Lecture 20: Local Hamiltonian ground state problems

Substitute lecturer: Richard Kueng, rkueng@caltech.edu

Ph219/CS219, Fall 2019
John Preskill
December 4, 2019

1 Agenda and Motivation

1. Motivation
2. Recapitulation: classical complexity
   (a) The problem class NP
   (b) Circuit-SAT reduces to 3-SAT
3. Classical local Hamiltonian ground state problems are NP-hard
4. The quantum Hamiltonian problem and the problem class QMA
   (a) The quantum k-local Hamiltonian problem
   (b) The problem class QMA

   The quantum Hamiltonian Ground state problem is QMA-complete
   (a) k-local Hamiltonian is in QMA
   (b) Roadmap
   (c) Constructing the Hamiltonian
   (d) Converting completeness into a small ground state energy
   (e) Converting soundness into a large ground state energy
   (f) Establishing locality
5. Summary

The task of computing ground state energy of structured Hamiltonians is an important problem in physics with numerous applications ranging from solid state physics to ab initio quantum chemistry. We actually expect that quantum computers can compute such energies even in cases where classical computers may struggle. In fact, this might be one of the most important applications of future quantum computers. On the other hand, we believe that even quantum computers may struggle to compute ground state energies in some cases – even if the Hamiltonian is structured. Today, we will explore the reasons for such beliefs.

A comment on structured Hamiltonians Physicists are often interested in translation-invariant geometrically local Hamiltonians, where all qubits interact the same way with their neighbors (modulo qubits at the boundary). Such Hamiltonians can provide good models for some real materials. However, Hamiltonians that lack translation invariance can also be interesting. For example, they may model a material with "disorder" due to imperfections (like dirt) in the sample. For such non-translation invariant Hamiltonians,

we have to specify how the local Hamiltonians vary from site to site in the system. Hamiltonians of this form are often called "spin glasses" for reasons that will become clear throughout the course of today's lecture.

2 Classical complexity theory

2.1 Recapitulation: the problem classes P and NP

Computation complexity is the study of how hard a certain computational task is. Digital computational tasks are formulated as *languages*. If f is a uniform family of Boolean functions $f : \{0,1\}^* \to \{0,1\}$, then the set of input strings accepted by f is called the associated *language*:

$$L = \{x \in \{0,1\}^* : \ f(x) = 1\}.$$

A language L is in P if and only if there exists a uniform classical circuit family C (traditionally: a deterministic Turing machine) such that

1. (Polynomial scaling) the size of $C(x)$ scales polynomially in all input sizes.
2. (Completeness) For all $x \in L$, $C(x) = 1$.
3. (Soundness) For all $x \notin L$, $C(x) = 0$.

P encompasses all computational tasks that are (classically) easy in the sense that they only require polynomial circuit size (runtime). If we increase the input size, the resulting overhead only increases polynomially.

Example 2.1 (Palindrome is in P)**.** A string $x = x_1 x_2 \cdots x_n \in \{0,1\}^*$ is in the language *palindrome* if it is equal to its own reversal: $x_1 x_2 \cdots x_n = x_n x_{n-1} \cdots x_1$. To see this, simply construct a circuit that reverses x and checks whether the reversal of x equals x itself. We leave it as an exercise to check that the size of such a circuit grows linearly in the input length n.

Another important problem class is NP. Roughly speaking, this problem class contains all problems that can be verified efficiently. A language L is in NP if and only if there exists a uniform classical circuit family $V(x, y)$ – the verifier – (traditionally: a deterministic Turing machine) such that

1. (Polynomial scaling) the size of the verifier $V(x, y)$ scales polynomially in the inputs x and y.
2. (Completeness) If $x \in L$, then there exists a witness y of polynomial size such that $V(x, y) = 1$.
3. If $x \notin L$, then no poly-size witness y can fool the verifier: $V(x, y) = 0$ for all y of polynomial size.

Example 2.2 (Circuit-SAT is in NP)**.** A Boolean circuit $C : \{0,1\}^n \to \{0,1\}$ comprised of poly(n) gates is in the language Circuit-SAT if there exists an input $x \in \{0,1\}^n$ such that $C(x) = 1$. This problem is in NP. Suppose that $C \in L$. Then, there exists a bit

string $x \in \{0,1\}^n$ such that $C(x) = 1$. This bit string verifies the language. Provided access to x, we can run the circuit C and verify membership (completeness). Conversely, if $C \notin L$, then no input x exists, where the circuit evaluates to one (soundness).

We emphasize that the problem class NP makes no assumption on how difficult it is to actually solve a problem. It merely ensures that proposed solutions can be efficiently checked for correctness. Note, that this feature is true, in particular, for problems in P: simply compute the answer in polynomial size (time) and output the solution as a witness. Therefore, P $\subset$ NP which intuitively makes sense. Checking correctness of a problem solution is not more difficult than coming up with a solution yourself. Interestingly, despite decades of active research, this is about as much as we know about the relationship between these two problem classes. Although widely believed to be true, the P $\neq$ NP conjecture is still open and, in fact, one of the \$1 million dollar problems from the Clay Mathematics Institute.

This lack of separation between intuitively meaningful problem classes makes it extremely difficult to determine the hardness of a computational task in absolute terms. Discontent with this state of the art, computer scientists have developed a rigorous formalism to assert relative hardness. Problem B is at least as hard as problem A if there exists a compiler – a polynomial-size uniform classical circuit family R – that maps x to $R(x)$ such that B accepts x if and only if A accepts $R(x)$. Such a compilation procedure *reduces* problem A to an instance of problem B. If we could find an ingenious way to solve problem B, we could also solve problem A with moderate overhead only.

A problem/language L is NP-*complete*, if it is contained in NP and every other NP-problem can be reduced to L. NP-complete problems are at the heart of the problem class NP: if you can solve only one of them efficiently, then you could solve all NP-problems efficiently as well ("one to rule them all"). This seems like a very strong condition, so it may seem surprising that NP-complete problems exist at all. Nonetheless, Karp came up with a list of 21 different NP-complete problems. This list contains famous problems like MAXCUT, traveling salesman, integer programming and has been greatly expanded since. The following example, however, is the quintessential example of an NP-complete problem.

Example 2.3 (k-SAT is NP-complete for $k \geq 3$)**.** The input to k-SAT is a Boolean formula on n variables $x_1, \ldots, x_n \in \{0,1\}$ that is a conjunction of polynomially many clauses that are comprised of k variables each, e.g.

$$f(x_1, \ldots, x_n) = (x_1 \vee x_5 \vee \bar{x}_2) \wedge (x_4 \vee \bar{x}_5 \vee x_3) \wedge \cdots \wedge (x_n \vee x_{n-1} \vee \bar{x}_1)$$

is an example of a 3-SAT formula in conjunctive normal form. Here, $\bar{x}_i$ denotes the negation of x_i. k-SAT is the language specified by the following rules:

1. SAT$(f) = 1$ if there exists $x = x_1 \cdots x_n \in \{0,1\}^n$ such that $f(x) = 1$.
2. SAT$(f) = 0$ otherwise.

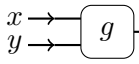
In words: A boolean formula f is in k-SAT if the formula obeys the structural constraints (conjunction of clauses with k terms each) and there exists a satisfiable assignment x, such that the formula evaluates to one (true).

Cook and Levin showed NP-completeness of k-SAT, by reducing the verification procedure associated with *any* NP-problem to an instance of a satisfiability problem. They cleverly converted the "there exists an efficient verifier"-part of the definition of an NP-problem into a "there exists an assignment that satisfies all the clauses" condition, i.e. k-SAT. We will see a variant of such an argument below, when we reduce Circuit-SAT to 3-SAT. Since we already know that Circuit-SAT is NP-complete, this reduction will provide an alternative proof of the Cook-Levin Theorem for $k = 3$.

Finally, we point out that it is relatively easy to reduce any k-SAT problem with $k \geq 3$ to 3-SAT by introducing dummy variables and reformulating the clauses in a given Boolean formula. Conversely, it is relatively straightforward to check that 1-SAT is an easy problem. More interestingly and less obvious is the insight that 2-SAT is also easy: 2-SAT $\in$ P, while 3-SAT is NP-complete.

2.2 Circuit-SAT reduces to 3-SAT

We will now show that Circuit-SAT (Example 2.2) reduces to 3-SAT (Example 2.3). To do so, we must find a compilation function that maps a classical circuit $C : \{0,1\}^n \to \{0,1\}$ – the input of Circuit-SAT – to a Boolean formula in conjunctive form, where each clause contains exactly 3 variables. To do so, we suppose that the circuit C is comprised of $m = \text{poly}(n)$ gates chosen form a universal gate set (e.g. AND, OR, NOT) such that each gate has at most 2 inputs and one output. The reduction of Circuit-SAT to 3-SAT is then based on reformulating the individual gates as clauses. To do so, we introduce a variable for the output of each gate and replace each gate with a clause:



$$\implies \quad C_g(x, y, z).$$

The three-variable clause $C_g(x, y, z)$ evaluates to true if and only if z is a valid output of the gate g for inputs x and y. Such a elementary reformulation is called a *gadget*. Some gates may only have a single input (like AND) and the associated clause only features two variables (input and output). Note that we can also regard input (no gate) and output (no gate) as trivial gates.

Replacing each gate by its corresponding clause – including an additional output variable – allows us to reformulate the circuit C as a Boolean function in conjunctive form

$$f_C : (x_1, \ldots, x_n, z_1, \ldots, z_{m'}) \to \{0,1\}$$

that evaluates to one if and only if the input $x = x_1 \cdots x_n$ is such that $C(x) = 1$. In turn, f_C is satisfiable if and only if there exists an input $x_1 \cdots x_n$ such that the original circuit evaluates to one. Conversely, if no such input exists, the Boolean function f_C cannot be satisfiable. Moreover, note that the number of additional (output) variables $z_1, \ldots, z_{m'}$ remains polynomial, because C is comprised of only polynomially many gates. This ensures that the reduction $C \mapsto f_C$ is indeed polynomial. Finally, note that the clauses have size at most 3 (and some of them have to have size 3, provided that the original circuit does contain gates with 2 inputs). So, checking whether f_C is satisfiable is a particular instance of 3-SAT.

Fact 2.4. *Circuit-*SAT *reduces to 3-*SAT*. Since Circuit-*SAT *is* NP*-complete, 3-*SAT *is also a* NP*-complete problem.*

It is worthwhile to emphasize the key idea behind this reduction. By including additional variables for each gate output, we found a 3-SAT formula that checks validity of the entire computing history of C: f_C is satisfiable if and only if the input evaluates to one ($C(x) = 1$) *and* all intermediate computational steps are logically correct. In so doing, we have replaced a dynamic problem – evaluating a circuit C for all possible inputs – by a static problem regarding the satisfiability of a single Boolean function. Moreover, note that it is easy to evaluate Boolean functions on concrete inputs. This implies that a valid history of the circuit computation is an efficiently checkable witness. This powerful idea can be extended to the quantum setting.

3 Classical local Hamiltonian ground state problems are NP-hard

Our reduction from Circuit-SAT to 3-SAT has important physical implications. Indeed, we can keep on going and reduce 3-SAT to a classical 3-local Hamiltonian ground state problem. Let

$$f(x_1, \ldots, x_n) = \bigwedge_{c=1}^{m} C_c(x_{c_1}, x_{c_2}, x_{c_3})$$

be a Boolean function in conjunctive form, where each clause contains exactly 3 variables. We can now replace each clause with a classical 3-local Hamiltonian

$$H_c(x_{c_1}, x_{c_2}, x_{c_3}) = \begin{cases} 0 & \text{if the clause } f_C(x_{c_1}, x_{c_2}, x_{c_3}) = 1 \text{ (true) for } x_{c_1}, x_{c_2}, x_{c_3} \in \{0,1\}, \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

The associated total Hamiltonian is then

$$H = \sum_{c=1}^{m} H_c(x_{c_1}, x_{c_2}, x_{c_3})$$

and acts on a classical system of n 0/1-variables $x_1, \ldots, x_n$ (e.g. spins) in a 3-local fashion. By construction,

$$\min_{x \in \{0,1\}^n} H(x) \quad \begin{cases} = 0 & \text{if } f(x_1, \ldots, x_n) \text{ is satisfiable,} \\ \geq 1 & \text{otherwise.} \end{cases}$$

Indeed, a satisfying assignment would ensure that all local Hamiltonian terms contribute zero-energy. Conversely, if the original Boolean function is not satisfiable, then there must be at least one local Hamiltonian that evaluates to one, regardless of the assignment. This allows us to conclude that the task of finding the ground state of a 3-local classical Hamiltonian is at least as hard as 3-SAT. Since 3-SAT is also NP-complete, the 3-local Hamiltonian problem is actually at least as hard as *any* problem in the class NP.

In fact, approximating the ground state energy to constant accuracy is NP-hard, even for classical Hamiltonians that are only 2-local: $H(x) = \sum_c H_c(x_{c_1}, x_{c_2})$. This statement follows from a different reduction argument. While 2-SAT is in P and thus

easy, a seemingly related problem is much harder. MAX-2-SAT is the problem of determining the maximum number of 2-clauses that can be satisfied in a given Boolean formula. This problem clearly allows for solving 2-SAT as a special case, but turns out to be much harder – NP-hard in fact. By replacing each 2-clause with a 2-local Hamiltonian a la (1) and computing the ground state energy of the associated classical Hamiltonian, we could effectively count the maximum number of satisfiable clauses.

Finally, we point out that one can further make such 2-local, classical Hamiltonians geometrically local without losing hardness. A concrete and prominent example is the *Ising spin-glass model* in three (or more) dimensions. Suppose that the binary variables x_i are actually classical "spins" sitting at the sites of a cubic lattice: $x_i \to z_i \in \{\pm 1\}$. The Ising-Hamiltonian takes on the following geometrically 2-local form:

$$H(J) = -\sum_{\langle ij \rangle} J_{ij} z_i z_j,$$

where $\langle ij \rangle$ labels the edge in the lattice that connects nearest neighbor sites only. The coupling constants $J_{ij} \in \{\pm 1\}$ encode the instance of the problem. If $J_{ij} = +1$, we say that the edge $\langle ij \rangle$ is ferromagnetic, because it is energetically favorable for the neighboring spins to align. Conversely, we say that an edge is antiferromagnetic if $J_{ij} = -1$.

If all the edges are ferromagnetic, then the Ising ground state problem is almost trivial. Simply align all the spins in one direction (up or down) to minimize energy. The problem becomes more interesting if some of the edges are antiferromagnetic. Such edges can generate frustration, meaning that it is impossible to minimize all Hamiltionians $-J_{ij} z_i z_j$ simultaneously and independently from each other. The resulting global optimization problem becomes "coupled" and greedy strategies (optimize different localized regions independently) are prone to get stuck in local minima, because it may require a lot of synchronized spin flips to further lower the energy. Viewed from this angle, it should not be surprising that frustrated Ising problems can be very difficult to solve. What should be more surprising, is the fact that such problems actually admit efficient solutions in one, or even two dimensions. We won't have time to talk about the fundamental difference between 2D and 3D Ising problems and refer to John's handwritten notes for further exposition.

Instead, we conclude this section by pointing out that there are also NP-hard spin glass problems in 2D, provided that we introduce additional local magnetic field terms in the Hamiltionian:

$$H = -\sum_{\langle ij \rangle} J_{ij} z_i z_j - \sum_i h_i z_i \quad \text{where} \quad J_{ij}, h_i \in \{-1, 0, 1\}.$$

The local magnetic field terms compound the frustration. Each spin wants to align with its local field, but by doing so the edge connecting said spin with its neighbors might become excited.

To summarize: it is often difficult to minimize local Hamiltonian problems with frustration, because it is impossible to locally relax spins into more energetically favorable configurations. In fact this is feature is the main reason why physicists call

them "glasses". Since it is so difficult to make local progress towards the ground state, these (hypothetical) materials remain in a transitionary, amorphous configuration for exceptionally long time scales.

4 The quantum Hamiltonian problem and the problem class QMA

The previous section highlights that classical 2-local Hamiltonian ground state problems can be as hard as any classical problem in NP. This is because of frustration: there is no way to satisfy all the local "clauses" simultaneously and there may be numerous local energy minima that do not minimize energy globally. The quantum version of this problem seems even more daunting: $H = \sum_a H_a$ and the H_a's might not commute with each other. This implies that it is impossible to simultaneously diagonalize the local Hamiltionians – a feature that compounds frustration effects even further. Indeed, the ground state could be highly entangled, with no succinct classical description.

Hence, we have strong reasons to believe that the quantum local Hamiltonian problem is even more difficult than its already difficult classical cousin. We devote the remainder of this lecture to rigorously pinpoint this intuition.

4.1 The quantum k-local Hamiltonian problem

Fix a n-qubit Hamiltonian

$$H = \sum_a H_a \tag{2}$$

that is k-local in the sense that each H_a acts nontrivially on at most k qubits. We furthermore assume that k is constant, $\|H_a\|_\infty \leq h$ (also constant) and each $2^k \times 2^k$ matrix H_a is specified up to $\text{poly}(n)$ bits of precision only. Moreover, we assume a *promise gap*. The ground state energy E_0 – i.e. the lowest eigenvalue of H – obeys either

$$E_0 \leq E_{\text{low}}, \quad \text{or} \quad E_0 > E_{\text{high}}, \quad \text{where} \quad E_{\text{high}} - E_{\text{low}} > \tfrac{1}{\text{poly}(n)}. \tag{3}$$

This promise gap prevents us from cheating when we want to establish hardness of this ground state problem. An exponentially small promise gap would potentially require exponentially large resolution to resolve and may therefore be difficult to detect with any computational device. Similar concerns are the reason for assuming that the local Hamiltonians (the problem description) are only specified to polynomial accuracy. The local Hamiltonian problem then corresponds to the following yes-no question.

Definition 4.1 (Local Hamiltonian)**.** Given a local Hamiltonian H (2) and a promise gap (3), decide whether $E_0 \leq E_{\text{low}}$, or $E_0 > E_{\text{high}}$, i.e.

$$f(H) = \begin{cases} 1 & \text{if } E_0 \leq E_{\text{low}}, \\ 0 & \text{if } E_0 > E_{\text{high}} \end{cases}$$

The promise gap in the problem formulation allows us to answer the yes/no question by determining the ground state energy E_0 up to $1/\text{poly}(n)$ accuracy only. We already know that the Local Hamiltonian problem is NP-hard, because it includes classical 3-local Hamiltonian as a special case.

4.2 The problem class QMA

Recall that NP is the classical complexity class that encompasses all problems that can be efficiently checked with classical circuits in a deterministic fashion. There is a class analogous to NP for randomized computation. This class is called MA and its quantum generalization is the class QMA.

A language L is in QMA if there exists a uniform quantum circuit family V – the quantum verifier – and a single qubit measurement $\{E_0, E_1\}$ such that

1. (Polynomial scaling) V has polynomial size.
2. (Probabilistic completeness) If $x \in L$, then there exists a quantum witness $|\psi_x\rangle$ such that $|\psi\rangle = V(|\psi_x\rangle \otimes |x\rangle \otimes |0\rangle^*)$ implies $\Pr[\text{accept}] = \langle\psi|E_1|\psi\rangle \geq 2/3$.
3. (Soundness) If $x \notin L$, then $\Pr[\text{accept}] = \langle\psi|E_1|\psi\rangle \leq 1/3$ for all potential witnesses $|\psi_x\rangle$.

Note that completeness and soundness conditions have been somewhat weakened. We are content if we can use the quantum witness $|\psi_x\rangle$ to correctly verify $x \in L$ with probability $2/3 > 1/2$, while we limit the probability of "false positives" by $1/3 < 1/2$. The exact value of these probabilities is somewhat arbitrary. As long as the probability of correctly verifying $x \in L$ is strictly larger than $1/2$, and the probability of a false positive is strictly smaller than $1/2$, we can repeat the probabilistic membership test several times and boost the probability of determining membership correctly: $2/3 \to 1 - \varepsilon$ and $1/3 \to \varepsilon$, where $\varepsilon > 0$ can be exponentially small. Such strategies allow us to make the defining properties 2. and 3. of QMA effectively stronger:

2.' (Boosted probabilistic completeness) If $x \in L$, then there exists a quantum witness $|\psi\rangle = V(|\psi_x\rangle \otimes |x\rangle \otimes |0\rangle^*)$ such that $\Pr[\text{accept}] = \langle\psi|E_1|\psi\rangle \geq 1 - \varepsilon$.

3.' (Boosted soundness) If $x \notin L$, then $\Pr[\text{accept}] = \langle\psi|E_1|\psi\rangle \leq \varepsilon$ for all potential witnesses $|\psi_x\rangle$.

Remark 4.2 (Subtlety regarding probability amplification). Classically, boosting the probability of success/failure is easy. Simply execute the same computation multiple times in parallel. In the quantum case, there is an additional complication, because the single witness for such a parallel computation may not be a product state that factorizes nicely into its individual constituents. Indeed, the total witness may be entangled and using individual constituents may result in seemingly mixed states. However, such effective mixed states cannot prompt us to accept the computation after many trials unless there actually is some proper (pure) input that is accepted with high probability in each trial.

5 The quantum Hamiltonian ground state problem is QMA-complete

QMA is the natural quantum generalization of MA – the probabilistic relaxation of NP. Also, recall that the 3-SAT problem is at the very heart of the classical problem class NP, because it encodes verification procedures of arbitrary NP problems (Cook-Levin). We now claim that the local Hamiltonian problem assumes such a central role for the quantum complexity class QMA: It is QMA complete for $k \geq 5$. Moreover, we shall see

that it is possible to encode any quantum verification procedure into the ground state of a local Hamiltonian. This deep result was established by Kitaev and subsequently improved to remain valid for geometrically 2-local Hamiltonians in 2D (qubits) and even geometrically 2-local Hamiltonians in 1D (for higher dimensional qudits, $d \geq 12$).

5.1 Roadmap

In order to establish this claim, we need to show 2 things:

1. The k-local Hamiltonian problem is itself in QMA.
2. *Any* QMA-problem can be reduced to a k-local Hamiltonian problem. We will show this for $k = 5$ without geometric locality, as in Kitaev's original work. This will require designing a local Hamiltonian H that reduces a given QMA question $x \in L$ in a question about ground state energies $E_0 = \min_\varphi \langle\varphi|H|\varphi\rangle$:

   i. Probabilistic completeness→"small" ground state energy: if $x \in L$, then $E_0 \leq \frac{\varepsilon}{T+1}$, where ε is the probability of a false negative ($\Pr[\text{accept}] = 1 - \varepsilon$) and T is the size of the verifying circuit ($T = \text{poly}(n)$).

   ii. Soundness→"large" ground state energy: $x \notin L$, then $E_0 \geq \text{const}\frac{1-\sqrt{\varepsilon}}{(T+1)^3}$.

   This discrepancy in ground state energies ensures a promise gap of size $1/\text{poly}(n)$ (provided that ε is sufficiently small). This is an essential feature of Hamiltonian ground state problems.

   The first step follows from a specifically engineering a Hamiltonian whose ground state energy is zero if we were able to verify membership perfectly, i.e. without probability of failure. Quantum mechanics is a linear theory and a relaxation to acceptance with very high probability remains benign. This part may be viewed as a relatively straightforward generalization of our classical reduction from Circuit-SAT to 3-SAT.

   The second step is new and therefore more challenging. In the classical case, soundness immediately followed form the binary nature of Boolean function. They are either true, or false. And if they are false, there can be no satisfying assignment – i.e. a false positive – whatsoever. The probabilistic nature of QMA thwarts this argument and we will explicitly have to verify that the ground state energy grows substantially if $x \notin L$.

   Finally, we will also have to make sure that the Hamiltonian H that encodes a given QMA problem is also local. We will follow Kitaev's original ideas and establish locality with $k = 5$.

5.2 k-local Hamiltonian is in QMA

This part has already been established in the previous lecture. A natural quantum witness of the k-local Hamiltonian problem is the ground state $|\psi_{\min}\rangle$ itself: $H|\psi_{\min}\rangle = E_0|\psi_{\min}\rangle$. Provided that we have access to this witness, we can compute the associated energy $E_0 = \langle\psi_0|H|\psi_0\rangle$ up to accuracy $1/\text{poly}(n)$ using the phase estimation algorithm.

5.3 Constructing the Hamiltonian

For this reduction, we will follow the strategy that we used to show that 3-SAT is NP-complete. The problem class QMA comes with a promise: there exists a quantum verifier and a poly-size quantum circuit that allows us to verify membership $x \in L$ by performing a particular quantum computation. This promise alone will allow us to construct a general witness that encodes the whole history of the verification procedure (inputs+computation+measurement). Suppose that $x = x_1 \cdots x_n$ is a input that belongs to a QMA language L. To set the stage, let us connect all the information that QMA provides us for free:

$$\begin{aligned} |\psi\rangle =& V(\overbrace{|\psi_x\rangle \otimes |x\rangle \otimes |0\rangle^*}^{|\psi_0\rangle}) \quad \text{such that} \quad \langle\psi|E_1|\psi\rangle \geq 1-\varepsilon \simeq 1, \\ V =& U_T U_{T-1} \cdots U_1 \quad U_t \text{ 2-local elementary gates}, \quad T = \text{poly}(n). \end{aligned}$$

Here, $|x\rangle$ is the input, $|\psi_x\rangle$ is the witness and $|0\rangle^*$ denotes a collection of scrap qubits initialized in the zero-state. The first crucial insight is as follows: instead of demanding a quantum witness $|\psi_x\rangle$ of x only, we can also demand a quantum witness for the entire verification procedure:

$$|\eta\rangle = \tfrac{1}{\sqrt{T+1}} \sum_{t=0}^{T} |\psi(t)\rangle \otimes |t\rangle, \quad \text{where} \quad |\psi(t)\rangle = (U_t U_{t-1} \cdots U_0)|\psi_0\rangle. \tag{4}$$

This is called a *history state*. Here, $|\psi_0\rangle$ is the original quantum verifier input and $|\psi(t)\rangle$ is the state obtained after the first t steps of the verification circuit. The state $|t\rangle$ corresponds to a clock register that records time flow, $t \in \{0, 1, \ldots, T\}$. For now, we assume that $|t\rangle$ simply corresponds to the computational basis of a $(T+1)$-dimensional Hilbert space. We will decompose this space into T clock qubits later on. Note that orthogonality $\langle t|s\rangle = \delta_{t,s}$ ensures that $|\eta\rangle$ is properly normalized.

Based on the history state (4), we can now construct a Hamiltonian that checks validity of the history state $|\eta\rangle$. In particular:

- Correct (classical) input: problem input $|x\rangle$ and scrap $|\psi_0\rangle$ must be correctly initialized.
- Correct (classical) output: the final time state $|\psi(T)\rangle$ must lead us to accept $x \in L$ with high probability: $|\langle 1|\psi(T)\rangle|^2 \geq 1-\varepsilon$.
- Proper computation history: $|\psi(t)\rangle = U_t|\psi(t-1)\rangle$ for all $t = 1, \ldots, T$.
- Correct encoding of time: the clock register must be properly encoded and updated.

We can now design individual Hamiltonians that introduce energy penalties, whenever any of these conditions is violated. The total Hamiltonian will then simply correspond to the sum of these constituents:

$$H = H_{\text{in}} + H_{\text{out}} + H_{\text{prop}} + H_{\text{clock}}. \tag{5}$$

- H_{in} enforces that the inputs are correct at time $t = 0$. For $x = x_1 \cdots x_n$, we get n contributions:

$$H_x^{(j)} = (|\bar{x}_j\rangle\langle\bar{x}_j|)^{(j)} \otimes \mathbb{I}^{(\text{else})} \otimes (|0\rangle\langle 0|)^{(\text{clock})} \quad j = 1, \ldots, n$$

  and correct initialization of $L = \text{poly}(n)$ scrap qubits is enforced by

$$H_{\text{scrap}}^{(l)} = (|1\rangle\langle 1|)^{(l)} \otimes \mathbb{I}^{(\text{else})} \otimes (|0\rangle\langle 0|)^{(\text{clock})} \quad l = 1, \ldots, L.$$

  The total penalizing Hamiltonian is $H_{\text{in}} = \sum_{j=1}^{n} H_x^{(j)} + \sum_{l=1}^{L} H_{\text{scrap}}^{(l)}$ and there is an energy penalty of 1 for each qubit that is not properly initialized.

- H_{out} imposes a penalty if the verifier circuit fails to accept a valid computation:

$$H_{\text{out}} = (|0\rangle\langle 0|)^{(\text{output})} \otimes \mathbb{I}^{(\text{else})} \otimes (|T\rangle\langle T|)^{(\text{clock})}.$$

  There is an energy penalty of 1 if the output qubit hast the value $|0\rangle$ rather than $|1\rangle$ (false negative).

- H_{prop} checks if each step in the verifying computation is executed correctly: $H_{\text{prop}} = \sum_{t=1}^{T} H_{\text{prop}}(t)$, where

$$H_{\text{prop}}(t) = \tfrac{1}{2}\Big(\mathbb{I} \otimes |t\rangle\langle t| + \mathbb{I} \otimes |t-1\rangle\langle t-1| - U_t \otimes |t\rangle\langle t-1| - U_t^\dagger \otimes |t-1\rangle\langle t|\Big).$$

  We leave it as an easy exercise to check that these terms are Hermitian. More importantly, each $H_{\text{prop}}(t)$ is constructed such that

$$\begin{aligned} |\psi(t-1)\rangle \otimes |t-1\rangle &\mapsto \tfrac{1}{2}(\psi(t-1) \otimes |t-1\rangle - \psi(t) \otimes |t\rangle), \\ |\psi(t)\rangle \otimes |t\rangle &\mapsto \frac{1}{2}(\psi(t) \otimes |t\rangle - \psi(t-1) \otimes |t-1\rangle). \end{aligned}$$

  If the entire Hamiltonian acts one a history state $|\eta\rangle$ that obeys all these sanity checks, then these contributions sum up to zero (telescoping sum):

$$H_{\text{prop}}|\eta\rangle = 0 \quad \text{if} \quad |\eta\rangle \text{ is a valid history state.}$$

5.4 Converting probabilistic completeness into a small ground state energy

The Hamiltonian H (5) is constructed in such a way that a valid history state of a proper verification procedure for $x \in L$ incurs zero energy penalties, provided that the circuit accepts with certainty. QMA, however, merely ensures that this happens with high probability $1 - \varepsilon$. In other words: with probability ε, we may still end up in a final state that is $|0\rangle$, not $|1\rangle$. The specific structure of the history state – in particular its normalization $1/\sqrt{T+1}$ – further suppresses potential energy penalties in this case:

$$\langle\eta|H|\eta\rangle = \langle\eta|H_{\text{out}}|\eta\rangle \leq \frac{\varepsilon}{T+1},$$

because H_{out} only acts on the $t = T$ portion of the history state. This highlights that our Hamiltonian construction still does what it is supposed to do, if we relax perfect verification to verification with high probability $1 - \varepsilon$ only.

5.5 Converting probabilistic soundness into a large ground state energy

On first sight, soundness seems straightforward. We have constructed our Hamiltonian such that it's lowest energy is zero if and only if everything works out correctly (and we can certify membership with probability one). If anything goes wrong, there will be an energy penalty that forces the ground state energy to be strictly larger than zero.

Unfortunately, such a naive argument is not enough. Not only do we have to take into account probabilistic acceptance/rejection, but we also need to establish a polynomial promise gap: if $x \notin L$, then the ground state energy of H must be at least $1/\mathrm{poly}(n)$ away from the maximal acceptance energy $\varepsilon/(T+1)$!

In order to achieve such a separation, we need to take a closer look at the ground spaces of different Hamiltonian contributions and the geometric relations between them. More concretely, we will divide the Hamiltonian H (5) into two relevant parts and ignore the clock Hamiltonian (it can only increase the gap further):

$$H = \underbrace{H_{\mathrm{in}} + H_{\mathrm{out}}}_{H_1} + \underbrace{H_{\mathrm{prop}}}_{H_2} + H_{\mathrm{clock}}.$$

Roughly speaking, the nullspace – i.e. the subspace of ground states – of H_1 will be somewhat aligned with the computational basis, because the penalizing terms are diagonal in the computational basis. Conversely, we will show that the eigenvectors of $H_2 = H_{\mathrm{prop}}$ are somewhat aligned with a "Fourier transform" of the computional basis instead. The relation between these nullspaces is governed by an uncertainty relation: succinct expansions in one basis become very spread out in the other basis and it is impossible to find a good balance between these extremes. We can either assure correct encoding and readout (nullspace of H_1), or a correct computational history (nullspace of H_2) – but not both. This follows from the soundness promise: if $x \notin L$, then it is impossible to fool the polynomial verification procedure with probability larger than ε.

Diagonalizing $H_2 = H_{\mathrm{prop}}$ We can explicitly determine the eigenvectors and eigenvalues of H_{prop} by performing a clever basis change (unitary transformation):

$$V = \sum_{t'=0}^{T} V_{t'} \otimes |t'\rangle\langle t'| \quad \text{where} \quad V_{t'} = U_{t'} U_{t'-1} \cdots U_1$$

The adjoint of this unitary freezes the motion in the history state ("rotating frame"):

$$V^\dagger |\eta\rangle = \tfrac{1}{\sqrt{T+1}} \left(\sum_{t'=0}^{T} V_{t'} \otimes |t'\rangle\langle t'| \right) \tfrac{1}{\sqrt{T+1}} \left(\sum_{t=0}^{T} V_t |\psi_0\rangle \otimes |t\rangle \right) = |\psi_0\rangle \otimes \left(\tfrac{1}{\sqrt{T+1}} \sum_{t=0}^{T} |t\rangle \right).$$

The Hamiltonian H_{prop} also transforms nicely under this basis change:

$$\begin{aligned}
&\tilde{H}_{\mathrm{prop}} = V^\dagger H_{\mathrm{prop}} V \\
=&\sum_{t=1}^{T} \tfrac{1}{2} \Big(V_t^\dagger V_t \otimes |t\rangle\langle t| + V_{t-1}^\dagger V_{t-1} \otimes |t-1\rangle\langle t-1| - V_t^\dagger U_t V_{t-1} \otimes |t\rangle\langle t-1| - V_{t-1}^\dagger U_t^\dagger V_t \otimes |t-1\rangle\langle t| \Big) \\
=&\mathbb{I} \otimes \sum_{t=1}^{T} \tfrac{1}{2} (|t\rangle\langle t| + |t-1\rangle\langle t-1| - |t\rangle\langle t-1| - |t-1\rangle\langle t|).
\end{aligned}$$

This rotated Hamiltonian only acts nontrivially on the clock, where it corresponds to a sum of overlapping 2×2 blocks. The total Hamiltonian is described by the following $(T+1) \times (T+1)$ matrix (with respect to the computational basis $|0\rangle, \ldots, |T\rangle$):

$$\tilde{H}_{\text{prop}} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & \cdots & & \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \cdots & \\ 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \cdots \\ & & \ddots & & & \\ & & & \ddots & & \\ & \cdots & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} \\ & & \cdots & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = \mathbb{I} - \frac{1}{2} \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & & \\ 1 & 0 & 1 & 0 & \cdots & \\ 0 & 1 & 0 & 1 & 0 & \cdots \\ & & \ddots & & & \\ & & & \ddots & & \\ & \cdots & 0 & 1 & 0 & 1 \\ & & \cdots & 0 & 1 & 1 \end{bmatrix}}_{M}.$$

The identity matrix does not affect eigenvectors. Every eigenvector of M will be an eigenvector of $\tilde{H}_{\text{prop}} = \mathbb{I} - \frac{1}{2}M$ with eigenvalue $1 - \lambda/2$. The matrix M is very structured. It looks a lot like an interaction matrix associated with a 1-dimensional chain of coupled classical oscillators (each oscillator j interacts with its left and right neighbor) without periodic boundary conditions. The first oscillator talks to itself and to its only neighbor, while the last oscillator also talks to itself and its only neighbor. As you might recall from your classical mechanics class, matrices of this form are diagonalized by a (finite dimensional) Fourier transform:

$$|\omega\rangle = \tfrac{1}{\sqrt{T+1}} \sum_{t=0}^{T} \mathrm{e}^{i\omega t} |t\rangle \quad \text{where} \quad \omega \in [0, 2\pi). \tag{6}$$

However, not every phase ω is allowed, because of boundary conditions. Relatively straightforward consistency checks involving the first and last basis vector (the boundary) demand

$$\omega(T+1) = \pi k \quad \Longleftrightarrow \quad \omega = \omega_k = \frac{\pi k}{T+1} \quad \text{where} \quad k \in \mathbb{N}.$$

The associated eigenvalue is

$$\lambda_k(M) = 2\cos(\omega_k) \quad \text{and} \quad \lambda_k\Big(\tilde{H}_{\text{prop}}\Big) = 1 - \cos(\omega_k).$$

Hence, we conclude that $\tilde{H}_{\text{prop}}$ has a 1-dimensional ground space ($k = 0$) whose energy is zero and the associated eigenvector is

$$|\omega = 0\rangle = \tfrac{1}{\sqrt{T+1}} \sum_{t=0}^{T} |t\rangle.$$

This is just the clock-part of the frozen history state $V^\dagger|\eta\rangle$ – a quick consistency check. More importantly, we conclude that every eigenvector of $\tilde{H}_{\text{prop}}$ corresponds to a uniform superposition of $|t\rangle$-states (6) and the Hamiltonian has a spectral gap:

$$\lambda_1(H_{\text{prop}}) = \lambda_1(\tilde{H}_{\text{prop}}) = 1 - \cos\left(\frac{\pi}{T_1}\right) = 2\sin^2\left(\frac{\pi}{2(T_1+1)}\right) \approx \frac{\pi^2}{2(T+1)^2} =: \Delta_2.$$

This spectral gap is inverse polynomial, because $T = \text{poly}(n)$.

Diagonalizing $H_1 = H_{\text{in}} + H_{\text{out}}$ Diagonalizing H_1 is much simpler by comparison. The nullspace is spanned by all history states that encode a valid input at $t = 0$ and are accepted (answer yes) at $t = T$. By construction, these state vectors are computational basis vectors in the respective Hilbert spaces. The remaining computational basis vectors incur an energy penalty of at least one: $\langle H_1 \rangle \geq 1$ for all vectors orthogonal to the nullspace. This allows us to conclude a spectral gap of size $\Delta_1 = 1$.

Relating the nullspaces of H_1 and H_2 Note that the nullspaces of H_1 and H_2 are not unrelated. If the verifier accepts the input with probability one, then there is a simultaneous null eigenvector $|\eta\rangle$ of both H_1 and H_2 – a valid history state of the accepting computation. However, by the QMA-promise this is cannot be the case if x does not belong to the language. The probability of accepting is very small and the different nullspaces cannot intersect nontrivially. This discrepancy can be characterized by an angle θ between the nullspaces – a straightforward generalization of the familiar angle between two vectors (lines). In the soundness regime, the probability of wrongfully accepting a computation is small and therefore the relative angle between the two different nullsapces must be reasonably large.

We can relate this angle to the ground state energy of the joint Hamiltonian $H_1 + H_2$. Let Π_1 be the projection onto the nullspace of H_1, and let Π_2 be the projection onto the nullspace of H_2. Moreover, let $\Delta = \min\{\Delta_1, \Delta_2\} \approx \pi^2/(2(T+1)^2)$ be the minimal spectral gap of these Hamiltonians. Then,

$$\langle H_1 + H_2 \rangle \geq \Delta \langle \mathbb{I} - \Pi_1 \rangle + \Delta \langle \mathbb{I} - \Pi_2 \rangle \geq 2\Delta \langle \mathbb{I} \rangle - \Delta \langle \Pi_1 + \Pi_2 \rangle = \Delta(2 - \langle \Pi_1 + \Pi_2 \rangle)$$

for any state. Now it is time to take into account the relative geometry between the two nullspaces. Suppose that Π_1 and Π_2 are such that the maximal overlap between different eigenvectors is bounded:

$$|\langle \psi_1 | \psi_2 \rangle| \leq \cos(\theta) \quad \text{for all} \quad |\psi_1\rangle \in \text{range}(\Pi_1),\ |\psi_2\rangle \in \text{range}(\Pi_2). \tag{7}$$

Then, a straighforward extension of a vector-angle argument ensures $\langle \Pi_1 + \Pi_2 \rangle \leq 1 + \cos(\theta)$. This allows us to conclude

$$\langle H_1 + H_2 \rangle \geq 2\Delta - \Delta \langle \Pi_1 + \Pi_2 \rangle \geq \Delta(1 - \cos(\theta)) = 2\Delta \sin^2(\theta/2) \quad \text{for any state.}$$

This almost allows us to establish the desired promise gap for the soundness case. There is no quantum witness that will allow us to accept the computation with high probability. If we want to minimize the total energy of the associated Hamiltonian, we can either try to get input and output right (land in the nullspace Π_1 of H_1), or get the computational history right (land in the nullspace Π_2 of H_2), but not both:

$$E_0 = \min\langle H \rangle = \min\langle H_1 + H_2 + H_{\text{clock}} \rangle \geq \min\langle H_1 + H_2 \rangle \geq \frac{\pi^2}{(T+1)^2} \sin^2(\theta/2). \tag{8}$$

In order to complete the argument, we must now show that the angle θ cannot be too small. To do so, we turn back to the definition of the angle θ (7):

$$\cos^2(\theta) = \max_{|\psi_1\rangle \in \text{ran}(\Pi_1), |\psi_2\rangle \in \text{ran}(\Pi_2)} |\langle \psi_1 | \psi_2 \rangle|^2 \leq \max_{|\psi_2\rangle \in \text{ran}(\Pi_2)} \langle \psi_1 | \Pi_1 | \psi_1 \rangle$$

Each $|\psi_2\rangle \in \mathrm{ran}(\Pi_2)$ is a valid history state $|\psi_2\rangle = \frac{1}{\sqrt{T+1}} \sum_{t=0}^{T} |\psi(t)\rangle \otimes |t\rangle$. Moreover, the projector Π_1 onto the nullspace of H_1 only cares about inputs and outputs. It acts trivially on clock states with $t \in \{1, \ldots, T-1\}$. After transforming into the convenient rotating frame basis (where these clock-degrees of freedom are frozen), we learn that

$$\langle \psi_2 | V V^\dagger \Pi_1 V V^\dagger | \psi_2 \rangle = \langle \tilde{\psi} | \tilde{\Pi}_1 | \tilde{\psi}_2 \rangle = \frac{T-1}{T+1} + \frac{1}{T+_1} \langle \psi_2' | \Pi_{\mathrm{in}} + \Pi'_{\mathrm{out}} | \psi_2' \rangle,$$

where ψ_2' is a state on the non-clock variables only, Π_1 still projects onto valid input states and Π'_{out} projects onto the rotated output

$$V_t^\dagger \Big(|1\rangle^{(\mathrm{out})} \otimes \text{anything else} \Big).$$

We can now recycle our angle argument:

$$\langle \psi_2' | \Pi_{\mathrm{in}} + \Pi'_{\mathrm{out}} | \psi_2' \rangle \leq \max \langle \Pi_{\mathrm{in}} + \Pi'_{\mathrm{out}} \rangle \leq (1 + \cos(\varphi)),$$

where φ is the angle between the ranges of Π_{in} and Π'_{out}. Here, we are finally at a stage where we can use the QMA soundness promise. For valid inputs (elements in the range of Π_{in}) associated with $x \notin L$, the probability of accepting is at most ε. This implies $\cos^2(\varphi) \leq \varepsilon$ and we can backtrack to get a bound on θ:

$$\cos^2(\theta) \leq \frac{T-1}{T+1} + \frac{1}{T+1}(1 + \cos(\varphi)) \leq \frac{T-1}{T+1} + \frac{1+\sqrt{\varepsilon}}{T+1} = 1 - \frac{1-\sqrt{\varepsilon}}{T+1}.$$

We can convert this upper bound on cos into a lower bound on sin:

$$\sin^2(\theta/2) \geq \tfrac{1}{4} \sin^2(\theta) = \tfrac{1}{4}(1 - \cos^2(\theta)) \geq \frac{1-\sqrt{\varepsilon}}{4(T+1)}.$$

Plugging this bound into our assertion about the minimum ground state energy (8) ensures

$$E_0 \geq \frac{\pi^2}{(T+1)^2} \sin^2(\theta/2) \geq \frac{\pi^2(1-\sqrt{\varepsilon})}{4(T+1)^3}.$$

It is instructive to compare this to the upper bound on the ground state energy for the case where the quantum verifier accepts:

$$E_0(x \in L) \leq \frac{\varepsilon}{T+1} \quad \text{while} \quad E_0(x \notin L) \geq \frac{1-\sqrt{\varepsilon}}{(T+1)^3}$$

for any language L in QMA. Viewed as a promise gap for the Hamiltonian ground state problem, this separation is inverse polynomial, provided that ε is small compared to $(T+1)^{-3}$.

5.6 Establishing locality of this general ground state problem

We have already achieved quite a lot. We managed to show that any QMA-problem can be reduced to the task of determine whether the ground state of a certain Hamiltonian is either below (accept x), or above (reject x) certain values separated by a promise gap. Moreover, the Hamiltonian decomposes nicely into different constituents that look almost local. Locality is only thwarted by the clock states which we so far have associated with the computational basis of a $(T+1)$-dimensional Hilbert space: $|t\rangle$ with $t = 0, \ldots, T$. However, we can encode these clock-states into a register composed precisely T qubits:

$$\begin{aligned} |t=0\rangle &= |000\cdots0\rangle, \\ |t=1\rangle &= |100\cdots0\rangle, \\ |t=2\rangle &= |110\cdots0\rangle, \\ &\vdots \\ |t=T\rangle &= |111\cdots1\rangle. \end{aligned}$$

This intuitive encoding is called "unary" and preserves orthogonality relations: $\langle t, s\rangle = \delta_{t,s}$. Moreover, terms in H that do depend on the clock state in question only compare $|t\rangle$ with itself, or its neighbors $|t+1\rangle$ and $t-1\rangle$. Unary encoding ensures that these contributions are at most 3-local. Indeed, the clock Hamiltonian – designed to enforce that time steps are properly updated – becomes 2-local:

$$H_{\text{clock}} = \sum_{t=1}^{T-1} (|01\rangle\langle 01|)_{t,t+1}.$$

Indeed, a $\{0,1\}$-string is a valid unary encoding of time if and only if a zero is never followed by a one. The above 2-local Hamiltonian enforces precisely that. For such valid encodings, the projection on $|t\rangle$ only needs to act on clock qubits numbered t and $t+1$:

$$|t\rangle\langle t| = (|10\rangle\langle 10|)_{t,t+1}.$$

This is another 2-local term. Finally, terms that advance or retard time act on three qubits each:

$$\begin{aligned} |t\rangle\langle t-1| &= (|110\rangle\langle 100|)_{t-1,t,t+1}, \\ |t-1\rangle\langle t| &= (|100\rangle\langle 110|)_{t-1,t,t+1}. \end{aligned}$$

These 3-local terms appear in H_{prop} as a tensor product with the elementary quantum gate U_t executed at time t: $U_t \otimes |t\rangle\langle t-1|$. Since we always assume that these gates are chosen from a universal gate set involving at most 2 qubits, the total term is effectively 5-local. All other Hamiltonian terms act on 4 or less qubits and we can conclude that *unary encoding of the clock register results in a 5-local Hamiltonian.*

6 Summary

We have shown that the 5-local Hamiltonian ground state problem is a "natural" QMA-complete problem. Much like 3-SAT is a natural NP-complete problem. Unfortunately, the family of problems that have been shown to be QMA-complete is still rather small and comprised of relatively artificial problems. This is in stark contrast to the classical case, where many practical problems turn out to be NP complete.

It is widely believed that the problem class QMA is strictly larger than the classical class NP. This indicates that the quantum local Hamiltonian problem is really harder than the classical local Hamiltonian problem.

Finally, we point out that it is possible to further reduce the locality while still maintaining QMA-completeness: $k = 5$ can be replaced by $k = 2$.