1 Local Hamiltonians

Recall that one postulate of quantum mechanics is that the evolution of a closed quantum system is characterized by a unitary transformation. That is, the state $|\psi\rangle$ of the system at time t_1 is related to the state $|\psi'\rangle$ of the system at time t_2 by a unitary operation U which depends only on time t_1 and t_2,

$$|\psi'\rangle = U|\psi\rangle.$$

Today we introduce a more refined version of this postulate, which describes the evolution of a quantum system in *continuous* time. It is stated as follows:

The time evolution of a state of a closed quantum system is described by *Shrödinger's equation*:

$$i\frac{d|\psi\rangle}{dt} = H|\psi\rangle.$$

H is a fixed Hermitian operator known as the Hamiltonian of the system. In specific, for an n-qubit system, its Hamiltonian H is a $2^n \times 2^n$ Hermitian matrix, i.e. $H = H^\dagger$.

Suppose H has a spectral decomposition

$$H = \sum_j \lambda_j |e_j\rangle\langle e_j|,$$

with eigenvalues λ_j's and corresponding eigenvectors $|e_j\rangle$'s. The states $|e_j\rangle$'s are conventionally referred to as energy eigenstates, or stationary states, and λ_j is the energy of the state $|e_j\rangle$. The lowest energy is known as the ground state energy for the system, and the corresponding energy eigenstate is known as the ground state.

Now suppose that at time $t = 0$ the initial state of the system is $|\psi(0)\rangle = |e_j\rangle$. Then a little calculus tells us that, at any time t, the system's state is given by $|\psi(t)\rangle = e^{-i\lambda_j t}|e_j\rangle$. So this explains why $|e_j\rangle$ are also called stationary states: their only change in time is to acquires an overall numerical factor.

Generally, suppose that at time $t = 0$ the initial state is $|\psi(0)\rangle = \sum_j \mu_j |e_j\rangle$, then at any time t the state of the system is given by $|\psi(t)\rangle = U(t)|\psi(0)\rangle = \sum_j \mu_j e^{-i\lambda_j t}|e_j\rangle$, where

$$U(t) = e^{-iHt} = \sum_{j=1}^{2^n} e^{-i\lambda_j t}|e_j\rangle\langle e_j|.$$

Remember that any unitary transformation U can be realized by a quantum circuit constructed from a universal set of quantum gates, i.e. $U = U_T U_{T-1} \ldots U_1$, where U_j's are local operations. However, a counting argument tells us that most unitary operations cannot be efficiently implemented in this way, i.e. they require the circuit to contain exponentially number of gates. The picture is also similar for Hamiltonians. Not all Hamiltonians can be easily implemented. The realistic Hamiltonians are local Hamiltonians. They are the

Hamiltonian that can be written as a sum over many local interactions. Specifically, suppose for a system of n particles,

$$H = \sum_{j=1}^{r} H_j,$$

where each H_j acts on at most a constant c number of particles (i.e. $H_j = A_j \otimes I$ for some c-particle operator A_j). Then we say that H is c-local. Such locality is quite physically reasonable, and originates in many systems from the fact that most interactions fall off with increasing distance of difference in energy.

Local Hamiltonians and quantum circuits can (approximately) simulate each other with polynomial overhead. To prove this, first observe that for any c-local Hamiltonian $H_j = A_j \otimes I$, the exponent $e^{-iH_j} = e^{-iA_j} \otimes I$ is a c-local unitary operations; conversely, for any c-local unitary operations U_j, we can also find a c-local Hamiltonian H_j such that $U_j = e^{-iH_j}$.

Suppose we are given a circuit $U = U_T U_{T-1} \dots U_1$, where U_j's are gates. We can write each gate $U_j = e^{-iH_j}$ for some local Hamiltonian H_j. Then $U = e^{-iH_T} e^{-iH_2} \dots e^{-iH_1}$. So U can be implemented by local Hamiltonians as follows: we first let set the Hamiltonian of the system to be H_1 and let it evolve for a unit time, then change its Hamiltonian into H_2 and let it evolve for another unit time, and so on.

The converse direction is a little complicated. Given a local Hamiltonian $H = \sum_{j=1}^{r} H_j$ and a time t, we want to use a quantum circuit to simulate the transition operator $U(t) = e^{-iHt}$. The difficulty arises from the fact that $e^{-iHt} \neq e^{-iH_1 t} e^{-iH_2 t} \dots e^{-iH_r t}$ in general, because H_j's may not commute with each other. Then, how can we use $e^{-iH_j t}$ to construct e^{-iHt}? Recall that e^x has the following Taylor expansion

$$e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!} = 1 + x + \frac{x^2}{2} + \dots.$$

So we get

$$e^{-iH_j t/n} = I - iH_j t/n + O(\frac{1}{n^2}),$$

and hence

$$e^{-i(H_1+H_2)t/n} = e^{-iH_1 t/n} e^{-iH_2 t/n} + O(\frac{1}{n^2}).$$

Noting that $e^{-i(H_1+H_2)t} = (e^{-i(H_1+H_2)t/n})^n$, we obtain

$$e^{-i(H_1+H_2)t} = (e^{-iH_1 t/n} e^{-iH_2 t/n})^n + O(\frac{1}{n}).$$

This formula is called Trotter expansion. It can be further generalized to

$$e^{-i(H_1+H_2+\dots+H_r)t} = (e^{-iH_1 t/n} e^{-iH_2 t/n} \dots e^{-iH_r t/n})^n + O(\frac{r}{n}).$$

This gives us the method of simulating e^{-iHt} with a quantum circuit, because each $e^{-iH_j t/n}$ is local and can be efficiently implemented by a quantum circuit. By choosing n to be large enough we can obtain a good approximation of e^{-iHt}.

2 QMA

The class NP (non-deterministic polynomial time) contains many thousand of the most important computational problems. Of these problems, the vast majority are NP-complete. This means that these are the hardest problems in NP. By this we mean that, if anyone of them can be solved by a polynomial time algorithm, then every problem in NP can be solved by a polynomial time algorithm. The cornerstone of this theory of NP-completeness is the Cook-Levin theorem, which states that 3-SAT is NP-complete.

A language L is in NP if there is a polynomial time proof checker C and a polynomial *poly*, with the following property: if $x \in L$ then there is a string y with $|y| \leq poly|x|$, such that $C(x,y) = 1$. If $x \notin L$, then for every y such that $|y| \leq poly(|x|)$, $C(x,y) = 0$.

Kitaev gave the quantum analogue of the Cook-Levin theorem by showing that QSAT the quantum analogue of 3-SAT is complete for the quantum analogue of NP, called BQNP or QMA.

QMA is the quantum generalization of MA — the probabilistic analogue of NP. To define MA, we simply replace the deterministic polynomial time proof checker with a probabilistic polynomial time proof checker C. Now if $x \in L$, then there is a string y with $|y| \leq poly|x|$, such that $C(x,y) = 1$ with probability at least $2/3$. If $x \notin L$, then for every y such that $|y| \leq poly(|x|)$, $C(x,y) = 0$ with probability at least $2/3$.

To define QMA, the quantum analogue of MA, we replace the probabilistic polynomial time proof checker by a quantum polynomial time proof checker. Equally important, the witness string y is now allowed to be a quantum witness, i.e., it can be a superposition over strings of length at most $poly(|x|)$.

Remark: If we require the witness y to be classical, but leave the verifier C to be quantum, then the corresponding class is usually called QCMA. It is obvious that QCMA $\subseteq$ QMA. But we do not know whether QCMA is equal to QMA.

BQP is trivially contained in BQNP since it can be simulated by the verifier alone. MA is also contained in BQNP since quantum machines can perform the classical computations of their classical counterparts. Kitaev's proof that QSAT is BQNP-complete implies a non-trivial upper bound, showing that $BQNP \subseteq P^{\#P}$.

A QMA-Complete Problem

Consider the following problem: **Local Hamiltonians or Q5SAT:** Let H_j (for $j = 1, \ldots r$) be 5-local Hamiltonians on n qubits (each specified by complex $2^5 \times 2^5$ matrices.). Assume that each H_j is scaled so that all eigenvalues λ of H_j satisfy $0 \leq \lambda \leq 1$. Let $H = \sum_{j=1}^{r} H_j$. There is a promise about H that either all eigenvalues of H are $\geq b$ or there is an eigenvalue of H that is $\leq a$, where $0 \leq a < b \leq 1$ and the difference $b - a$ is at least inverse polynomial in n, i.e., $b - a \geq \frac{1}{poly(n)}$. The problem asks whether H has an eigenvalue $\leq a$.

The Connection with 3-SAT

In 3-SAT, we are given a formula f on n variables in 3-CNF (conjunctive normal form.) That is, f is a conjunction of many clauses c_i:

$$f(x_1, x_2, \ldots, x_n) = c_1 \wedge c_2 \wedge \ldots \wedge c_m \ ,$$

where each clause c_j is a disjunction of three variables or their negations. For example, c_j may be $(x_a \vee \overline{x_b} \vee x_c)$.

We would like to make a corresponding Hamiltonian H_i for each clause c_i. H_i should penalize an assignment which does not satisfy the clause c_i. In the example where $c_j = (x_a \vee \overline{x_b} \vee x_c)$, we want to penalize the assignment state $|010\rangle$. If our notion of *penalize* is to have a positive eigenvalue, then we can let $H_j = |010\rangle\langle 010|$, and define the other H_i's similarly, i.e., each H_i has a 1 eigenvalue with a corresponding eigenvector that causes clause c_i to be false.

Finally, we let

$$H = \sum_{i=1}^{m} H_i,$$

so that H is a sum of 3-local Hamiltonians. It is not hard to see that the smallest eigenvalue of H is the minimum (over all assignments) number of unsatisfied clauses. In particular, H has a 0 eigenvalue exactly when there is a satisfying assignment for f.

For general QSAT instances, the Hamiltonians H_j cannot be simultaneously diagonalized in general, and the problem appears much harder.

Membership in QMA

We can assume without loss of generality that each H_j is just a projection matrix $|\phi_j\rangle\langle\phi_j| \otimes I$. The prover would like to provide convincing and easily verifiable evidence that $H = \sum_j H_j$ has a small eigenvalue $\lambda \leq a$. The proof consists of (a tensor product of) polynomial in n copies of the corresponding eigenvector $|\eta\rangle$, which satisfies $\lambda = \sum_j \langle\eta|H_j|\eta\rangle$. Given a single copy of $|\eta\rangle$, the verifier can flip a coin with bias $\frac{\lambda}{r}$ as follows:

1. Pick $H_j = |\phi_j\rangle\langle\phi_j| \otimes I$ at random,
2. Measure $|\eta\rangle$ by projecting onto $|\phi_j\rangle$.

This succeeds with probability $\frac{\lambda}{r}$. Given the promise that $\lambda \leq a$ or $\lambda \geq b$, it suffices for the verifier to repeat this test $\frac{r^2}{(b-a)^2}$ times to conclude with high confidence that $\lambda \leq a$. Thus polynomial in n copies of $|\eta\rangle$ are sufficient. Note that since the verifier is performing each test randomly and independently, the prover gains no advantage by sending an entangled state to the verifier.

QMA-Completeness

To show that QSAT is complete in QMA, we need to show that the universal BQNP problem reduces to it. That is, given a quantum circuit $U = U_L U_{L-1} \ldots U_1$ and a promise that exactly one of the following holds:

1. $\exists|\eta\rangle$, U accepts on input $|\eta\rangle$ with probability $\geq p_1 = 1 - \varepsilon$,
2. $\forall|\eta\rangle$, U accepts on input $|\eta\rangle$ with probability $\leq p_0 = \varepsilon$,

The challenge is to design an instance of QSAT which allows us to distinguish the above two cases. i.e. we wish to specify a sum of local Hamiltonians that has an eigenvector with small eigenvalue if and only if case 1 happens.

The construction of the local Hamiltonian is analogous to Cook's theorem. The quantum analogue of the accepting tableau in Cook's theorem will be the computational history of the quantum circuit:

$$\begin{aligned} |T\rangle &= \frac{1}{\sqrt{L+1}} \sum_{t=0}^{L} |\phi_t\rangle \otimes |t\rangle \\ &= \frac{1}{\sqrt{L+1}} \sum_{t=0}^{L} U_t U_{t-1} \ldots U_1 |\phi_0\rangle \otimes |t\rangle, \end{aligned}$$

where $|\phi_0\rangle$ is a valid initial state and $|\phi_i\rangle = U_i |\phi_{i-1}\rangle$. Thus the computation history $|T\rangle$ is an element of $(\mathscr{C}^2)^{\otimes n} \otimes \mathscr{C}^{L+1}$. It is a superposition over time steps of the state of the qubits as the quantum circuit operates on them.

Now the idea of the QMA-completeness proof is to design the Hamiltonian H such that:

1. if $\exists |\eta\rangle$, U accepts on input $|\eta\rangle$ with probability at least $1-\varepsilon$, then the corresponding computational history $|T\rangle$ is an eigenvector of H with eigenvalue at most $\frac{\varepsilon}{L+1}$,
2. if U rejects on every input with probability at least $1-\varepsilon$, then all the eigenvalues of H are at least $\frac{c(1-\varepsilon)}{(L+1)^3}$ for some constant c.

Our Hamiltonian will be the sum of three terms,

$$H = H_{initial} + H_{final} + H_{propagate}.$$

The first two terms are simple and express the condition that the computational history starts with a valid input state, and ends in an accepting state.

We consider the first m qubits of U's state to be the input qubits and the remaining $n-m$ qubits to be the clean work qubits. The design of the $H_{initial}$ component should then reflect that at time 0, all of the work bits are clear:

$$H_{initial} = \sum_{s=m+1}^{n} \Pi_s^{(1)} \otimes |0\rangle \langle 0|,$$

where $\Pi_s^{(1)}$ denotes projection onto the s-th qubit with value $|1\rangle$.

Assume that the state of the first qubit at the output determines whether or not the input is accepted. Then H_{final} needs to indicate that at time L the first qubit is a $|1\rangle$:

$$H_{final} = \Pi_1^{(0)} \otimes |L\rangle \langle L|.$$

The most complicated component of H is $H_{propagate}$, which captures transitions between time steps.

$$H_{propagate} = \sum_{t=1}^{L} H_t,$$

where

$$H_t = -\frac{1}{2}U_t \otimes |t\rangle\langle t-1| - \frac{1}{2}U_t^\dagger \otimes |t-1\rangle\langle t| + \frac{1}{2}I \otimes (|t\rangle\langle t| + |t-1\rangle\langle t-1|).$$

The fact that the computational history is a superposition over time steps is quite crucial here. To check that the correct operation has been applied in step t, it suffices to restrict attention to the $|t-1\rangle$ and $|t\rangle$ clock states. Now the quantum register is in a superposition over its state at time $t-1$ and at time t. Locally checking this superposition is sufficient to determine whether its clock t component is the result of applying the quantum gate U_t to the clock $t-1$ component. This is precisely what the Hamiltonian H_t above is designed to do.

Next we show that if U accepts on input $|\eta\rangle$ with probability at least $1-\varepsilon$, then the corresponding computational history $|T\rangle = \frac{1}{\sqrt{L+1}}\sum_{t=0}^{L}|\phi_t\rangle \otimes |t\rangle$ is an eigenvector of H with eigenvalue at most $\frac{\varepsilon}{L+1}$. We analyze the contribution from each component of H. First, $|T\rangle$ starts with qubits $m+1$ through n clear, so $H_{initial}$ does not contribute to $H|T\rangle$. Second, $|T\rangle$ is a computation of U, i.e. $|\phi_t\rangle = U_t|\phi_{t-1}\rangle$ for all t, so we get:

$$\begin{aligned}
H_t|T\rangle &= \frac{1}{\sqrt{L+1}}(-\frac{1}{2}U_t|\phi_{t-1}\rangle|t\rangle - \frac{1}{2}U_t^\dagger|\phi_t\rangle|t-1\rangle + \frac{1}{2}|\phi_t\rangle|t\rangle + \frac{1}{2}|\phi_{t-1}\rangle|t-1\rangle) \\
&= \frac{1}{\sqrt{L+1}}(-\frac{1}{2}|\phi_t\rangle|t\rangle - \frac{1}{2}|\phi_{t-1}\rangle|t-1\rangle + \frac{1}{2}|\phi_t\rangle|t\rangle + \frac{1}{2}|\phi_{t-1}\rangle|t-1\rangle) \\
&= 0,
\end{aligned}$$

for no contribution from $H_{propergate}$. Finally, U accepts with probability at least $1-\varepsilon$, so only H_{final} contributes a penalty to the sum, which is at most $\frac{\varepsilon}{L+1}$.

The hard part of the proof lies in showing the converse, i.e. if U rejects on every input with high probability, then all the eigenvalues of H are high. Here we only sketch the proof. For more technical details, please refer to [1, 2].

The idea is to write H as a sum of two Hamiltonians, $H_1 = H_{initial} + H_{final}$, $H_2 = H_{propagate}$, and to use the following geometrical lemma, which gives a lower bound on the lowest eigenvalue of a sum of two Hamiltonians, given some conditions on the eigenvalues and eigenspaces of the two Hamiltonians.

Lemma 8.1: *Let H_1 and H_2 be two Hermitian positive semi-definite matrices, and let N_1 and N_2 be the eigenspaces of the eigenvalue 0, respectively. If the angle between N_1 and N_2 is some $\theta > 0$, and the second eigenvalues of both H_1 and H_2 are at least λ, then the minimal eigenvalue of $H_1 + H_2$ is at least $\lambda \sin^2(\theta/2)$.*

Proof: Suppose $|\delta\rangle$ is an arbitrary eigenvector of $H_1 + H_2$. For at least one of the subspaces N_1 or N_2, the angle between $|\delta\rangle$ and this subspace is at least $\frac{\theta}{2}$. Without loss of generality, let this subspace be N_1. Then we have

$$\langle\delta|(H_1+H_2)|\delta\rangle = \langle\delta|H_1|\delta\rangle + \langle\delta|H_2|\delta\rangle \geq \langle\delta|H_1|\delta\rangle.$$

Suppose $|\mu\rangle, |\mu^\perp\rangle$ are the projections of $|\delta\rangle$ onto N_1 and its orthogonal complement $N_1^\perp$ respectively. Then we have

$$\langle\delta|H_1|\delta\rangle = \langle\mu^\perp|H_1|\mu^\perp\rangle \geq \lambda\||\mu^\perp\rangle\|^2 \geq \lambda\sin^2(\theta/2),$$

where the first equality follows from the fact that N_1 and and its complement are invariant to the application of H_1, the second follows from the definition of H_1 and λ, and the last follows from $\||\mu^\perp\rangle\|^2 \geq sin^2(\theta/2)$ because the angle between N_1 and $|\delta\rangle$ is at least $\theta/2$. □

To use the geometrical lemma, we need to give lower bounds on the second eigenvalues of H_1 and H_2, as well as a lower bound on θ.

We will first bound the second eigenvalues of H_1 and H_2.

Lemma 8.2: *The second eigenvalue of H_1 is at least* 1.

Proof: Since $H_{initial}$ and H_{final} are projections, and the eigenspaces of the eigenvalue 1 of $H_{initial}$ and H_{final} are orthogonal (because they operate on different times), the second eigenvalue of $H_1 = H_{initial} + H_{final}$ is simply the minimal second eigenvalue of the two. □

Lemma 8.3: *The second eigenvalue of H_2 is at least $\frac{1}{2(L+1)^2}$.*

Proof: It turns out that for this argument it is simpler to look at $H_{propagate}$ in a rotated basis. The eigenvalues of a matrix are not changed when looked at in a different basis. Hence we define the unitary matrix R as follows:

$$R = \sum_{t=0}^{L} U_t U_{t-1} ... U_1 \otimes |t\rangle\langle t|.$$

It is easy to check that

$$R^\dagger H_{propagate} R = \frac{1}{2} \sum_{t=1}^{L} (I \otimes |t\rangle\langle t| + I \otimes |t-1\rangle\langle t-1| - I \otimes |t-1\rangle\langle t| - I \otimes |t\rangle\langle t-1|).$$

We can write $H_{propagate} = I \otimes A$ where A is a $(L+1) \times (L+1)$ matrix of the form:

$$A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$= I - \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = I - B.$$

So the second smallest eigenvalue of $H_{propagate}$, is simply 1 minus the second largest eigenvalue of B. Interestingly, the matrix B is the stochastic matrix corresponding to a simple random walk on a line with $L+1$ vertices, with a loop at both ends. By Cheeger inequality, we can bound B's second largest eigenvalue λ_2 by the conductance ϕ of the underlying graph,

$$1 - \lambda_2 \geq \phi^2/2,$$

The conductance of this graph is $\frac{1}{L+1}$, which gives $1 - \lambda_2 \geq \frac{1}{2(L+1)^2}$. This implies the desired result. □

It is left to give a lower bound on the angle between the two null spaces.

Lemma 8.4: *The angle between N_1 and N_2 satisfies $sin^2(\theta/2) \geq \frac{c(1-\varepsilon)}{L+1}$ for some constant c.*

Proof Sketch: We note that N_1, the null space of $H_1 = H_{initial} + H_{final}$, is the space spanned by all states with valid input (i.e. all work qubits are clear) and accepting output (i.e. the first qubit is $|1\rangle$); N_2, the null space of $H_2 = H_{propagate}$, is the space spanned by all valid computations starting with arbitrary input. We now want to bound the angle θ between N_1 and N_2, which is the minimal angle between any two states from both spaces. For any such two states, their inner product is the sum of contributions from every time leaf $t = 0, 1, \ldots, L$. Due to the fact that U rejects on any input with high probability, the contribution from the $t = 0$ and $t = T$ leafs together is far from the maximal possible contribution $\frac{2}{L+1}$. A careful computation tells us that their inner product is $\leq 1 - \frac{c'(1-\varepsilon)}{L+1}$, which implies $\sin^2\theta \geq \frac{c'(1-\varepsilon)}{L+1}$. □

Putting the above lemmas together, we obtain the soundness of our construction.

To make our Hamiltonian truly 5-local, we need to move from operators on the entire clock to local operators. To achieve this, we represent the time in unary representation on L qubits which serve as the clock qubits. Specifically, any time t is represented by the state $|11\ldots100\ldots0\rangle$ which begins with t 1's. To modify the Hamiltonian accordingly, we replace all operators on the clock space by operators that operate on three qubits at most. We apply the following modifications:

$$\begin{aligned}
|t\rangle\langle t-1| &\longmapsto |110\rangle\langle 100| \otimes I \\
|t-1\rangle\langle t| &\longmapsto |100\rangle\langle 110| \otimes I \\
|t\rangle\langle t| &\longmapsto |110\rangle\langle 110| \otimes I \\
|t-1\rangle\langle t-1| &\longmapsto |100\rangle\langle 100| \otimes I
\end{aligned}$$

where in all these cases $|110\rangle\langle 100|$ or the similar terms operate on qubits $t-1, t, t+1$ of the clock qubits and the identity I operates on the remaining $L-3$ clock qubits. To avoid referring to qubits 0 and $L+1$ which do not exist, we make two exceptions: for $t = 1$, we drop the first bit of the 3-bit operator; for $t = L$, we drop the last bit of the 3-bit operator.

In addition, we introduce a new term H'_{clock} which checks that the clock bits are a valid unary representation and penalizes them if they are not. It can be done locally as follows:

$$H'_{clock} = \sum_{t=2}^{L} |01\rangle\langle 01|_{t-1,t} \otimes I.$$

Our final Hamiltonian is defined to be

$$H' = H'_{initial} + H'_{final} + H'_{propagate} + H'_{clock},$$

where $H'_{initial}, H'_{final}, H'_{propagate}$ are the modified version of corresponding original terms.

It is easy to see that the following statement go through with these modifications: If U accepts on input $|\eta\rangle$ with probability at least $1-\varepsilon$, and $|T'\rangle$ is the corresponding computational history, then $\langle T'|H'|T'\rangle \leq \frac{\varepsilon}{L+1}$.

For the converse direction, observe that H' keeps the subspace that is spanned by all states in which the clock qubits are valid unary representations invariant, and let us call this subspace $\mathscr{W}$. Its orthogonal complement $\mathscr{W}^{\perp}$ is also invariant under the operation of H'. H' operates on $\mathscr{W}$ just as the previous H did, and hence on this subspace the lower bound on the eigenvalues holds as before; on the orthogonal subspace $\mathscr{W}^{\perp}$ the

eigenvalue of H' is at least 1 since H'_{clock} detects at least one violation. Hence, overall, the original lower bound still holds.

Upper bound on QMA

One consequence of the previous proof of QMA-completeness is the following:

Theorem 8.1: $QMA \subseteq P^{\#P}$.

Replace H with $I-H$, so it either has an eigenvalue greater than or equal to $a' = 1-a$ or all eigenvalues are smaller than $b' = 1-b$. Consider the trace of H^k. This is either at least a'^k or at most Nb'^k. We can make sure that $a'^k >> Nb'^k$, by choosing $k >> n^d logN$. So we just need to estimate $Tr(H^k)$ in $P^{\#P}$.

To see this, write $Tr(H^k) = Tr((\sum_j H_j)^k) = Tr(\sum_{j_1,\ldots j_k} H_{j_1} \cdots H_{j_k}) = \sum_{j_1,\ldots j_k} Tr(H_{j_1} \cdots H_{j_k})$. Each trace in this sum is itself just a sum of exponentially many easy to compute contributions, and thus the entire sum is easily seen to be estimated in $P^{\#P}$.

References

[1] Dorit Aharonov, Tomer Naveh, Quantum NP - A Survey, http://arxiv.org/abs/quant-ph/0210077.

[2] A. Yu Kitaev, A. Shen, M. N. Vyalyi, Classical and Quantum Computation, American Mathematical Society, 2002.