CMSC 858L: Quantum Complexity
Unit 4: QMA

Instructor: Daniel Gottesman

Fall 2026

1 Local Hamiltonians

Working our way up to larger complexity classes, our next stop is QMA. Recall the definition of QMA:

Definition 1. *Fix a particular efficiently computable approximately universal gate set $\mathcal{G}$. Let QMA be the set of languages L such that there exists a uniform family of polynomial-size quantum circuits $Q_{n,m}$ (with gates drawn from $\mathcal{G}$) that takes two inputs $(x, |\psi\rangle)$ with the following properties: For any instance x,*

1. *If $x \in L$, then exists an n-qubit state $|\psi\rangle_x$ with $n = O(\text{poly}(|x|))$ and $\text{Prob}(M(Q_{|x|,n}|x\rangle \otimes |\psi\rangle_x) = 1) \geq 2/3$.*
2. *If $x \notin L$, then for any n-qubit state $|\psi\rangle_x$ with $n = O(\text{poly}(|x|))$, $\text{Prob}(M(Q_{|x|,n}|x\rangle \otimes |\psi\rangle_x) = 0) \geq 2/3$.*

In the first case, $|\psi\rangle_x$ is the witness *for $x \in L$.*

As the most natural quantum analogue of NP, a natural question to ask about it is what does a QMA-complete problem look like? The first one was found by Kitaev, and that is the *local Hamiltonian* problem. A *Hamiltonian* is a Hermitian (i.e. self-adjoint) operator which represents the energy of a system. Hamiltonians have an important role in physics because they determine almost all the properties of the system in question, including how they evolve in time given an initial state, their behavior at thermal equilibrium, and properties of excitations. The spectrum of the Hamiltonian (the list of eigenvalues) already says a lot about a Hamiltonian, even without information about the eigenvectors themselves. At very low energies, the lowest eigenvalues are the most important, and the *ground state*, the lowest-energy state, most of all. (However, this to an extent dependent on the energies of other low-lying states.)

But what is a *local* Hamiltonian?

Definition 2. *A* k-local Hamiltonian *is a Hermitian operator H which can be written in the form*

$$H = \sum_S H_S \otimes I, \tag{1}$$

where the sum is taken over subsets S of at most k qubits and H_S is a Hamiltonian acting only on the qubits in S.

Thus, a k-local Hamiltonian is a Hamiltonian that is a sum of terms each of which acts on at most k qubits. Here, k is usually assumed to be a constant independent of the number of qubits in the system.

Note that we are not making any constraints on where these k qubits are located in physical space. They can be far apart from each other. If we want a Hamiltonian where all terms only interact qubits that are physically near each other, we will call it a *geometrically local* Hamiltonian or just one that is *local in D dimensions*, where D is the geometric dimension of the space containing the qubits' physical locations. I want to use the term *geometric dimension* here (or *spatial dimension*) because sometimes we also want to talk

about local Hamiltonians using qudits which have a Hilbert space of dimension d. Then the Hilbert space dimension d of each qudit is called the *local dimension*. The local Hamiltonian problem is unfortunately full of terminology like this that can be interpreted two ways. There is actually one more case of confusing terminology that we will get to soon.

The local Hamiltonian problem is then to determine the ground state energy (the lowest eigenvalue) of a local Hamiltonian. However, because the energy is a real number, we are going to want to limit the precision we need:

Definition 3. *The language* k-LOCAL HAMILTONIAN *considers instances of the form* (H, E, Δ) *with the promise that* $0 < \Delta = \Omega(1/\text{poly}(n))$ *(where* n *is the size of the instance) and the lowest eigenvalue (*ground state energy*) of* H *is either* $\leq E$ *or* $\geq E + \Delta$*. A "yes" instance is* (H, E, Δ) *such that the ground state energy of* H *is* $\leq E$*.*

Δ is essentially the precision that we need to specify the Hamiltonian's ground state energy to. k-LOCAL HAMILTONIAN is a decision problem, but clearly there is a related search problem: Find the ground state energy to precision Δ. If we have an algorithm for the decision problem, we can solve the search problem by trying different values of E until we narrow down the ground state energy. Note that there is a potential hiccup depending on what the algorithm does when the promise is not satisfied (i.e., the ground state energy is between E and $E + \Delta$, but we can deal with this by trying a few different (E, Δ) pairs at each step before updating our window on E.

Δ is sometimes called the *promise gap* of the instance. This should be distinguished from the *spectral gap* of the Hamiltonian, which is the difference between the lowest two eigenvalues. (Yes, this is the other example of confusing terminology I was mentioning.) The spectral gap has an important impact on the physics of the system, but the promise gap is just something somewhat artificial introduced to make this a problem in QMA rather than some much higher class. (Although precision is a perfectly reasonable physics concept too.) Spectral gap vs. promise gap is a bit more confusing generally than the terminology "local" and "dimension" because they both refer to something about the energies. In some example instances, the two are actually the same, but that is not generally true.

Now, what is the complexity of k-LOCAL HAMILTONIAN? The first thing we can say is:

Theorem 1. *k-LOCAL HAMILTONIAN is in QMA.*

Proof sketch. There exist efficient quantum algorithms to simulate the dynamics of any k-local Hamiltonian on an initial state provided to us. Thus, if Merlin claims that an instance of k-LOCAL HAMILTONIAN is a "yes" instance, he can provide us with a witness which is a copy of the ground state $|\psi\rangle$. We then create an ancilla in the state $\sum_t |t\rangle$ and conditionally run the simulation of the Hamiltonian H on $|\psi\rangle$ for a time t. If $|\psi\rangle$ is indeed an eigenstate of energy E_0, then the Hamiltonian gives a phase $e^{-iE_0 t/\hbar}$ after a time t. The ancilla thus ends up unentangled with the witness and in the state $\sum_t e^{-iE_0 t/\hbar}|t\rangle$. The quantum Fourier transform will then let us measure E_0. There are some details to fill out here (which I will skip), but we don't get E_0 exactly, but merely some approximation $\tilde{E}_0$ to E_0 whose accuracy is inverse polynomial in the size of the simulation and ancilla. We pick our circuits so that the accuracy is less than $\Delta/2$ with probability $\geq 2/3$.

If we find $\tilde{E}_0 < E + \Delta/2$, then we can accept the witness. In the case that Merlin is being honest and has provided us with a valid witness with energy at most E, we will accept it with high probability. On the other hand, if this is a "no" instance and Merlin provided us a different eigenstate, then $E_0 \geq E + \Delta$, so with probability at least $2/3$, $\tilde{E}_0 > E + \Delta/2$, and we will reject the witness.

But what if we have a "no" instance and Merlin gives us something which is not an eigenstate? We can expand an arbitrary state $|\psi\rangle = \sum_i \alpha_i |\psi_i\rangle$, where $|\psi_i\rangle$ run over an orthonormal basis of eigenstates with energies E_i. Since it is a "no" instance, $E_i \geq E + \Delta$ for all i. What happens when we run our simulation? We now get the state

$$\sum_i \alpha_i \sum_t e^{-iE_i t/\hbar}|t\rangle|\psi_i\rangle. \tag{2}$$

This is entangled between the ancilla and system. Since the eigenstates are orthogonal to each other, if we trace out the witness, the state of our ancilla is a mixture over pure states $\sum_t e^{-iE_i t/\hbar}|t\rangle$ with probability $|\alpha_i|^2$. When we measure, we get an estimate of E_i.

However, there is still one more complication: The Fourier transform doesn't give us a direct estimate of E_i, but rather of a multiple of $1/E_i$; you need multiple such estimates to actually learn E_i. If we ask for multiple copies of the witness, we could measure it multiple times, but we would be getting different E_i each time, making it hard to extract the answer. However, we can just measure it repeatedly on the same copy: In the case where the witness is genuinely an eigenstate, this doesn't disturb the state (modulo imperfections in the simulation), and in the case where it is superposition we get

$$\sum_i \alpha_i \left(\sum_t e^{-iE_i t/\hbar}|t\rangle\right) \otimes \left(\sum_t e^{-iE_i t/\hbar}|t\rangle\right) \cdots \otimes |\psi_i\rangle. \tag{3}$$

That is, in this case, we get the *same* E_i for each ancilla used. We can put them together to get a single value of E_i out. Which one is random, but they are all at least $E + \Delta$, so will always fail the witness check with probability at least 2/3. □

Note that the fact that the Hamiltonian is local is playing an important role here, as is the accuracy Δ. If the Hamiltonian is arbitrary, we don't have an efficient algorithm to simulate the Hamiltonian and therefore have no way to estimate the energy. If Δ is too small, we don't have a way to efficiently learn the energy to sufficient accuracy to distinguish between the yes and no instances. Also note that we really need a quantum simulation here, unless the ground state has an efficient classical description and we have an efficient classical algorithm to compute the energy with this Hamiltonian of a state provided. This problem is not generally in NP or MA.

2 QMA-Completeness of LOCAL HAMILTONIAN

2.1 Completeness of $O(\log n)$-LOCAL HAMILTONIAN

It is actually OK for the Hamiltonian is $O(\log n)$-local rather than k-local for constant k. That is, the Hamiltonian is a sum of terms, each of which acts on $O(\log n)$ qubits, where n is the total number of qubits in the system. The instance size is polynomial in n, so we could equally have said logarithmic in the instance size. In this case, the "local" matrices are still polynomial in size and there is still an efficient quantum simulation. This is convenient, because it is a bit easier to prove that the $O(\log n)$-LOCAL HAMILTONIAN problem is QMA-complete.

The main technique for proving QMA-completeness of local Hamiltonian problems is the *history state*. The idea is to construct hard subsets of instances by designing a Hamiltonian whose ground state represents the history of some circuit. Given a circuit with T gates, let $|\psi_0\rangle$ be the initial state of the circuit and $|\psi_t\rangle$ $(t = 1, \ldots, T)$ be the state of the circuit after the ith gate. Then, at its simplest, the *history state* of this circuit is the state

$$\sum_{t=0}^{T} |\psi_t\rangle|t\rangle. \tag{4}$$

That is, the history state has two registers, one containing the state of the circuit at time t and the other a *clock register* containing the time, and then we have a superposition over all times.

We can create a Hamiltonian that has history states as ground states by including terms that cause transitions between $|\psi_{t-1}\rangle|t-1\rangle$ and $|\psi_t\rangle|t\rangle$.

Lemma 1. *If a circuit consists of a sequence of gates U_t, $t = 1, \ldots, T$, then the Hamiltonian*

$$H_P = \sum_{t=1}^{T} H_t = \frac{1}{2}\sum_{t=1}^{T} \left[I \otimes (|t\rangle\langle t| + |t-1\rangle\langle t-1|) - U_t \otimes |t\rangle\langle t-1| - U_t^\dagger \otimes |t-1\rangle\langle t|\right] \tag{5}$$

has as 0 *energy states all states of the form* $\sum_{t=0}^{T} |\psi_t\rangle|t\rangle$ *for arbitrary initial states* $|\psi_0\rangle$ *and no other states. The eigenvalues of* H_P *are* $1-\cos(\pi k/(T+1))$ *for* $k = 0, \ldots, T$.

We need the $U_t^\dagger \otimes |t-1\rangle\langle t|)$ term to ensure that the Hamiltonian is Hermitian.

Proof. Let us calculate what the Hamiltonian does to any history state $\sum_{t=0}^{T} |\psi_t\rangle|t\rangle$. First consider the action of a single term:

$$I \otimes |t\rangle\langle t| \left(\sum_{t=0}^{T} |\psi_t\rangle|t\rangle \right) = |\psi_t\rangle|t\rangle \tag{6}$$

$$(U_t \otimes |t\rangle\langle t-1|) \left(\sum_{t=0}^{T} |\psi_t\rangle|t\rangle \right) = U_t|\psi_{t-1}\rangle|t\rangle = |\psi_t\rangle|t\rangle \tag{7}$$

$$(U_t^\dagger \otimes |t-1\rangle\langle t|) \left(\sum_{t=0}^{T} |\psi_t\rangle|t\rangle \right) = U_t^\dagger|\psi_t\rangle|t-1\rangle = |\psi_{t-1}\rangle|t-1\rangle. \tag{8}$$

Then

$$H_P \left(\sum_{t=0}^{T} |\psi_t\rangle|t\rangle \right) = \frac{1}{2} \sum_{t=1}^{T} [(|\psi_t\rangle|t\rangle + |\psi_{t-1}\rangle|t-1\rangle) - |\psi_t\rangle|t\rangle - |\psi_{t-1}\rangle|t-1\rangle] = 0. \tag{9}$$

Thus, all history states are eigenstates of 0 eigenvalue.

To understand the other eigenvalues, for any specific $|\psi_0\rangle$, define a subspace $S_{|\psi_0\rangle}$ spanned by all states of the form

$$|\psi_t\rangle|t\rangle = \prod_{i=t}^{1} U_i|\psi_0\rangle|t\rangle. \tag{10}$$

Then all terms in the Hamiltonian preserve the subspace $S_{|\psi_0\rangle}$ (meaning that if we have a state $|\phi\rangle \in S_{|\psi_0\rangle}$, then for each term H_t in the Hamiltonian, $H_t|\phi\rangle \in S_{|\psi_0\rangle}$). Thus, any eigenvectors can be taken to be in $S_{|\psi_0\rangle}$ for some $|\psi_0\rangle$. (If eigenvectors from two different such subspaces have the same eigenvalue, then a superposition of those two eigenvectors is also an eigenvector, but we can choose a basis of eigenvectors such that each element of the basis is in one subspace $S_{|\psi_0\rangle}$.)

Let us look at $S_{|\psi_0\rangle}$ and consider the basis $\{|v_t\rangle = |\psi_t\rangle|t\rangle\}$ for $t = 0, \ldots, T$. Using this basis, what does the Hamiltonian look like? The diagonal terms are $\frac{1}{2}|v_0\rangle\langle v_0| + \sum_{t=1}^{T-1} |v_t\rangle\langle v_t| + \frac{1}{2}|v_T\rangle\langle v_T|$ (since the middle terms $|v_t\rangle\langle v_t|$ get $\frac{1}{2}$ contributions from H_t and H_{t+1}, whereas the first and last terms only get one such contribution, from either H_1 or H_T). The off-diagonal terms are $-\frac{1}{2}\sum_{t=1}^{T}(|v_t\rangle\langle v_{t-1}| + |v_{t-1}\rangle\langle v_t|)$. We can write this as a matrix too:

$$H_P = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{11}$$

We can recognize this as a quantum walk on a line. A prospective eigenvector will be a wave along the line:

$$|\phi\rangle = \sum_{t=0}^{T} (\alpha e^{i\omega t} + \beta e^{-i\omega t})|v_t\rangle. \tag{12}$$

To see when something of this form is an eigenvector, compute

$$H_P|\phi\rangle = \frac{1}{2}(\alpha+\beta-\alpha e^{i\omega}-\beta e^{-i\omega})|v_0\rangle + \sum_{t=1}^{T-1}(\alpha e^{i\omega t}+\beta e^{-i\omega t})(1-\frac{1}{2}e^{i\omega}-\frac{1}{2}e^{-i\omega})|v_t\rangle$$
$$+\frac{1}{2}\left[\alpha e^{i\omega T}(1-e^{-i\omega})+\beta e^{-i\omega T}(1-e^{i\omega})\right]|v_T\rangle \tag{13}$$
$$= i(-\alpha e^{i\omega/2}+\beta e^{-i\omega/2})\sin(\omega/2)|v_0\rangle + \sum_{t=1}^{T-1}(\alpha e^{i\omega t}+\beta e^{-i\omega t})(1-\cos\omega)|v_t\rangle$$
$$+ i(\alpha e^{i\omega(T-1/2)}-\beta e^{-i\omega(T-1/2)})\sin(\omega/2)|v_T\rangle. \tag{14}$$

We have eigenvalue λ if:

$$i(-\alpha e^{i\omega/2}+\beta e^{-i\omega/2})\sin(\omega/2) = \lambda(\alpha+\beta) \tag{15}$$
$$(\alpha e^{i\omega t}+\beta e^{-i\omega t})(1-\cos\omega) = \lambda(\alpha e^{i\omega t}+\beta e^{-i\omega t}) \tag{16}$$
$$i(\alpha e^{i\omega(T-1/2)}-\beta e^{-i\omega(T-1/2)})\sin(\omega/2) = \lambda(\alpha e^{i\omega T}+\beta e^{-i\omega T}) \tag{17}$$

The equations for $t = 1, \ldots, T-1$ tell us $\lambda = 1-\cos\omega = 2\sin^2(\omega/2)$. Thus, to satisfy the equations for $t = 0$ and $t = T$, we need that

$$\frac{i}{2}(-\alpha e^{i\omega/2}+\beta e^{-i\omega/2}) = \sin(\omega/2)(\alpha+\beta) \tag{18}$$
$$\frac{i}{2}(\alpha e^{i\omega(T-1/2)}-\beta e^{-i\omega(T-1/2)}) = \sin(\omega/2)(\alpha e^{i\omega T}+\beta e^{-i\omega T}). \tag{19}$$

The first can be solved by letting $\alpha = e^{i\omega}\beta$ since then we have

$$e^{i\omega}e^{i\omega/2}-e^{-i\omega/2} = (e^{i\omega/2}-e^{-i\omega/2})(e^{i\omega}+1). \tag{20}$$

The last equation then becomes

$$e^{i\omega(T+1/2)}-e^{-i\omega(T-1/2)} = -(e^{i\omega/2}-e^{-i\omega/2})(e^{i\omega(T+1)}+e^{-i\omega T}) \tag{21}$$
$$= -e^{i\omega(T+3/2)}+e^{i\omega(T+1/2)}-e^{-i\omega(T-1/2)}+e^{-i\omega(T+1/2)}, \tag{22}$$

or

$$e^{i\omega(T+3/2)} = e^{-i\omega(T+1/2)} \tag{23}$$

This means that $2\omega(T+1)$ is an integer multiple of 2π, or $\omega = \pi k/(T+1)$ for integer k.

This gives $T+1$ eigenvalues $1-\cos(\pi k/(T+1))$, which matches the dimension of the subspace, so these are all the eigenvalues. This argument also gives us the explicit eigenstates, and we can see that we don't get a proper history state unless $\omega = 0$; otherwise, we get a history state with some additional phases between terms. □

But how is this related to proving that the local Hamiltonian problem is QMA-complete? We can prove QMA-completeness by reducing from an arbitrary QMA language to $O(\log n)$-LOCAL HAMILTONIAN. What do we know about an arbitrary QMA language? It has an efficient checking circuit. We will therefore pick a Hamiltonian whose ground state is a history state for the language we are reducing from.

Theorem 2. *$O(\log n)$-LOCAL HAMILTONIAN is QMA-complete.*

Proof. We wish to reduce from a language L to $O(\log n)$-LOCAL HAMILTONIAN. Let C be the checking circuit for L for an instance of size n, and assume that the acceptance probability of this checking circuit has been amplified to close to 1 (we will need it to be polynomially close). We will discuss the best ways to do this later. We define a Hamiltonian H as follows:

$$H = H_P + H_i + H_f. \tag{24}$$

Here H_P is the Hamiltonian defined above whose ground states are history states for C.

$H_i = \sum_j |1\rangle_j\langle 1| \otimes |0\rangle\langle 0|$, where the sum is taken over all qubits j which are ancilla states for C and the second tensor factor is the clock register. This ensures that at time 0, the ancilla states for C start out in the $|0\rangle$ state; otherwise they will suffer an energy penalty. Note that we don't put a constraint on all the input qubits to C, since C also takes the witness as input, and we don't know what that is supposed to be.

$H_f = |0\rangle_0\langle 0| \otimes |T\rangle\langle T|$, where the first register is a projector on the output qubit of C and the second tensor factor is the clock register. This term gives an energy penalty if the output qubit of the circuit is 0, indicating that the checking circuit does not accept.

Note that each of H_P, H_i, and H_f are positive semi-definite. This means that the energy of a state for H is at least as large as the energy for each of the three pieces.

This Hamiltonian is $O(\log n)$-local, since each gate for C acts on a constant number of qubits. The clock requires $\log T$ qubits, and $T = O(\text{poly } n)$, so the clock has $O(\log n)$ qubits in it. The Hamiltonian is the sum of terms each of which is an operator on the first register which acts on a constant number of qubits tensor an operator of the form $|t\rangle\langle t|$ or $|t\rangle\langle t'|$, which can potentially act on all qubits in the clock. Thus, each term of the Hamiltonian may act on up to $O(\log n)$ qubits.

Let us compute the energy of a history state $|\psi\rangle = \frac{1}{\sqrt{T+1}} \sum_t |\psi_t\rangle|t\rangle$ for C with correctly-initialized ancillas for some choice of witness. (I have normalized the state correctly since this will be important for our calculation.) $H_P|\psi\rangle = 0$ by lemma 1. Since the initial state $|\psi_0\rangle$ of the circuit has 0 for all ancilla qubits, $H_i|\psi\rangle = 0$ as well. However, H_f does not necessarily give 0 energy because the witness might not be accepted. Suppose the final state $|\psi_T\rangle$ for a particular witness is accepted with probability p. Then $|\psi_T\rangle = \sqrt{p}|1\rangle|\psi_T^1\rangle + \sqrt{1-p}|0\rangle|\psi_T^0\rangle$ and

$$H_f|\psi\rangle = \frac{1}{\sqrt{T+1}} \sum_{t=0}^{T} H_f|\psi_t\rangle|t\rangle \tag{25}$$

$$= \frac{1}{\sqrt{T+1}} (|0\rangle_0\langle 0| \otimes |T\rangle\langle T|)(\sqrt{p}|1\rangle|\psi_T^1\rangle + \sqrt{1-p}|0\rangle|\psi_T^0\rangle) \otimes |T\rangle \tag{26}$$

$$= \frac{1}{\sqrt{T+1}} \sqrt{1-p}|0\rangle|\psi_T^0\rangle \otimes |T\rangle \tag{27}$$

$$\langle\psi|H_f|\psi\rangle = \frac{1-p}{T+1}. \tag{28}$$

Thus, the energy of this state is not 0 unless the witness for L is accepted by C with probability 1. The history state is also not an eigenstate of H unless the witness is accepted with probability 1.

Now, let us consider a "yes" instance of L. Then there exists a witness that is accepted with probability $p \geq 1 - O(1/\text{poly}(n))$. Therefore, we can construct a history state using that witness and the expectation value of its energy for H is $(1-p)/(T+1) \leq O(1/[(T+1)\text{poly}(n)]) = E$. While this history state is not an eigenstate, there must be a ground state and its energy must be less than or equal to that of the history state, so in particular, we see that the ground state energy is at most E.

For a "no" instance of L, we can make a history state out of any witness, but C will always accept with probability $p \leq O(1/\text{poly}(n))$. This implies that the energy of that history state is at least $(1 - O(1/\text{poly}(n))/(T+1) \geq E + c_1/(T+1)$ for some constant c_1. We want this to be a "no" instance of $O(\log n)$-LOCAL HAMILTONIAN, which means that the energy of every eigenstate is at least $E + \Delta$. This is equivalent to saying that all states have their energy at least this big. However, so far we have only checked history states formed with initial states with correctly initialized ancillas. What about states that are not of this form?

For instance, consider a history state $|\psi\rangle$ with one or more ancillas incorrectly initialized to $|1\rangle$ rather than $|0\rangle$. Then

$$\langle\psi|H|\psi\rangle \geq \langle\psi|H_i|\psi\rangle = \frac{1}{T+1} \sum_j \langle\psi_0|1\rangle_j\langle 1|\psi_0\rangle \otimes \langle 0|0\rangle = \frac{d}{T+1}, \tag{29}$$

where d is the number of ancillas initialized to $|1\rangle$ instead of $|0\rangle$. This is greater than E by at least $c_2/(T+1)$ for some constant c_2.

We could also have history states with ancillas initially in a superposition or an entangled state, but these are then superpositions of history states with correct ancillas and $|1\rangle$ ancillas. Let us write this history state as $|\psi\rangle = \alpha|\phi_v\rangle + \beta|\phi_v^\perp\rangle$, where $|\phi_v\rangle$ is a normalized history state with correctly initialized ancillas and $|\phi_v^\perp\rangle$ is a normalized superposition of history states with at least one ancilla qubit initialized to $|1\rangle$. Then

$$\begin{aligned}\langle\psi|H_i|\psi\rangle =&|\alpha|^2\langle\phi_v|H_i|\phi_v\rangle + |\beta|^2\langle\phi_v^\perp|H_i|\phi_v^\perp\rangle + \alpha^*\beta\langle\phi_v|H_i|\phi_v^\perp\rangle + \alpha\beta^*\langle\phi_v^\perp|H_i|\phi_v\rangle &(30)\\ \geq&|\beta|^2(E + c_2/(T+1)), &(31)\end{aligned}$$

where I've used the argument above and the fact that $H_i|\phi_v\rangle = 0$ since the ancillas are initialized properly for $|\phi_v\rangle$.

We can also bound $\langle\psi|H_f|\psi\rangle$, but it is a bit trickier because neither $|\phi_v\rangle$ nor $|\phi_v^\perp\rangle$ has to be an eigenstate of H_f. However, we do know two things. First, $|\phi_v\rangle$ has only a small chance of being accepted, so $H_f|\phi_v\rangle$ is very close to $|\phi_{v,T}\rangle$, a state at time T derived from correctly initialized ancillas. Specifically, we can write $H_f|\phi_v\rangle = |\phi_{v,T}\rangle + \eta|\xi\rangle$, with $\eta = O(1/\text{poly}(n))$. ($|\phi_{v,T}\rangle$ and $|\xi\rangle$ do not have to be orthogonal, but we assume both are normalized.) Moreover, we know that states with orthogonal ancillas at $t = 0$ are orthogonal at all times, so $\langle\phi_{v,T}|\phi_v^\perp\rangle = 0$. Thus,

$$|\langle\phi_v|H_f|\phi_v^\perp\rangle| = \eta|\langle\xi|\phi_v^\perp\rangle| = O(1/\text{poly}(n))/(T+1). \tag{32}$$

(The $1/(T+1)$ comes because $|\xi\rangle$ is only at time T and $|\phi_v^\perp\rangle$ is a history state.) We then get

$$\langle\psi|H_f|\psi\rangle \geq |\alpha|^2(E + c_1/(T+1)) - O(1/\text{poly}(n))/(T+1) \geq |\alpha_2|(E + c_3/(T+1)) \tag{33}$$

for some constant c_3. Let $c = \min(c_2, c_3)$. Then, for any history state $|\psi\rangle$, we have that

$$\langle\psi|(H_i + H_f)|\psi\rangle \geq |\alpha|^2(E + c/(T+1)) + |\beta|^2(E + c/(T+1)) = E + c/(T+1). \tag{34}$$

But what if we don't have a history state at all? Such a state will be penalized by H_P, but how much? We will determine this using a lemma by Kitaev:

Lemma 2. *Let K_1, K_2 be positive semi-definite operators with no shared 0 eigenstates and suppose that all eigenstates of both K_1 and K_2 with nonzero eigenvalue have eigenvalue at least v. Let θ be the minimum angle between $|\psi_1\rangle$ and $|\psi_2\rangle$ where $|\psi_i\rangle$ is an eigenstate of K_i with eigenvalue 0. Then*

$$\langle\phi|(K_1 + K_2)|\phi\rangle \geq 2v\sin^2\theta/2 \tag{35}$$

for all state $|\phi\rangle$.

We will apply this lemma by letting $K_1 = H_P$ and $K_2 = H_i + H_f$. The 0 eigenstates of H_P are the history states. H_i and H_f are positive semi-definite, so any 0 eigenstate of K_2 must be a 0 eigenstate of both H_i and H_f. The history states which are 0 eigenstates of H_i are those with correctly initialized ancillas, but we know that these states are not also 0 eigenstates of H_f, so indeed K_1 and K_2 have no shared 0 eigenstates.

By lemma 1, the smallest non-zero eigenvalue of $K_1 = H_P$ is $1 - \cos(\pi/(T+1)) = \Theta(1/T^2)$. $H_i + H_f$ is easy to diagonalize, and we see that the smallest non-zero eigenvalues are 1.

To determine the angle between 0 eigenstates of K_1 and K_2, let $|\psi_1\rangle$ be a history state, a 0 eigenstate of K_1, and let $|\psi_2\rangle$ be a 0 eigenstate of K_2. Choose $|\psi_1\rangle$ and $|\psi_2\rangle$ that achieve the minimum angle θ. Then

$$\cos\theta = \langle\psi_1|\psi_2\rangle = \langle\psi_1|(I - K_2)|\psi_2\rangle. \tag{36}$$

Now,

$$\|(I - K_2)|\psi_1\rangle\| = \sqrt{\langle\psi_1|(I - K_2)|\psi_1\rangle} \leq \sqrt{1 - (E + c/(T+1))} \tag{37}$$

since K_2 is a projector and $I - K_2$ is a projector onto the orthogonal space. Then

$$\cos\theta \leq \|(I - K_2)|\psi_1\rangle\| \leq 1 - \Omega(1/T) \tag{38}$$

and $\sin^2\theta/2 = \Omega(1/T)$. By lemma 2, we then have that

$$\langle\phi|H|\phi\rangle \geq 2\Theta(1/T^2)\Omega(1/T) = \Omega(1/\text{poly}(n)) = \Delta. \tag{39}$$

Thus, the instance (H, E, Δ) is a "no" instance. It follows that we have a reduction from L to $O(\log n)$-LOCAL HAMILTONIAN for arbitrary $L \in$ QMA, and therefore $O(\log n)$-LOCAL HAMILTONIAN is QMA-complete. □

2.2 Kitaev Lemma

We have now proved that $O(\log n)$-LOCAL HAMILTONIAN is QMA-complete except for lemma 2.

Proof of lemma. Let $|\phi\rangle$ be an arbitrary state and let Π_i be a projector onto the 0 eigenvalue space of K_i. Then

$$\langle\phi|(K_1 + K_2)|\phi\rangle \geq v\langle\phi|[(I - \Pi_1) + (I - \Pi_2)]|\phi\rangle = (2 - \langle\phi|(\Pi_1 + \Pi_2)|\phi\rangle)v. \tag{40}$$

We wish to upper bound $\langle\phi|(\Pi_1 + \Pi_2)|\phi\rangle$.

Let us focus on $|\phi\rangle$ which is an eigenstate of $\Pi_1 + \Pi_2$, $(\Pi_1 + \Pi_2)|\phi\rangle = \lambda|\phi\rangle$. We can write $|\phi\rangle$ as

$$|\phi\rangle = \alpha_i|\eta_i\rangle + \beta_i|\xi_i\rangle \tag{41}$$

for $i = 1, 2$, $|\eta_i\rangle$ in subspace projected on by Π_i, and $|\xi_i\rangle$ a state orthogonal to the subspace projected on by Π_i. Then

$$\langle\phi|(\Pi_1 + \Pi_2)|\phi\rangle = |\alpha_1|^2 + |\alpha_2|^2 = \lambda. \tag{42}$$

We also have

$$\begin{aligned}
\lambda^2 &= (\langle\phi|(\Pi_1 + \Pi_2))((\Pi_1 + \Pi_2)|\phi\rangle) && (43)\\
&= (\alpha_1^*\langle\eta_1| + \alpha_2^*\langle\eta_2|)(\alpha_1|\eta_1\rangle + \alpha_2|\eta_2\rangle) && (44)\\
&= |\alpha_1|^2 + |\alpha_2|^2 + 2\text{Re}(\alpha_1^*\alpha_2\langle\eta_1|\eta_2\rangle) && (45)\\
&= \lambda + 2\text{Re}(\alpha_1^*\alpha_2\langle\eta_1|\eta_2\rangle). && (46)
\end{aligned}$$

Now, we know that

$$|\langle\eta_1|\eta_2\rangle| \leq \cos\theta, \tag{47}$$

since θ is the minimum angle between 0 eigenvectors of K_1 and K_2. Thus,

$$\lambda^2 - \lambda \leq 2|\alpha_1\alpha_2|\cos\theta \leq (|\alpha_1|^2 + |\alpha_2|^2)\cos\theta = \lambda\cos\theta. \tag{48}$$

That is,

$$\lambda \leq 1 + \cos\theta. \tag{49}$$

Plugging this back into eq. (40), we find

$$\langle\phi|(K_1 + K_2)|\phi\rangle \geq (1 - \cos\theta)v = 2v\sin^2\theta/2. \tag{50}$$

□

2.3 Completeness of 5-LOCAL HAMILTONIAN

Recall that for $O(\log n)$-LOCAL HAMILTONIAN, we had a Hamiltonian $H = H_P + H_i + H_f$, where H_P enforces the ground state to be a history state $\sum_t |\psi_t\rangle|t\rangle$ for the checking circuit, H_i enforces the $t = 0$ terms to have correctly initialized ancillas, and H_f gives an energy penalty if the final $t = T$ time is not an accepting state for the check circuit.

The $O(\log n)$ size of the terms in the previous result come purely from the clock. If we can find a better way to encode the clock so that we can identify the time specified by the clock using only a few qubits, we can instead show the completeness of k-LOCAL HAMILTONIAN with constant k.

For instance, we could encode the clock in unary. This is of course a very inefficient way of storing it, but the maximum time T is still polynomial in n, so the number of qubits needed for a unary clock encoding is acceptable. Then the time 0 is $00\ldots0$, time 1 is $10\ldots0$, and time T is $11\ldots1$. The advantage of this encoding is that time t is $1\ldots10\ldots0$, beginning with t 1's, which always differs from the encoding of time $t-1$ in only one place, the tth bit, which is 1 for t and 0 for time $t-1$.

In the Hamiltonian above, we had terms involving $|t\rangle\langle t-1|$. We might want to replace this by just a single qubit operator $|1\rangle\langle 0|$ on the tth qubit. Indeed, this will convert the time $t-1$ to time t, as desired. Unfortunately, it will also convert time s (for $s < t-1$) to something that is not a valid time encoding, so using this replacement will have undesireable side effects. Instead, we need an operator that acts *only* on the times t and $t-1$. The thing that singles out those two times is that they have a 1 at the $(t-1)$th place and a 0 at the $(t+1)$th place; they are the only two times that have those two properties. Thus, if we replace $|t\rangle\langle t-1|$ by $|110\rangle\langle 100|$ for qubits $t-1$, t, and $t+1$, this will only convert time $t-1$ to time t and have no effect on other times. (We might be tempted to just use $|11\rangle\langle 10|$ for qubits $t-1$ and t, but to keep the Hamiltonian hermitian, we need to include also the adjoint of this $|10\rangle\langle 11|$, which will convert any time greater than t into some invalid time encoding.)

Similarly, we can replace $|t\rangle\langle t|$ with $|10\rangle\langle 10|$ on qubits t and $t+1$. Once we have done this everywhere in H_P, we can immediately see that history states using the new time encoding are once again ground states for the new H_P. We can similarly revise H_i and H_f. All terms of these Hamiltonians are now 5-local.

However, closer examination will reveal that $H_P+H_i+H_f$ has 0 energy states that are not history states. In particular, any state where the clock qubits are not in a valid clock encoding, such as $|\psi\rangle|00100\ldots0\rangle$, will not activate any of the terms of H_P, H_i, or H_f, and so have 0 energy. The solution is straightforward: We need an additional constraint term in the Hamiltonian which forces the clock to be in a valid encoding. One property of valid clock encodings is that they never have a 1 to the right of a 0. Any state that is not a valid clock encoding, on the other hand, does have a 1 to the right of a 0. Thus, if we have a Hamiltonian

$$H_C = \sum_i I \otimes |01\rangle_{i,i+1}\langle 01| \tag{51}$$

(where the non-trivial action is on qubits i and $i+1$ of the clock), then the only states that have 0 energy for H_C are those with valid clock encodings.

If we consider the Hamiltonian

$$H = H_P + H_i + H_f + H_C, \tag{52}$$

the argument above for the proof of QMA-completeness of $O(\log n)$-LOCAL HAMILTONIAN shows that in a "yes" instance of L, the history state for a good witness for L has energy less than E, whereas in a "no" instance of L, any state in the subspace spanned by states with valid clock encodings has energy at least $E+\Delta$. We also note that any state in the subspace spanned by states with invalid clock encodings has energy at least 1 (due to H_C), which is greater than $E+\Delta$. The Hamiltonian preserves these two subspaces, so any state which is a superposition over the two of them will have energy at least as big as the lesser energy $E+\Delta$. This proves:

Theorem 3. *5-LOCAL HAMILTONIAN is QMA-complete.*

2.4 Completeness of 2-LOCAL HAMILTONIAN

5-LOCAL HAMILTONIANS are still rather complicated. Can we work with simpler Hamiltonians? By improving our clock construction further, we can prove the completeness of 3-LOCAL HAMILTONIANS (although I will skip most details of the proof). For instance, we could still use a unary clock encoding $|11\cdots10\cdots0\rangle$ but then terms of H_P will simply be

$$\frac{1}{2}\left[I \otimes (|10\rangle_{t,t+1}\langle 10| + |10\rangle_{t-1,t}\langle 10|) - U_t \otimes |1\rangle_t\langle 0| - U_t^\dagger \otimes |0\rangle_t\langle 1|\right], \tag{53}$$

where the subscript indicates which clock qubits are addressed. In particular, the terms with a U involve only the tth clock qubit, changing it from 0 to 1 (or from 1 to 0 for $U^\dagger$). This is just a 3-LOCAL HAMILTONIAN.

Unfortunately, as we saw before, this H_P will involve transitions between valid clock encodings and invalid clock encodings, for instance time 1 $|100\cdots0\rangle$ could transition to $|100010\cdots0\rangle$. But we still have the H_C term, so this invalid encoding will have an energy penalty. The Hamiltonian then becomes *frustrated*, meaning it is not possible to satisfy all Hamiltonian terms at once, and the ground states won't have 0 energy even when we are reducing from a problem where the witness is accepted with 100% probability. The actual ground states will be some compromise between satisfying the modified H_P term and the H_C term. In particular, if we increase the relative strength of H_C compared to the other terms, then the lowest energy states will have to come very close to satisfying H_C and thus have a valid clock encoding. Within this large component of the ground state, the only relevant transitions are ones between consecutive valid clock encodings, and each is associated with doing the appropriate gate U_t. Therefore, the ground state will be close to a valid history state for the circuit, as required.

But this approach will not work for 2-LOCAL HAMILTONIAN, since U_t itself is already 2-local, and we then don't have any room to tie it to advancing the clock. Instead, we use another trick, known as *perturbation theory gadgets*. The idea is that, given any 3-local Hamiltonian H, we can make a 2-local Hamiltonian H' that "looks like" the 3-local Hamiltonian on the low energy states.

Suppose we have a 3-qubit Hamiltonian term of H which is the product of three Paulis, which we will assume without loss of generality to be $H_{123} = X_1X_2X_3$ on qubits 1, 2, and 3. The new Hamiltonian will have the original 3 qubits but also 3 additional qubits A, B, and C to tie together the qubits 1, 2, and 3. In particular, we will have terms in H':

$$H'_{ABC} = 3I - Z_AZ_B - Z_AZ_C - Z_BZ_C. \tag{54}$$

This will force all three qubits to be in the same state in the standard basis, either $|000\rangle_{ABC}$ or $|111\rangle_{ABC}$ (or some superposition). Then we also have a term

$$H'_{123} = X_1X_A + X_2X_B + X_3X_C. \tag{55}$$

What are the low-energy states of the sum $H'_{123} + \lambda H'_{ABC}$ when λ is large? The first priority to have low energy is to satisfy H'_{ABC} as much as possible; H'_{123} is a small perturbation. Thus, the low energy states should be close to superpositions of $|000\rangle_{ABC}$ and $|111\rangle_{ABC}$. But within this subspace, states that do a better job of satisfying H'_{123} will be lower energy than ones which don't. However, a single term of H'_{123} applied to such a state will give us a state that violates H'_{ABC}, so is severely penalized. However, there are terms $X_1X_AX_2X_BX_3X_C$ that are the product of 3 terms of H'_{123} that respect H'_{ABC} and so are allowed in the low-energy subspace. But notice: this term includes $X_1X_2X_3$, the original term H_{123} from H! The best attempt to satisfy H'_{123} within the low-energy space of H'_{ABC} is to simulate the H_{123} term.

One can show that by replacing each 3-local product term in H with a 2-local Hamiltonian using these perturbation theory gadgets, the resulting Hamiltonian has low-energy states which map to the low-energy states of H. This in turn shows that it is possible to reduce 3-LOCAL HAMILTONIAN to 2-LOCAL HAMILTONIAN, which must also be QMA-complete.

Naturally, 1-LOCAL HAMILTONIAN is not going to be QMA-complete. It is just a sum of single-qubit terms, and each can be separately diagonalized in constant time. The ground state will just be a tensor product of the ground states of the single-qubit Hamiltonians, and therefore the ground state energy can be easily computed in P.

Note, though, that more general Hamiltonians which are a sum of commuting terms are not necessarily easy, although probably they are not QMA-complete. For one thing, k-SAT can be phrased as a commuting Hamiltonian where each clause can be enforced through a diagonal Hamiltonian term on k qubits. Thus, commuting Hamiltonians are NP-hard. However, it's not even clear that the general k-local commuting Hamiltonian problem is in NP. (This is an open problem.) Again, all terms of the Hamiltonian can be simultaneously diagonalized, but the eigenstates might be highly entangled states. (For instance, stabilizer states, which we saw earlier, are eigenstates of commuting Hamiltonians formed from their stabilizers.) Therefore, classically specifying a single one of these to be the witness might be difficult.

In general, when dealing with Hamiltonian complexity, one can consider different kinds of Hamiltonians as we have seen above. The single-qubit (or single-qudit) Hamiltonians are essentially trivial. Then commuting Hamiltonians are more complicated but still easier to deal with than the non-commuting but unfrustrated Hamiltonians (as in the $O(\log n)$-LOCAL HAMILTONIAN and 5-LOCAL HAMILTONIAN problems), which in turn are easier to deal with than general frustrated Hamiltonians, as with the 3-LOCAL HAMILTONIAN and 2-LOCAL HAMILTONIAN problems. That is, it is easier to prove things about the "less complicated" Hamiltonian types on this list; but of course looking at a wider range of possible Hamiltonians may enable you to find examples that have some particular interesting property.

2.5 Further variations of LOCAL HAMILTONIAN

We now have a full classification of the complexity of the k-LOCAL HAMILTONIAN problem for different values of k. But because k-local doesn't correspond to geometric locality, this is still far from the more physically interesting Hamiltonians. What about Hamiltonians that are constrained to short-range or nearest-neighbor interactions in a small number of spatial dimensions?

It turns out that these problems can be QMA-complete as well. To prove this, we need to continue to work on the clock. We can use a checking circuit that can be laid out in 1 dimension without much difficulty — use SWAP gates to move qubits around as needed and then back, which only introduces polynomial overhead. But if all qubits need to interact with the clock, which would be stored at a particular location, that would introduce violations of geometric locality. Instead, the usual approach is to find some way of encoding the clock in the spatial position of some counter that moves around between the qubits and performs the appropriate gates at the appropriate "time." Typically, the strategy is to come up with some sort of classical process that can be implemented via geometrically local transition rules and geometrically local constraints and then encode the history of that with a geometrically local Hamiltonian.

This usually involves increasing the local dimension — the size of the individual qudits in the system. In 2 spatial dimensions, the local dimension can be reduced again to 2 (qubits) using perturbation theory gadgets, showing that 2-LOCAL HAMILTONIAN is QMA-complete using Hamiltonians with nearest-neighbor interactions in 2 spatial dimensions. In 1 spatial dimension, there is not enough room to apply the perturbation theory gadgets, so the problem is open for 2-local qubit Hamiltonians in 1 spatial dimension, but we know that 2-local Hamiltonians using large enough qudits in 1 spatial dimension are QMA-complete. (I believe that the best current result is for a local dimension of 8.)

Note that for classical SAT-type problems (including MAX-k-SAT and variations with more than two possible states), the 1-dimensional problem is in P. (It can be solved by dynamic programming.) The 2-dimensional version is NP-complete again. The superposition in a quantum state in some ways takes the place of a spatial dimension.

Even so, these Hamiltonians are not that physically realistic since they involve highly structured but different terms in different locations. One can go even further and focus on translationally-invariant Hamiltonians instead, which use the same Hamiltonian term at every site, translated appropriately throughout the system. However, there is an immediate complexity issue, which is that specifying a single Hamiltonian term is just constant size. The only thing that gives us a family of problems of different sizes is the number of qudits in the system. This is the sort of thing that complexity theorists would normally want to specify in unary. However, in that case, there is just one problem instance of any given size, so the problem is automatically solvable in P/poly since the advice for a given size can just specify the answer to the single instance of that size.

The solution to still have an interesting complexity problem is to instead specify the number of qudits in binary, meaning an n-qudit instance has size $\log n$. However, measuring the energy takes a time polynomial in n, which is exponential in $\log n$; the problem is no longer in QMA. Instead, it is in a class $\mathrm{QMA}_{\mathrm{EXP}}$, which is like QMA, but has exponentially large checking circuits. In fact, using similar ideas to those we have discussed above, one can show that 1-dimensional translationally-invariant Hamiltonians with large enough qudits are $\mathrm{QMA}_{\mathrm{EXP}}$-complete. The idea is to use a quantum Turing machine (which has translation-invariant transition rules) and encode the instance in the number of qudits n, which becomes the input to the Turing

machine. Unfortunately, this whole process uses incredibly large (but constant size) qudits, so the existing constructions are still not at all realistic.

Many other ground state properties are QMA-complete or QMA-hard, basically anything that is measurable on a ground state is a good candidate. In some cases, the problems can be even harder. For instance, it turns out to be undecideable to determine if a system has a spectral gap as the number of particles goes to infinity (i.e., if the limit of the difference between the lowest two eigenvalues is non-zero).

3 Other QMA-complete problems

3.1 Density Matrix Consistency

Another related QMA-complete problem is *density matrix consistency* and variations.

Definition 4. *Let $\mathcal{S}$ be a set of subsets $\{S_\mu\}$, where each S_μ is a set of at most k qubits. If ρ is a density matrix, then let $\rho(\mu) = \mathrm{Tr}_{i \notin S_\mu} \rho$ (i.e., ρ on only the qubits in S_μ). Then k-CONSISTENCY is a language consisting of instances $(\mathcal{S}, \{\rho_\mu\}, \epsilon)$ (with all quantities specified to polynomially many bits of precision) with the promise that either:*

1. *$\exists$ density matrix ρ such that $\rho(\mu) = \rho_\mu$ ("yes" instances)*
2. *$\forall$ density matrices σ, $\|\sigma(\mu) - \rho_\mu\|_1 \geq \epsilon$ for some μ. ("no" instances).*

In other words, the k-CONSISTENCY problem asks if there is a *global* density matrix consistent with a set of local density matrices we are provided with. It might not be clear why this is even a question, let alone why it might be hard. Certainly if $k = 1$, so all the local density matrices are single-qubit ones, then there is always a global density matrix consistent with them, the tensor product of the local values. Once $k \geq 2$, however, the local density matrices can be specified on overlapping sets of qubits, and that offers a possibility for inconsistency. Sometimes this is trivial and obvious, such as if $\rho_{12} = |00\rangle\langle 00|$ and $\rho_{23} = |11\rangle\langle 11|$: Qubit 2 can't be both $|0\rangle$ (as required by ρ_{12}) and $|1\rangle$ (as required by ρ_{23}). Sometimes it is less obvious, depending on the global structure of the density matrices. For instance, suppose $\rho_{i,i+1} = \frac{1}{2}(|01\rangle\langle 01| + |10\rangle\langle 10|)$, and qubit N is the same as qubit 0. That is, adjacent qubits on a ring of N qubits have the opposite value. This is possible when N is even but not when N is odd. Even more subtle global structures are possible.

However, this problem is in QMA. The basic idea is that the witness is the global density matrix ρ that is consistent with all ρ_μ. To verify, choose a set S_μ at random and make a random measurement Π on $\rho(\mu)$. Accept if the measurement outcome is 1; reject otherwise. If the witness is a good one, it will be accepted with probability $\mathrm{Tr}\,\Pi\rho_\mu$, a known quantity, and if it is not a good witness (in the sense of condition 2 of the definition of k-CONSISTENCY), there is a polynomially small chance we have picked a μ such that $\|\rho(\mu) - \rho_\mu\|_1 \geq \epsilon$, in which case, there is a constant probability that we pick a measurement for which $\mathrm{Tr}\,\Pi\rho_\mu$ is significantly different from $\mathrm{Tr}\,\Pi\rho(\mu)$. Overall, there is thus a polynomially small difference in acceptance probability between the "yes" and "no" instances. This isn't the 2/3 vs. 1/3 difference we require for the usual definition of QMA, but it can be turned into that through the amplification techniques we will discuss later.

I won't prove QMA-completeness of this problem, but we can understand qualitatively the relationship with LOCAL HAMILTONIAN: Let ρ be some ground state of a k-local Hamiltonian H (either a pure ground state or a mixed state which is a mixture of degenerate ground states). Since H is k-local, it is a sum $\sum_\mu H_\mu$, where H_μ acts on the subset S_μ of qubits. Then the ground state energy of H is

$$\mathrm{Tr}(\rho H) = \sum_\mu \mathrm{Tr}(H_\mu \rho) = \sum_\mu \mathrm{Tr}(H_\mu \rho(\mu)). \tag{56}$$

That is, the ground state energy depends on only the local density matrices $\rho(\mu)$. One way we might try to find the ground state energy is by separately optimizing $\mathrm{Tr}(H_\mu \rho_\mu)$ for each μ, but once we've done that,

we need to make sure that the separate ρ_μ we have found are all consistent with a single global ρ — the k-CONSISTENCY problem.

In fact, this is not good enough, since usually separately optimizing the terms $\mathrm{Tr}(H_\mu \rho_\mu)$ *doesn't* give us a globally valid state. There might be multiple solutions for each term, leaving ambiguity about which set of choices we should make (and there are exponentially many sets of choices if each subset of qubits has 2 or more choices), or it might be that the Hamiltonian is frustrated and it is instead a question of compromising by violating each term as little as possible. However, the overall goal, to minimize $\sum_\mu \mathrm{Tr}(H_\mu \rho_\mu)$, is a *convex optimization* problem, and there are good algorithms for that. The wrinkle here is that we must do so under the constraint that the ρ_μ are globally consistent, but if we somehow have an oracle for the k-CONSISTENCY problem, we are still able to run this convex optimization algorithm and find the local density matrices of that Hamiltonian's ground state. Thus, if we can solve k-CONSISTENCY, we can solve k-LOCAL HAMILTONIAN and therefore any problem in QMA when $k \geq 2$.

Note that this reduction is a *Turing reduction* rather than the *Karp reductions* we discussed before: we run a general program which makes use of the hypothetical subroutine to solve k-CONSISTENCY, as opposed to converting instances of k-LOCAL HAMILTONIAN to instances of k-CONSISTENCY. However, there is also a different reduction of essentially the same problem has been found which is a Karp reduction.

3.2 Circuit Non-Equivalence

Another type of QMA-complete problem are properties of quantum circuits. One good example is *circuit non-equivalence*, in which we are given two circuits and wish to determine if they compute the same unitary or not:

Definition 5. *The CIRCUIT NON-EQUIVALENCE problem consists of instances of the form (C_1, C_2, a, b) where C_1 and C_2 are polynomial-size quantum circuits on n qubits (involving efficiently computable gates) computing the unitaries U_1 and U_2 respectively, and $b - a \geq \Omega(1/\mathrm{poly}(n))$ with the following promise: Either*

1. *$\forall \phi$, $\|U_1 - e^{i\phi} U_2\|_{sup} \geq b$ ("yes" instances)*
2. *$\exists \phi$ such that $\|U_1 - e^{i\phi} U_2\|_{sup} \leq a$ ("no" instances).*

It is not surprising that this is in QMA: The witness is simply a few copies of a state $|\psi\rangle$ that distinguishes the two unitaries, one for which $\|U_1|\psi\rangle - e^{i\phi} U_2|\psi\rangle\| \geq b$. There is in this case a measurement that distinguishes the states $U_i|\psi\rangle$, so to verify the witness, just run the two circuits on $|\psi\rangle$ and make the measurement to see difference results. The complication is again that you really need multiple copies to run both circuits and see the difference in statistics, but this can be dealt with.

This problem is QMA-complete. In fact, the special case NON-IDENTITY CHECK, where one of the circuits is the identity, is also QMA-complete. It is not hard to show this directly from the definition of QMA: Let V be the verification circuit for some QMA language, amplified so that the success probability is very high. Then we run V, copy the output qubit to an ancilla, then run $V^\dagger$. This is C_1. C_2 is just the identity. For a "yes" instance, there exists some input state (a valid witness) such that the overall effect is to flip the ancilla qubit, which is definitely not the identity. For the "no" instance, the ancilla isn't changed and then $V^\dagger$ inverts V, returning the system to its initial state, an overall identity operation.

4 QCMA

Recall that QCMA is like QMA, but with the difference that the witness must be a classical string instead of a quantum state. (However, the verification circuit is still quantum, which is what distinguishes QCMA from MA.) I think it is fair to say that most people believe that QCMA $\neq$ QMA, but there is not overwhelming evidence of this. There are known oracle separations between QCMA and QMA, but of course that is far from definitive. Mostly I would say this belief is based on the understanding that quantum states can

be very complicated and providing one is potentially much more powerful than providing simply classical information.

If we take a QMA-complete problem and add extra promises to make sure that the witness is classically describable, we tend to get a QCMA-complete problem. For instance, if we consider k-local Hamiltonians with the additional promise that the ground state can be created by a polynomial-size quantum circuit, then it is a QCMA-complete problem. Or we could use NON-IDENTITY CHECK with the extra constraint that U_1 acts differently from the identity on some basis state input.

5 Witness Amplification

We have been using the fact that we can amplify the acceptance probability of a QMA witness. The analogous property for classical MA is quite straightfoward: If we have a randomized algorithm to test the witness w with probability p, we can simply run the algorithm multiple times and take the majority result. The probability of getting the right answer for a single test is independent over multiple runs of the algorithm, so the chance of a majority of the trials giving the wrong answer is small.

For QMA, the roadblock to this strategy is that the witness is used up the first time we run a checking circuit. If we want to amplify the success probability by repeating, the most obvious way to do so is to require Merlin to give us multiple copies of the witness, and then we can run the checking circuit multiple times. This works, but does require some additional proof to verify that Merlin cannot cheat by giving us a large entangled state instead of the tensor product of witnesses.

However, surprisingly, there is a way to magnify the acceptance probability of a witness *without* having to increase the size of the witness. The main idea is to run the checking circuit, measure or otherwise record the output, and then reverse the checking circuit and see if the ancilla qubits return to their initial $|00\cdots0\rangle$ state.

In particular, imagine that we start with the state $|\psi\rangle \otimes |00\cdots0\rangle$ (the second register being ancilla qubits needed for the checking circuit), perform a unitary checking circuit C, concluding with a measurement, which we can assume is a projective measurement of a single qubit. Let $M_i = |i\rangle\langle i| \otimes I$ be a projection onto the measured final qubit having value i and let $P_0 = I \otimes |00\cdots0\rangle\langle 00\cdots0|$, $P_1 = I - P_0$ be projections onto the subspace with ancillas all 0 and the orthogonal subspace.

Consider the operators

$$A_{ij} = P_j C^\dagger M_i C. \tag{57}$$

(Note that these are generally not projectors.) If the checking circuit is accepted with 100% probability, then

$$A_{10}|\psi\rangle \otimes |00\cdots0\rangle = |\psi\rangle \otimes |00\cdots0\rangle \tag{58}$$

$$A_{00}|\psi\rangle \otimes |00\cdots0\rangle = 0 \tag{59}$$

since the projection M_1 is trivial and M_0 annihilates the state. Conversely, if the checking is always rejected,

$$A_{00}|\psi\rangle \otimes |00\cdots0\rangle = |\psi\rangle \otimes |00\cdots0\rangle \tag{60}$$

$$A_{10}|\psi\rangle \otimes |00\cdots0\rangle = 0. \tag{61}$$

Note that in both cases, for the outcome that can actually occur, we get the original witness back by reversing C.

In the case where the circuit is accepted with high probability, but not quite 1, we don't always succeed in getting back to $|\psi\rangle \otimes |00\cdots0\rangle$, but we can get close. Instead, however, let us first focus on eigenstates of the operators A_{i0}, albeit ones with eigenvalues less than 1. Thus, for instance, we have

$$A_{i0}|\psi\rangle \otimes |00\cdots0\rangle = \lambda_i |\psi\rangle \otimes |00\cdots0\rangle. \tag{62}$$

Note that any eigenstate with non-zero eigenvalue of A_{i0} must be of the form $|\psi\rangle \otimes |00\cdots 0\rangle$ since A_{i0} ends with a projector onto such states and that since $A_{00} + A_{10} = P_0$, any eigenvector of A_{10} will also be an eigenvector of A_{00}. Furthermore,

$$(\langle\psi| \otimes \langle 00\cdots 0|)A_{i0}(|\psi\rangle \otimes |00\cdots 0\rangle) = (\langle\psi| \otimes \langle 00\cdots 0|)C^\dagger M_i C(|\psi\rangle \otimes |00\cdots 0\rangle), \tag{63}$$

which is p_i, the probability that the circuit C gives outcome i on input $|\psi\rangle$. Thus, $\lambda_i = p_i$.

Suppose we actually run the procedure implied by A_{i0}. That is, run C, measure the output qubit to have value i (which occurs with probability p_i) and then run $C^\dagger$ and measure to see if the ancilla qubits have been reset to all 0. What is the overall probability that this happens, including the output qubit having value i? It is

$$\begin{aligned}
(\langle\psi| \otimes \langle 00\cdots 0|)A_{i0}^\dagger A_{i0}(|\psi\rangle \otimes |00\cdots 0\rangle) &= (\langle\psi| \otimes \langle 00\cdots 0|)C^\dagger M_i C P_0 C^\dagger M_i C(|\psi\rangle \otimes |00\cdots 0\rangle) && (64)\\
&= \sum_\ell |(\langle\psi_\ell| \otimes \langle 00\cdots 0|)C^\dagger M_i C(|\psi\rangle \otimes |00\cdots 0\rangle)|^2 && (65)\\
&= |(\langle\psi| \otimes \langle 00\cdots 0|)C^\dagger M_i C(|\psi\rangle \otimes |00\cdots 0\rangle)|^2 && (66)\\
&= p_i^2. && (67)
\end{aligned}$$

Here, $|\psi_\ell\rangle$ runs over a basis of possible witness states including $|\psi\rangle$ and the third line follows because $|\psi\rangle|00\cdots 0\rangle$ is an eigenstate of A_{i0}.

Thus, the conditional probability of the ancillas being 0 conditioned on having outcome bit i is p_i. Also, note that for eigenstates $|\psi\rangle$, if we do happen to get the ancillas being in the state all 0, then we have restored the ancilla and can just repeat the process on the same state to get more data as to the values of p_0 and p_1.

But what about if we find the ancillas are *not* all 0? In that case, we are instead performing A_{i1}. We know that the conditional probability of ancillas not being 0 conditioned on the output qubit being i is $p_{1\oplus i}$, but what is the final state? It is not generally going to be $|\psi\rangle$, which might pose a problem for repeating the measurement.

Let $|\phi_i\rangle = A_{i1}|\psi\rangle \otimes |00\cdots 0\rangle$. Note that $\langle\phi_i|\phi_i\rangle = p_i(1-p_i) = p_0 p_1$ because $P_1 = I - P_0$. Suppose we go ahead and perform C on $|\phi_i\rangle$ and measure the output qubit. What state do we get if we get outcome k?

$$\begin{aligned}
M_k C A_{i1}|\psi\rangle \otimes |00\cdots 0\rangle &= M_k C(I - P_0)C^\dagger M_i C|\psi\rangle \otimes |00\cdots 0\rangle && (68)\\
&= M_k M_i C|\psi\rangle \otimes |00\cdots 0\rangle - M_k C A_{i0}|\psi\rangle \otimes |00\cdots 0\rangle && (69)\\
&= \delta_{ik} M_k C|\psi\rangle \otimes |00\cdots 0\rangle - p_i M_k C|\psi\rangle \otimes |00\cdots 0\rangle. && (70)
\end{aligned}$$

That is, whether $i = k$ or not, we get $M_k C|\psi\rangle \otimes |00\cdots 0\rangle$, the same state we get from the correct initial state on outcome probability k. If $i \neq k$, the normalization is p_i and the overall probability of this sequence of results is $p_i^2 p_k$. If $i = k$, the normalization is $1 - p_i = p_{i\oplus 1}$ and the overall probability of this sequence of results is $p_i p_{i\oplus 1}^2$. Since the probability of getting i and then P_1 is $p_0 p_1$, in either case, the conditional probability of getting outcome k is $p_{k\oplus 1}$. That is, the probabilities are reversed from the original circuit: outcome 1 happens with probability p_0 and outcome 0 happens with probability p_1! And this is true regardless of what the measurement outcome was the first time we ran C.

One consequence is that after running C and measuring the output qubit in value k, we always get the same state, which depends on k but not on whether we started with $|\psi\rangle \otimes |00\cdots 0\rangle$ or $|\phi_i\rangle$ (for either i). And the probability of getting k didn't depend on whether we started with $|\phi_0\rangle$ or $|\phi_1\rangle$, only on whether the previous measurement result was P_0 or P_1. If we continue alternating C followed by a measurement of M_i and then $C^\dagger$ followed by measurement of P_j, the probability of getting outcome i or j only depends on the previous measurement outcome and not on any earlier ones.

Moreover, the pattern is very simple: After measuring P_j, we get the following conditional probabilities:

$$\begin{aligned}
&\text{Prob}(i = 0|j = 0) = p_0\\
&\text{Prob}(i = 1|j = 0) = p_1\\
&\text{Prob}(i = 0|j = 1) = p_1\\
&\text{Prob}(i = 0|j = 1) = p_0.
\end{aligned}$$

That is, $\text{Prob}(i = j) = p_0$ and $\text{Prob}(i \neq j) = p_1$. Similarly, after measuring M_i, we get conditional probabilities:

$$\begin{aligned}
\text{Prob}(j = 0|i = 0) &= p_0 \\
\text{Prob}(j = 0|i = 1) &= p_1 \\
\text{Prob}(j = 1|i = 0) &= p_1 \\
\text{Prob}(j = 1|i = 1) &= p_0.
\end{aligned}$$

Once again, $\text{Prob}(j = i) = p_0$ and $\text{Prob}(j \neq i) = p_1$. Each outcome ($i$ or j) has the same value as the previous one with probability p_0 and has the opposite value with probability p_1. When we repeat this a number of times, we get a good estimate of p_0 and p_1 by counting the number of changes of outcome. This will enable us to determine if $p_0 \geq 2/3$ or $\leq 1/3$ with high probability.

This analysis was done under the assumption that $|\psi\rangle \otimes |00\cdots0\rangle$ was an eigenvector of A_{i0}. Since $A_{00} + A_{10} = P_0$, the projector on the space $|\psi\rangle \otimes |00\cdots0\rangle$, we can choose a basis of eigenvectors $|\psi_\alpha\rangle$ of A_{i0} in the subspace $|\psi\rangle \otimes |00\cdots0\rangle$ and write any witness state as a superposition of eigenvectors. Since $A_{i0}|\psi_\alpha\rangle = p_i|\psi_\alpha\rangle$ and $A_{k0}A_{i1}|\psi_\alpha\rangle \propto |\psi_\alpha\rangle$ as well, the different eigenvectors of A_{i0} don't get mixed up under repeatedly applying A_{ij}. Thus, if the initial state is a superposition of different eigenvectors, each of them acts independently, meaning we see a sequence of measurement outcomes as if we had just a single eigenvector chosen randomly from the superposition.

6 Quantum PCP

A major result of classical complexity theory is the PCP theorem (for "probabilistically checkable proofs"), which states that every language in NP has a verification procedure (with appropriate witness) that only checks the witness in a constant number of locations. One application of the PCP theorem is to show that many NP-complete problems are hard to even approximate: For instance, suppose you have a k-SAT instance. Recognizing that finding a satisfying instance is going to be hard, you might instead wish to find a solution that satisfies a fraction f of all the clauses in the problem. However, PCP shows that this is not possible if $f > 7/8$, or at least not unless P = NP.

The analogous statement for QMA is open, and is one of the biggest open problems in the field. The Quantum PCP conjecture is most often phrased in terms of a local-Hamiltonian problem:

Conjecture 1. *The following language is QMA-complete: Instances of the form (H, E, Δ), for an n-qubit system, with quantities specified to polynomial accuracy, H a k-local Hamiltonian (for constant k), $\Delta = O(1)$ with the promise that either:*

1. *The ground state energy of H is at most En ("yes" instance)*
2. *The ground state energy of H is at least $(E + \Delta)n$ ("no" instance).*

Note that this is just a standard k-LOCAL HAMILTONIAN problem, but we are requiring that the promise gap is *linear in n*. (As opposed to $O(1/\text{poly}(n))$ as before.)

Again, this is an open problem; the above is a conjecture, and it is really not clear if it is true or not. There's no real evidence one way or the other. Note, though, that the hard-to-approximate instances of SAT are special cases of the language above with a Hamiltonian diagonal in the standard basis, so the language is at least NP-hard.

If the quantum PCP conjecture is true, that means that the ground state energy *per particle* is hard to approximate for some systems to even constant accuracy, let alone accuracy increasing with any parameter. It also implies that it is hard to find the thermal state for some systems at temperature greater than 0. To understand this statement, note that thermodynamics tells us that at thermal equilibrium, we have a mixture of states with different energies. The relative population of a particular eigenstate with energy E is

$$\text{Prob}(E) \propto e^{-\beta E}, \tag{71}$$

where $\beta = 1/kT$. T is the temperature in Kelvin and k is known as Boltzmann's constant. To get the actual population for this particular eigenstate, divide the above by the *partition function* $Z = \sum e^{-\beta E}$.

In particular, as $T \to 0$, the population is dominated by the ground state, whereas as $T \to \infty$, all states are equally populated. A thermal state is one that satisfies the correct distribution (71). Therefore at infinite temperature, finding the thermal state is easy: It is simply the maximally mixed state. However, at 0 temperature, for at least some systems, it is hard since the thermal state is just the ground state.

What about non-zero but low temperature? One might think that the thermal state will have a large admixture of the ground state, and that therefore this case will be hard as well: If you could create the thermal state, then with some high probability you could also create the ground state by simply measuring the energy of a thermal state, repeating polynomially many times until you find a solution.

However, there is one additional factor to take into account, which is the degeneracy of low-energy states. (71) only gives the weighting for a *single* state. But if we have n_E states of energy E, *each* of them has relative population $e^{-\beta E}$, meaning the total population of energy E states is

$$\left(\sum_E n_E e^{-\beta E} \right) / Z. \tag{72}$$

While the $e^{-\beta E}$ term decreases exponentially with E, it is possible (and often will be true) that the n_E term increases more than exponentially with E (in particular, n_E can scale with the size of the system while E does not). If this is the case, then at any non-zero temperature T, the ground state is overwhelmed by the much larger numbers of higher-energy states. This means that measuring the energy in a thermal state of this sort is unlikely to produce ground states. That route to trying to show hardness of the thermal state hits a dead end.

However, if the Hamiltonian satisfies the conditions and results of the quantum PCP conjecture, then finding the thermal state is QMA-hard. We will show that for a Hamiltonian satisfying the promise of the quantum PCP conjecture, the LOCAL HAMILTONIAN problem reduces to finding the energy of a thermal state for that Hamiltonian. Under the quantum PCP conjecture, the k-LOCAL HAMILTONIAN with this promise would be QMA-complete, which means that finding the energy of a thermal state would also be QMA-complete.

Suppose you have an algorithm to measure the energy of the thermal state. Under the quantum PCP promise, this tells you about the ground state energy: If it is a "no" instance of the LOCAL HAMILTONIAN problem, then all eigenstates have energy at least $(E+\Delta)n$, and certainly we will get at least this value when we measure the energy of the thermal state. On the other hand, if we have a "yes" instance of the LOCAL HAMILTONIAN problem, then the ground state has energy En, which we can shift to 0, so its relative probability is 1. We don't know the distribution of higher-energy states, and there may be some or many with energies below $(E + \Delta)n$. There will also be some or many with energies above $(E + \Delta)n$. However, we know that the probability of such states in the thermal state is at most $n_E e^{-\beta(E+\Delta)n}$. Even if $n_E \propto 2^n$ (the total number of states), the overall population with energy at least $(E + \Delta)n$ is at most $e^{(2-\beta(E-\Delta))n}$, which for low enough temperature (large enough β), will always be negligible. Thus, the thermal state will be dominated by states with energy below $(E + \Delta)n$. This means that the energy of the thermal state will certainly be lower than $(E+\Delta)n$. Consequently, if we accept whenever the measured energy for the witness is less than $(E + \Delta)n$ and reject otherwise, we will have answered the instance of the local Hamiltonian problem.

7 References

For the local Hamiltonian problem, see chapter 14 of Kitaev, Shen, and Vyalyi, *Classical and Quantum Computation*.

For QMA-completeness of the local Hamiltonian problem, see chapter 14 of Kitaev, Shen, and Vyalyi, *Classical and Quantum Computation*.

For QMA-completeness of the 5-LOCAL HAMILTONIAN problem, see chapter 14 of Kitaev, Shen, and Vyalyi, *Classical and Quantum Computation*. For QMA-completeness of the 3-LOCAL HAMILTONIAN problem, see Kempe and Regev, "3-Local Hamiltonian is QMA-complete," quant-ph/0302079. For QMA-completeness of the 2-LOCAL HAMILTONIAN problem, see Kempe, Kitaev, and Regev, "The Complexity of the Local Hamiltonian Problem," quant-ph/0406180.

A survey of QMA-complete problems (as of 2012) is given in Bookatz, "QMA-complete problems," arXiv:1212.6312.

Hamiltonians in 2 spatial dimensions are discussed in Aharonov, van Dam, Kempe, Landau, Lloyd, and Regev, "Adiabatic Quantum Computation is Equivalent to Standard Quantum Computation," quant-ph/0405098 and Oliveira and Terhal, "The complexity of quantum spin systems on a two-dimensional square lattice," quant-ph/0504050. Hamiltonians in 1 spatial dimension are discussed in Aharonov, Gottesman, Irani, and Kempe, "The power of quantum systems on a line," arXiv:0705.4077 and Hallgren, Nagaj, and Narayanaswami, "The Local Hamiltonian problem on a line with eight states is QMA-complete," arXiv:1312.1469. Translationally-invariant Hamiltonians are discussed in Gottesman and Irani, "The Quantum and Classical Complexity of Translationally Invariant Tiling and Hamiltonian Problems," arXiv:0905.2419.

Consistency of local density matrices was proven QMA-complete in Liu, "Consistency of Local Density Matrices is QMA-complete," quant-ph/0604166. (This was proved for Karp reductions in Broadbent and Grilo, "QMA-hardness of Consistency of Local Density Matrices with Applications to Quantum Zero-Knowledge," arXiv:1911.07782. Inequivalence of circuits was proven QMA-complete by Janzing, Wocjan, and Beth, "Non-identity check is QMA-complete," quant-ph/0305050.

You can read more about the Quantum PCP conjecture in Aharonov, Arad, and Vidick, "The Quantum PCP Conjecture," arXiv:1309.7495, although there has been some progress since then, most notably the recent resolution of the NLTS conjecture. Witness amplification was introduced in Marriott and Watrous, "Quantum Arthur-Merlin Games," cs/0506068. Vidick and Watrous, "Quantum Proofs," arXiv:1610.01664 is a good reference for QMA and quantum interactive proofs (which we will discuss later).