{4}$$

It will also be helpful to note that

$$\begin{aligned} A^\dagger\, |\gamma_0\rangle &= -\sqrt{p}\, |\delta_0\rangle + \sqrt{1-p}\, |\delta_1\rangle \\ A^\dagger\, |\gamma_1\rangle &= \sqrt{1-p}\, |\delta_0\rangle + \sqrt{p}\, |\delta_1\rangle \end{aligned} \tag{5}$$

which follows from the equations 4 along with the fact that A is unitary.

With the above equations 4 and 5 in hand, it is now possible to calculate the probability associated with each sequence of measurement outcomes. The procedure B begins in state $|\phi\rangle = |\delta_1\rangle$, and the procedure A is performed. After the measurement described by $\{\Pi_0, \Pi_1\}$ the (renormalized) state of register R becomes $|\gamma_0\rangle$ or $|\gamma_1\rangle$ according to whether the outcome is 0 or 1, with associated probabilities $1 - p$ and p, respectively. If instead the procedure B were to start in state $|\delta_0\rangle$, the renormalized states after measurement would be the same, but the probabilities would be reversed; probability p is associated with outcome 0 and probability $1 - p$ with outcome 1. For the second step of the loop the situation is similar. If the register R is in state $|\gamma_1\rangle$, the transformation $A^\dagger$ is applied, and the state is measured with respect to the measurement $\{\Delta_0, \Delta_1\}$, the renormalized state after measurement will be either $|\delta_1\rangle$ or $|\delta_0\rangle$, with associated probabilities p and $1 - p$. If instead the initial state were $|\gamma_0\rangle$ rather than $|\gamma_1\rangle$, the renormalized states after the measurement would again be the same, but the probabilities would be reversed. These transition probabilities are illustrated in 2. In all cases we see that the probability of obtaining the same outcome as for the

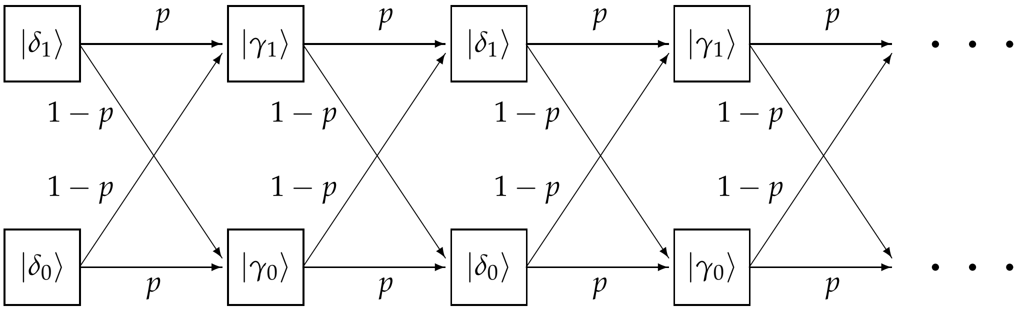

Figure 2: Transition probabilities for verification procedure B.

previous measurement is p, and the probability of the opposite outcome is $1-p$. The probability associated with a given sequence $z = (z_1, \ldots, z_N)$ is therefore $p^{w(z)}(1-p)^{N-w(z)}$ as claimed, as each z_i is 1 if the measurement outcomes y_{i-1} and y_i are equal, and is 0 otherwise. (Setting $y_0 = 1$ includes the first measurement outcome in this pattern.)

At this point we are ready to consider the completeness and soundness properties of the procedure B. Suppose first that the input x is in L, which implies that the procedure A can be made to accept with probability at least a. As an arbitrary state $|\psi\rangle$ is accepted by A with probability $\langle\psi| Q |\psi\rangle$, we therefore have $\langle\psi| Q |\psi\rangle \geq a$ for some choice of $|\psi\rangle$. Because Q is positive semidefinite it is the case that $\langle\psi| Q |\psi\rangle$ is bounded above by the largest eigenvalue of Q. Consequently there must exist a unit eigenvector $|\psi\rangle$ of Q having associated eigenvalue $p \geq a$. The procedure B has been shown to accept such a choice of $|\psi\rangle$ with probability at least $1 - 2^{-r}$ as required.

Now let us consider the soundness of the procedure B. If the input x is not contained in L, then every choice for the state $|\psi\rangle$ causes A to accept with probability at most b. Therefore, every eigenvalue of the operator Q is at most b. We have shown that if $|\psi\rangle$ is an eigenvector of Q, then the procedure B will accept $|\psi\rangle$ with probability less than 2^{-r}. Unfortunately, we may not assume that Merlin chooses $|\psi\rangle$ to be an eigenvector of Q. Nevertheless, the previous analysis can be extended to handle this possibility.

Specifically, let $\{|\psi_1\rangle, \ldots, |\psi_{2^m}\rangle\}$ be a complete orthonormal collection of eigenvectors of Q, with p_j denoting the eigenvalue corresponding to $|\psi_j\rangle$ for $j = 1, \ldots, 2^m$. An arbitrary unit vector $|\psi\rangle$ may be written as

$$|\psi\rangle = \sum_{j=1}^{2^m} \alpha_j |\psi_j\rangle$$

for $\alpha_1, \ldots, \alpha_{2^m} \in \mathbb{C}$ satisfying $\sum_j |\alpha_j|^2 = 1$. Given such a state $|\psi\rangle$ as input, the procedure B obtains each sequence $z = (z_1, \ldots, z_N)$ with probability

$$\sum_{j=1}^{2^m} |\alpha_j|^2 p_j^{w(z)} (1-p_j)^{N-w(z)}$$

and so the probability of acceptance is

$$\sum_{j=1}^{2^m} |\alpha_j|^2 \sum_{N\cdot\frac{a+b}{2} \leq i \leq N} \binom{N}{i} p_j^i (1-p_j)^{N-i} < 2^{-r}.$$

This does not follow from linearity because measurements are nonlinear. Instead, to see that it is indeed the case, one may repeat the analysis given previously in somewhat more generality. Specifically, let $|\phi_j\rangle = |\psi_j\rangle\,|0^k\rangle$ and

$$|\gamma_{j,0}\rangle = \frac{\Pi_0 A \Delta_1 |\phi_j\rangle}{\sqrt{1-p_j}}, \quad |\gamma_{j,1}\rangle = \frac{\Pi_1 A \Delta_1 |\phi_j\rangle}{\sqrt{p_j}}, \quad |\delta_{j,0}\rangle = \frac{\Delta_0 A^\dagger \Pi_1 |\gamma_{j,1}\rangle}{\sqrt{1-p_j}}, \quad |\delta_{j,1}\rangle = \frac{\Delta_1 A^\dagger \Pi_1 |\gamma_{j,1}\rangle}{\sqrt{p_j}},$$

for each $j = 1, \ldots, 2^m$. As before, each of these vectors is a unit vector, $|\delta_{j,1}\rangle = |\phi_j\rangle$, and

$$\begin{aligned} A\ |\delta_{j,0}\rangle &= -\sqrt{p_j}\ |\gamma_{j,0}\rangle + \sqrt{1-p_j}\ |\gamma_{j,1}\rangle\,, \\ A\ |\delta_{j,1}\rangle &= \sqrt{1-p_j}\ |\gamma_{j,0}\rangle + \sqrt{p_j}\ |\gamma_{j,1}\rangle\,, \\ A^\dagger\ |\gamma_{j,0}\rangle &= -\sqrt{p_j}\ |\delta_{j,0}\rangle + \sqrt{1-p_j}\ |\delta_{j,1}\rangle\,, \\ A^\dagger\ |\gamma_{j,1}\rangle &= \sqrt{1-p_j}\ |\delta_{j,0}\rangle + \sqrt{p_j}\ |\delta_{j,1}\rangle\,. \end{aligned}$$

Moreover, each of the sets $\{|\gamma_{j,0}\rangle\}$, $\{|\gamma_{j,1}\rangle\}$, $\{|\delta_{j,0}\rangle\}$, and $\{|\delta_{j,1}\rangle\}$ is an orthonormal set. Because of this fact, when B is performed on the state $|\psi\rangle$, a similar pattern to the single eigenvector case arises independently for each eigenvector $|\psi_j\rangle$. This results in the stated probability of acceptance, which completes the proof. □

Applications of strong error reduction

Two applications of 4 will now be discussed. The first is a simplified proof that QMA is contained in the class PP.

Theorem 5. QMA $\subseteq$ PP.

Proof. Let $L \subseteq \Sigma^*$ be a language in QMA. By 4 there exists a function $m \in poly$ such that

$$L \in \mathrm{QMA}_m\left(1 - 2^{-(m+2)}, 2^{-(m+2)}\right).$$

Let A be a verification procedure that witnesses this fact. Specifically, each circuit A_x acts on $k+m$ qubits, for some $k \in poly$, and satisfies the following. If $x \in L$, then there exists an m qubit state $|\psi\rangle$ such that

$$\Pr[A_x \text{ accepts } |\psi\rangle] \geq 1 - 2^{-m-2},$$

while if $x \notin L$, then

$$\Pr[A_x \text{ accepts } |\psi\rangle] \leq 2^{-m-2}$$

for every m qubit state $|\psi\rangle$.

For each $x \in \Sigma^*$, define a $2^m \times 2^m$ matrix Q_x as

$$Q_x = \left(I_m \otimes \langle 0^k|\right) A_x^\dagger \Pi_1 A_x \left(I_m \otimes |0^k\rangle\right).$$

Each Q_x is positive semidefinite, and $\langle\psi|\, Q_x\, |\psi\rangle = \Pr[A_x \text{ accepts } |\psi\rangle]$ for any unit vector $|\psi\rangle$ on m qubits. The maximum probability with which A_x can be made to accept is the largest eigenvalue of Q_x. Because the trace of a matrix is equal to the sum of its eigenvalues and all eigenvalues of Q_x are nonnegative, it follows that if $x \in L$, then $\mathrm{tr}(Q_x) \geq 1 - 2^{-m-2} \geq 3/4$, while if $x \notin L$, then $\mathrm{tr}(Q_x) \leq 2^m 2^{-m-2} \leq 1/4$.

Now, based on a straightforward modification of the method of [FR99] discussed previously, we have that there exists a polynomially-bounded FP function g and GapP functions f_1 and f_2 such that the real and imaginary parts of the entries of Q_x are represented by f_1, f_2, and g in the sense that

$$\Re(Q_x[i,j]) = \frac{f_1(x,i,j)}{2^{g(x)}} \quad \text{and} \quad \Im(Q_x[i,j]) = \frac{f_2(x,i,j)}{2^{g(x)}}$$

for $0 \leq i,j < 2^m$. Define

$$h(x) = \sum_{i=0}^{2^m-1} f_1(x,i,i).$$

Because GapP functions are closed under exponential sums, we have $h \in$ GapP. It holds that $h(x) = 2^{g(x)} \operatorname{tr}(Q_x)$, and therefore

$$x \in L \;\Rightarrow\; h(x) \geq \frac{3}{4} 2^{g(x)} \quad \text{and} \quad x \notin L \;\Rightarrow\; h(x) \leq \frac{1}{4} 2^{g(x)}.$$

Because $2^{g(x)}$ is an FP function, it follows that $2h(x) - 2^{g(x)}$ is a GapP function that is positive if $x \in L$ and negative if $x \notin L$. Thus, $L \in$ PP as required. □

Remark 6. A simple modification of the above proof yields QMA $\subseteq$ A$_0$PP. Specifically, the GapP function $2h$ and the FP function $2^{g(x)}$ satisfy the required properties to prove $L \in$ A$_0$PP, namely

$$x \in L \;\Rightarrow\; 2h(x) \geq 2^{g(x)} \quad \text{and} \quad x \notin L \;\Rightarrow\; 2h(x) \leq \frac{1}{2} 2^{g(x)}.$$

The second application concerns one-message quantum Arthur–Merlin games where Merlin sends only a logarithmic number of qubits to Arthur. Classical one-message Arthur–Merlin games with logarithmic-length messages from Merlin to Arthur are obviously equivalent in power to BPP, because Arthur could simply search through all possible messages in polynomial time in lieu of interacting with Merlin. In the quantum case, however, this argument does not work, as one may construct exponentially large sets of pairwise nearly-orthogonal quantum states on a logarithmic number of qubits, such as those used in quantum fingerprinting [BCWdW01]. Nevertheless, logarithmic length quantum messages can be shown to be useless in the context of QMA using a different method, based on the strong error reduction property of QMA proved above.

For any choice of $a, b : \mathbb{N} \to [0,1]$, define $\text{QMA}_{\log}(a,b)$ to be the class of all languages contained in $\text{QMA}_m(a,b)$ for $m(n) = O(\log n)$, and let

$$\text{QMA}_{\log} = \text{QMA}_{\log}(2/3, 1/3).$$

The choice of the constants 2/3 and 1/3 is arbitrary, which follows from 4.

Theorem 7. $\text{QMA}_{\log} = \text{BQP}$.

Proof. The containment BQP $\subseteq \text{QMA}_{\log}$ is trivial, so it suffices to prove $\text{QMA}_{\log} \subseteq$ BQP. Assume $L \in \text{QMA}_m$ for m logarithmic, and assume A is a QMA verification procedure that witnesses this fact and has completeness and soundness error less than $2^{-(m+2)}$. Let

$$Q_x = \left(I_m \otimes \langle 0^k|\right) A_x^\dagger \Pi_1 A_x \left(I_m \otimes |0^k\rangle\right).$$

Similar to the proof of 5, we have

$$x \in L \;\Rightarrow\; \operatorname{tr}(Q_x) \geq 3/4, \qquad x \notin L \;\Rightarrow\; \operatorname{tr}(Q_x) \leq 1/4.$$

We will describe a polynomial-time quantum algorithm B that decides L with bounded error. The algorithm B simply constructs a totally mixed state over m qubits and runs the verification procedure A using this state in place of Merlin's message. Running the verification procedure on the totally mixed state is equivalent to running the verification procedure on m qubits initialized to some uniformly generated standard basis state, which is straightforward to simulate using Hadamard transforms and reversible computation. The totally mixed state on m qubits corresponds to the density matrix $2^{-m} I_m$, from which it follows that the probability of acceptance of B is given by

$$\Pr[B \text{ accepts } x] = \operatorname{tr}\left(Q_x \, 2^{-m} I_m\right) = 2^{-m} \operatorname{tr}(Q_x).$$

Given that m is logarithmic in $|x|$, we have that the probabilities with which B accepts inputs $x \in L$ and inputs $x \notin L$ are bounded away from one another by the reciprocal of some polynomial. This difference can be amplified by standard methods, implying that $L \in \mathrm{BQP}$. □

4 QAM

A QAM verification procedure A consists of a polynomial-time generated family

$$\left\{ A_{x,y} \,:\, x \in \Sigma^*,\, y \in \Sigma^{s(|x|)} \right\}$$

of quantum circuits together with functions $m, s \in \mathit{poly}$. As for QMA verification procedures, each circuit $A_{x,y}$ acts on two collections of qubits: $m(|x|)$ qubits sent by Merlin and $k(|x|)$ qubits corresponding to Arthur's workspace. The notion of a circuit $A_{x,y}$ accepting a message $|\psi\rangle$ is defined in the same way as for QMA. In the present case, the string y corresponds to a sequence of coin-flips sent by Arthur to Merlin, on which Merlin's message may depend.

Definition 8. The class $\mathrm{QAM}(a,b)$ consists of all languages $L \subseteq \Sigma^*$ for which there exists a QAM verification procedure A satisfying the following conditions.

1. If $x \in L$ then there exists a collection of states $\{|\psi_y\rangle\}$ on m qubits such that

$$\frac{1}{2^s} \sum_{y \in \Sigma^s} \Pr[A_{x,y} \text{ accepts } |\psi_y\rangle] \;\geq\; a.$$

2. If $x \notin L$ then for every collection of states $\{|\psi_y\rangle\}$ on m qubits it holds that

$$\frac{1}{2^s} \sum_{y \in \Sigma^s} \Pr[A_{x,y} \text{ accepts } |\psi_y\rangle] \;\leq\; b.$$

Similar to QMA, one may consider the cases where a and b are constants or functions of $n = |x|$, and in the case that a and b are functions of the input length it is assumed that $a(n)$ and $b(n)$ can be computed deterministically in time polynomial in n. Also as before, let $\mathrm{QAM} = \mathrm{QAM}(2/3, 1/3)$.

Error reduction for QAM

The first fact about QAM that we prove is that completeness and soundness errors may be reduced by running many copies of a given game in parallel. The proof is similar in principle to the proof of Lemma 14.1 in [KSV02], which corresponds to our 3.

Theorem 9. *Let $a, b : \mathbb{N} \to [0,1]$ and $q \in poly$ satisfy*

$$a(n) - b(n) \geq \frac{1}{q(n)}$$

for all $n \geq \mathbb{N}$. Then $\mathrm{QAM}(a,b) \subseteq \mathrm{QAM}(1 - 2^{-r}, 2^{-r})$ for every $r \in poly$.

Proof. Let $L \in \mathrm{QAM}(a,b)$, and let A be a QAM verification procedure witnessing this fact. We consider a new QAM verification procedure that corresponds to playing the game described by $\{A_{x,y}\}$ in parallel N times. The new procedure accepts if and only if the number of acceptances of the original game is at least $N \cdot \frac{a+b}{2}$. Although Merlin is not required to play the repetitions independently, we will show that playing the repetitions independently in fact gives him an optimal strategy. The theorem then follows by choosing an appropriately large value of N and applying a Chernoff-type bound.

Assume hereafter that the input x is fixed, and define

$$\begin{aligned} Q_y^{(0)} &= (I \otimes \langle 0^k|) A_{x,y}^{\dagger} \Pi_0 A_{x,y} (I \otimes |0^k\rangle), \\ Q_y^{(1)} &= (I \otimes \langle 0^k|) A_{x,y}^{\dagger} \Pi_1 A_{x,y} (I \otimes |0^k\rangle) \end{aligned}$$

for each $y \in \Sigma^s$. We have $Q_y^{(1)} = I - Q_y^{(0)}$, and consequently $Q_y^{(0)}$ and $Q_y^{(1)}$ share a complete set of orthonormal eigenvectors. Let $\{|\psi_{y,1}\rangle, \ldots, |\psi_{y,2^m}\rangle\}$ be such a set, and let

$$p_{y,1}^{(z)}, \ldots, p_{y,2^m}^{(z)}$$

be the corresponding eigenvalues for $Q_y^{(z)}$, $z \in \{0,1\}$. As $Q_y^{(0)}$ and $Q_y^{(1)}$ are positive semidefinite and sum to the identity, $p_{y,i}^{(0)}$ and $p_{y,i}^{(1)}$ are nonnegative real numbers with $p_{y,i}^{(0)} + p_{y,i}^{(1)} = 1$ for each y and i. Assume without loss of generality that the eigenvectors and eigenvalues are ordered in such a way that

$$p_{y,1}^{(1)} \geq \cdots \geq p_{y,2^m}^{(1)}.$$

This implies that the maximum acceptance probability of $A_{x,y}$ is $p_{y,1}^{(1)}$.

Under the assumption that Arthur's coin-flips for the N repetitions are given by strings

$$y_1, \ldots, y_N \in \Sigma^s,$$

if Merlin plays the repetitions independently, and optimally for each repetition, his probability of convincing Arthur to accept is

$$\sum_{\substack{z_1,\ldots,z_N \in \Sigma \\ z_1 + \cdots + z_N \geq N \cdot \frac{a+b}{2}}} p_{y_1,1}^{(z_1)} \cdots p_{y_N,1}^{(z_N)}. \tag{6}$$

Without any assumption on Merlin's strategy, the maximum probability with which Merlin can win $N \cdot \frac{a+b}{2}$ repetitions of the original game when Arthur's coin-flips are given by $y_1, \ldots, y_N$ is equal to the largest eigenvalue of

$$\sum_{\substack{z_1,\ldots,z_N \in \Sigma \\ z_1 + \cdots + z_N \geq N \cdot \frac{a+b}{2}}} Q_{y_1}^{(z_1)} \otimes \cdots \otimes Q_{y_N}^{(z_N)}. \tag{7}$$

Therefore, to prove the proposition it suffices to show that these quantities are equal.

All of the summands in equation 7 share the complete set of orthonormal eigenvalues given by

$$\left\{ |\psi_{y_1,i_1}\rangle \cdots |\psi_{y_N,i_N}\rangle \;:\; i_1,\ldots,i_N \in \{1,\ldots,2^m\} \right\},$$

and so this set also describes a complete set of orthonormal eigenvectors of the sum. The eigenvalue associated with $|\psi_{y_1,i_1}\rangle \cdots |\psi_{y_N,i_N}\rangle$ is

$$\sum_{\substack{z_1,\ldots,z_N\in\Sigma\\ z_1+\cdots+z_N\geq N\cdot\frac{a+b}{2}}} p^{(z_1)}_{y_1,i_1}\cdots p^{(z_N)}_{y_N,i_N}. \tag{8}$$

Define $u_1(X) = X$, $u_0(X) = 1 - X$, and let

$$f(X_1,\ldots,X_N) = \sum_{\substack{z_1,\ldots,z_N\in\Sigma\\ z_1+\cdots+z_N\geq N\cdot\frac{a+b}{2}}} u_{z_1}(X_1)\cdots u_{z_N}(X_N).$$

The quantity in equation 8 is equal to

$$f\left(p^{(1)}_{y_1,i_1},\ldots,p^{(1)}_{y_N,i_N}\right).$$

The function f is multi-linear and nondecreasing in each variable everywhere on the unit hypercube. Thus, the maximum of the quantity in equation 8 is

$$f\left(p^{(1)}_{y_1,1},\ldots,p^{(1)}_{y_N,1}\right),$$

which is equal to the quantity in equation 6. This completes the proof. □

An upper bound on QAM

We now observe that the upper bound

$$\mathrm{QAM} \subseteq \mathrm{BP}\cdot\mathrm{PP}$$

holds. The following fact concerning the maximum probabilities of acceptance of $A_{x,y}$ for random y will be used. Here we let $\mu(A_{x,y})$ denote the maximum probability that $A_{x,y}$ can be made to accept (maximized over all choices of Merlin's message $|\psi_y\rangle$).

Proposition 10. *Suppose that*

$$\left\{ A_{x,y} \;:\; x\in\Sigma^*,\, y\in\Sigma^{s(|x|)} \right\}$$

is a QAM *verification procedure for a language L that has completeness and soundness errors bounded by 1/9. Then for any $x\in\Sigma^*$ and for $y\in\Sigma^s$ chosen uniformly at random,*

$$x\in L \Rightarrow \Pr[\mu(A_{x,y})\geq 2/3] \;\geq\; 2/3$$

$$x\notin L \Rightarrow \Pr[\mu(A_{x,y})\leq 1/3] \;\geq\; 2/3.$$

Proof. Suppose that $x\in L$. Let $z(y) = 1-\mu(A_{x,y})$, and let Z be a random variable whose value is $z(y)$ for a uniformly chosen $y\in\Sigma^s$. The assumption of the proposition implies that $E[Z]\leq 1/9$. By Markov's inequality we have

$$\Pr[Z > 1/3] \leq \frac{E[Z]}{1/3} \leq 1/3,$$

and therefore

$$\Pr[\mu(A_{x,y}) \geq 2/3] = \Pr[Z \leq 1/3] \geq 2/3.$$

The proof for $x \notin L$ is similar. □

Theorem 11. QAM $\subseteq$ BP $\cdot$ PP.

Proof. Let $L \in$ QAM, and let

$$A = \left\{ A_{x,y} \,:\, x \in \Sigma^*, \, y \in \Sigma^{s(|x|)} \right\}$$

be a QAM verification procedure for L with completeness and soundness errors bounded by $1/9$. Such a procedure exists by 9. By a straightforward modification of the proof of 5, one may conclude that there exists a language $K \in$ PP such that

$$\mu(A_{x,y}) \geq 2/3 \Rightarrow (x,y) \in K,$$

$$\mu(A_{x,y}) \leq 1/3 \Rightarrow (x,y) \notin K.$$

It is possible that $\mu(A_{x,y}) \in (1/3, 2/3)$ for some values of y, but in this case no requirement is made on whether or not $(x,y) \in K$. The theorem now follows from 10. □

5 QMAM

A QMAM verification procedure A consists of a polynomial-time generated family

$$\left\{ A_{x,y} \,:\, x \in \Sigma^*, \, y \in \Sigma^{s(|x|)} \right\}$$

of quantum circuits, together with functions $m_1, m_2, s \in poly$. The functions m_1 and m_2 specify the number of qubits in Merlin's first and second messages to Arthur, while s specifies the number of random bits Arthur sends to Merlin. Each circuit $A_{x,y}$ acts on $m_1(|x|) + m_2(|x|) + k(|x|)$ qubits, where as before $k(|x|)$ denotes the number of qubits corresponding to Arthur's workspace.

In the QMAM case, it becomes necessary to discuss possible actions that Merlin may perform rather than just discussing states that he may send. This is because Merlin's strategy could involve preparing some quantum state, sending part of that state to Arthur on the first message, and transforming the part of that state he did not send to Arthur (after receiving Arthur's coin-flips) in order to produce his second message.

Definition 12. A language $L \subseteq \Sigma^*$ is in QMAM(a,b) if there exists a QMAM verification procedure A such that the following conditions are satisfied.

1. If $x \in L$ then for some l there exists a quantum state $|\psi\rangle$ on $m_1 + m_2 + l$ qubits and a collection of unitary operators $\{U_y \,:\, y \in \Sigma^s\}$ acting on $m_2 + l$ qubits such that

$$\frac{1}{2^s} \sum_{y \in \Sigma^s} \Pr[A_{x,y} \text{ accepts } (I_{m_1} \otimes U_y) \, |\psi\rangle] \;\geq\; a.$$

2. If $x \notin L$ then for every l, every quantum state $|\psi\rangle$ on $m_1 + m_2 + l$ qubits, and every collection of unitary operators $\{U_y \,:\, y \in \Sigma^s\}$ acting on $m_2 + l$ qubits,

$$\frac{1}{2^s} \sum_{y \in \Sigma^s} \Pr[A_{x,y} \text{ accepts } (I_{m_1} \otimes U_y) \, |\psi\rangle] \;\leq\; b.$$

The same assumptions regarding a and b apply in this case as in the QMA and QAM cases.

In the above definition, the circuit $A_{x,y}$ is acting on $m_1 + m_2$ qubits sent by Merlin in addition to Arthur's k workspace qubits, while $(I_{m_1} \otimes U_y)\,|\psi\rangle$ is a state on $m_1 + m_2 + l$ qubits. It is to be understood that the last l qubits of $(I_{m_1} \otimes U_y)\,|\psi\rangle$ remain in Merlin's possession, so $A_{x,y}$ is effectively tensored with the identity acting on these qubits.

Equivalence of QMAM and QIP

We now prove QMAM = QIP. Because quantum Arthur–Merlin games are a restricted form of quantum interactive proof systems, QMAM $\subseteq$ QIP is obvious. To prove the opposite containment, we will require the following lemmas. The first lemma is a corollary of Uhlmann's Theorem (see [NC00]).

Lemma 13. *Suppose the pair of registers (V, M) is in a mixed state for which the reduced state of V is σ. If the pair (V, M) is measured with respect to a binary valued measurement described by orthogonal projections $\{\Lambda_0, \Lambda_1\}$, then the probability of obtaining the outcome 1 is at most $F(\sigma, \rho)^2$ for some $\rho \in \mathcal{S}_{\mathsf{V}}(\Lambda_1)$.*

The second lemma is a simple property of the fidelity function.

Lemma 14 ([NS02, SR02])**.** *For any choice of density matrices ρ, ξ, and σ, we have*

$$F(\rho, \sigma)^2 + F(\sigma, \xi)^2 \leq 1 + F(\rho, \xi).$$

Theorem 15. *Let $L \in$ QIP and let $r \in$ poly. Then L has a three message quantum Arthur–Merlin game with completeness error 0 and soundness error at most $1/2 + 2^{-r}$. Moreover, in this quantum Arthur–Merlin game, Arthur's message consists of a single coin-flip.*

Proof. Let $L \in$ QIP, which implies that L has a three-message quantum interactive proof system with completeness error 0 and soundness error $\varepsilon(n) = 2^{-2r(n)}$ on inputs of length n.

Consider a QMAM verification procedure A that corresponds to the following actions for Arthur. (It will be assumed that the input x is fixed, and it will be clear that the family of quantum circuits corresponding to this verification procedure can be generated in polynomial-time given that the same is true of the verifier being simulated.)

1. Receive register V from Merlin.
2. Flip a fair coin and send the result to Merlin.
3. Receive register M from Merlin. If the coin flipped in step 2 was HEADS, apply V_2 to (V, M) and accept if the first qubit of V (i.e., the output qubit of the quantum interactive proof system) is 1, otherwise reject. If the coin in step 2 was TAILS, apply $V_1^\dagger$ to (V, M) and accept if all qubits of V are set to 0, otherwise reject.

Suppose first that $x \in L$, so that some prover, whose actions are described by a state $|\psi\rangle$ and a unitary operator U can convince V to accept with certainty. Then Merlin can convince Arthur to accept with certainty as follows:

1. Prepare state $|0^k\rangle$ in register V and state $|\psi\rangle$ in registers (M, P). Apply V_1 to registers (V, M), and send V to Arthur.
2. If Arthur flips HEADS, apply U to (M, P) and send M to Arthur. If Arthur flips TAILS, send M to Arthur without applying U.

Now assume $x \notin L$, so that no prover can convince V to accept with probability exceeding ε. Suppose that the reduced density matrix of register V sent by Merlin is σ. By 13 and 14, the probability that Arthur can be made to accept is at most

$$\frac{1}{2}F(\rho,\sigma)^2 + \frac{1}{2}F(\xi,\sigma)^2 \leq \frac{1}{2} + \frac{1}{2}F(\rho,\xi)$$

maximized over $\rho \in \mathcal{S}_{\mathsf{V}}(V_1 \Delta_1 V_1^{\dagger})$ and $\xi \in \mathcal{S}_{\mathsf{V}}(V_2^{\dagger}\Pi_1 V_2)$. By 1 this probability is at most

$$\frac{1}{2} + \frac{\sqrt{\varepsilon}}{2} \leq \frac{1}{2} + 2^{-r(|x|)},$$

which completes the proof. □

Corollary 16. *For any function $r \in poly$ we have* QIP $\subseteq$ QMAM$(1, 1/2 + 2^{-r})$.

Error reduction for QMAM

Now, suppose that we have a QMAM protocol for a language L with perfect completeness and soundness error b, and we repeat the protocol N times in parallel, accepting if and only if all N of the repetitions accept. It is clear that this resulting protocol has perfect completeness, because Merlin can play optimally for each parallel repetition independently and achieve an acceptance probability of 1 for any $x \in L$. In the case that $x \notin L$, Merlin can gain no advantage whatsoever over playing the repetitions independently, and so the soundness error decreases to b^N as we would hope. This follows from the fact that the same holds for arbitrary three-message quantum interactive proof systems [KW00], of which three-message quantum Arthur–Merlin games are a restricted type. This implies the following corollary.

Corollary 17. *For any function $r \in poly$ we have* QIP $=$ QMAM$(1, 2^{-r})$.

More than three messages

Finally, we note that one may define quantum Arthur–Merlin games having any polynomial number of messages in a similar way to three-message quantum Arthur–Merlin games. Such games are easily seen to be equivalent in power to three-message quantum Arthur–Merlin games. Specifically, polynomial-message quantum Arthur–Merlin games will be special cases of quantum interactive proof systems, and can therefore be parallelized to three-message interactive proofs and simulated by three-message quantum Arthur–Merlin games as previously described.

6 Open questions

Many interesting questions about quantum Arthur–Merlin games remain unanswered, including the following questions.

- Are there interesting examples of problems in QMA or QAM that are not known to be in AM? A similar question may be asked for QMAM vs. PSPACE.
- The question of if there exists an oracle relative to which BQP is outside of the polynomial-time hierarchy appears to be a difficult problem. In fact it is currently not even known if there is an oracle relative to which BQP $\not\subseteq$ AM. Is there an oracle relative to which QMA or QAM is not contained in AM? If so, what about QMA or QAM versus PH? Such results might shed some light on the problem of BQP versus the polynomial-time hierarchy.
- [NW94] proved almost-NP $=$ AM. Is it the case that almost-QMA $=$ QAM?

Acknowledgements

Thanks to Dorit Aharonov, Oded Regev, and Umesh Vazirani for their comments on error reduction for QMA, Ashwin Nayak for helpful references, and Alexei Kitaev for discussions about quantum proof systems. This research was supported by Canada's NSERC, the Canadian Institute for Advanced Research (CIAR), and the Canada Research Chairs program.

References

[ADH97] L. Adleman, J. DeMarrais, and M. Huang. Quantum computability. *SIAM Journal on Computing*, 26(5):1524–1540, 1997.

[AN02] D. Aharonov and T. Naveh. Quantum NP – a survey. Available as arXiv.org e-Print quant-ph/0210077, 2002.

[AR03] D. Aharonov and O. Regev. A lattice problem in quantum NP. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.

[Bab85] L. Babai. Trading group theory for randomness. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, pages 421–429, 1985.

[BCWdW01] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf. Quantum fingerprinting. *Physical Review Letters*, 87(16): article 167902, 2001.

[BM88] L. Babai and S. Moran. Arthur-Merlin games: a randomized proof system, and a hierarchy of complexity classes. *Journal of Computer and System Sciences*, 36(2):254–276, 1988.

[For97] L. Fortnow. Counting complexity. In L. Hemaspaandra and A. Selman, editors, *Complexity Theory Retrospective II*, pages 81–107. Springer, 1997.

[FR99] L. Fortnow and J. Rogers. Complexity limitations on quantum computation. *Journal of Computer and System Sciences*, 59(2):240–252, 1999.

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

[GS89] S. Goldwasser and M. Sipser. Private coins versus public coins in interactive proof systems. In S. Micali, editor, *Randomness and Computation*, volume 5 of *Advances in Computing Research*, pages 73–90. JAI Press, 1989.

[GW05] G. Gutoski and J. Watrous. Quantum interactive proofs with competing provers. In *Proceedings of the 22nd Symposium on Theoretical Aspects of Computer Science*, volume 3404 of *Lecture Notes in Computer Science*, pages 605–616. Springer, 2005.

[JWB03] D. Janzing, P. Wocjan, and T. Beth. "Identity check" is QMA-complete. Available as arXiv.org e-Print quant-ph/0305050, 2003.

[Kit97] A. Kitaev. Quantum computations: algorithms and error correction. *Russian Mathematical Surveys*, 52(6):1191–1249, 1997.

[Kit99] A. Kitaev. "Quantum NP". Talk at AQIP'99: Second Workshop on Algorithms in Quantum Information Processing, DePaul University, January 1999.

[KKR04] J. Kempe, A. Kitaev, and O. Regev. The complexity of the local hamiltonian problem. In *Proceedings of the 24th Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 372–383, 2004.

[KM03] H. Kobayashi and K. Matsumoto. Quantum multi-prover interactive proof systems with limited prior entanglement. *Journal of Computer and System Sciences*, 66(3), 2003.

[KMY03] H. Kobayashi, K. Matsumoto, and T. Yamakami. Quantum Merlin-Arthur proof systems: Are multiple Merlins more helpful to Arthur? In *Proceedings of the 14th Annual International Symposium on Algorithms and Computation*, 2003.

[Kni96] E. Knill. Quantum randomness and nondeterminism. Technical Report LAUR-96-2186, Los Alamos National Laboratory, 1996. Available as arXiv.org e-Print quant-ph/9610012.

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

[KSV02] A. Kitaev, A. Shen, and M. Vyalyi. *Classical and Quantum Computation*, volume 47 of *Graduate Studies in Mathematics*. American Mathematical Society, 2002.

[KW00] A. Kitaev and J. Watrous. Parallelization, amplification, and exponential time simulation of quantum interactive proof system. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 608–617, 2000.

[LFKN92] C. Lund, L. Fortnow, H. Karloff, and N. Nisan. Algebraic methods for interactive proof systems. *Journal of the ACM*, 39(4):859–868, 1992.

[NC00] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[NS02] A. Nayak and P. Shor. Bit-commitment based coin flipping. Available as arXiv.org e-Print quant-ph/0206123, 2002.

[NW94] N. Nisan and A. Wigderson. Hardness vs. randomness. *Journal of Computer and System Sciences*, 49(2):149–167, 1994.

[RS04] R. Raz and A. Shpilka. On the power of quantum proofs. In *Proceedings of the 19th Annual IEEE Annual Conference on Computational Complexity*, pages 260–274, 2004.

[RW04] B. Rosgen and J. Watrous. On the hardness of distinguishing mixed-state quantum computations. arXiv.org e-Print cs.CC/0407056, 2004.

[Sha92] A. Shamir. IP = PSPACE. *Journal of the ACM*, 39(4):869–877, 1992.

[SR02] R. Spekkens and T. Rudolph. Degrees of concealment and bindingness in quantum bit-commitment protocols. *Physical Review A*, 65: article 123410, 2002.

[Ter04] B. Terhal. Is entanglement monogamous? *IBM Journal of Research and Development*, 2004.

[Vya03] M. Vyalyi. QMA=PP implies that PP contains PH. Electronic Colloquium on Computational Complexity Report TR03-021, 2003.

[Wat00] J. Watrous. Succinct quantum proofs for properties of finite groups. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, pages 537–546, 2000.

[Wat03] J. Watrous. PSPACE has constant-round quantum interactive proof systems. *Theoretical Computer Science*, 292(3):575–588, 2003.