Frontiers of Quantum complexity *Instructor: Anand Natarajan, Anurag Anshu*

Lecture 2: QMA in PP and the Local Hamiltonian problem

September 9, 2025 *Scribe: Kabir Peshawaria*

1 The PP Complexity Class

Last class we discussed four complexity classes and showed the following relationships:

$$\mathsf{NP} \subseteq \mathsf{MA} \subseteq \mathsf{QCMA} \subseteq \mathsf{QMA}.$$

Since we have a better grasp of *classical* complexity theory, a natural question to ask is

"In which classical complexity class does QMA *reside?"*.

This question helps us better understand how powerful QMA protocols can be. Today, we show that $\mathsf{QMA} \subseteq \mathsf{PP}$, which we now formally define.

Definition 1.1. A computational problem $\{\mathcal{I}, \mathcal{O}, f\}$ is in the complexity class PP if

- For every valid input $x \in \mathcal{I}$, the output $f(x)$ is either 0 or 1.
- There is a (uniform and classical) polynomial-time randomized algorithm A such that for every $x \in \mathcal{I}$,
$$\text{Prob}[A(x) = f(x)] > 0.5.$$
This probability is taken over the random coins of algorithm A.

Remark 1.2. This class looks very similar to BPP, but this is not the case. Note that algorithm A can have advantage only exponentially small over a coin-flip. Thus, one cannot boost the error probability arbitrarily close to 0 with polynomial many instantiations.

1.1 Warmup: NP contained in PP

We start by showing a weaker result, namely that $\mathsf{NP} \subseteq \mathsf{PP}$.

Lemma 1.3. $\mathsf{NP} \subseteq \mathsf{PP}$

Proof. Suppose $(\mathcal{I}, \{0,1\}, g) \in \mathsf{NP}$, where $g : \mathcal{I} \to \{0,1\}$. By definition, there exists a (univariate) polynomial p and polynomial-time verifier $V : \{0,1\}^n \times \{0,1\}^{p(n)} \to \{0,1\}$ and $\forall x \in g^{-1}(1)\ \exists y \in \{0,1\}^{p(|x|)}$ such that $V(x,y) = 1$. Furthermore, $\forall x \in g^{-1}(0)\ \forall y \in \{0,1\}^{p(|x|)}$, $V(x,y) = 0$.

It suffices to specify a polynomial time randomized algorithm $A(x)$. Flip a fair coin. On heads, output 1 with probability $1 - \frac{0.5}{2^{p(|x|)}}$ and 0 otherwise. On tails, sample $y \in_R \{0,1\}^{p(|x|)}$ uniformly at random and output $V(x,y)$.

If $g(x) = 1$, then the existence of a witness y implies

$$\text{Prob}[A(x) = 1] \geq \frac{1}{2}\left(1 - \frac{0.5}{|\{0,1\}^{p(|x|)}|} + \frac{1}{|\{0,1\}^{p(|x|)}|}\right) \geq \frac{1}{2}.$$

Conversely, if $g(x) = 0$, then the non-existence of witness y implies

$$\text{Prob}[A(x) = 0] \geq \frac{1}{2}\left(1 + \frac{0.5}{|\{0,1\}^{p(|x|)}|}\right) \geq \frac{1}{2}.$$

□

Remark 1.4. It is worth noting that the PP randomized algorithm A only needs to succeed with probability more than $\frac{1}{2}$, but it cannot be *arbitrarily* close to $\frac{1}{2}$. This is because A is still a polynomial time algorithm. So, it can only flip polynomially many coins, and there are only $2^{\mathsf{poly}(n)}$ many possible computation paths. Thus, it would be reasonable to define PP such that $\text{Prob}[A(x) = f(x)] \geq \frac{1}{2} + \left(\frac{1}{2}\right)^{\mathsf{poly}(n)}$.

2 QMA is contained in PP

Let $\{\mathcal{I}, \{0,1\}, f\}$ be a decision promise problem in QMA. Consider the corresponding QMA verification algorithm Q that takes n input qubits, a ancilla qubits, and m witness qubits. Citing the Marriott-Watrous soundness amplification result from last lecture, we can assume WLOG that the Q satisfies:

- If $f(x) = 1$, there is some witness making Q accept with probability at least $1 - 2^{-(m+2)}$.
- If $f(x) = 0$, for all witnesses, Q accepts with probability at most $2^{-(m+2)}$.

Lemma 2.1 (Warmup)**.** *There is a quantum polynomial-time algorithm A such that for all $x \in \mathcal{I}$,*

$$\text{Prob}[A(x) = f(x)] > \frac{1}{2}.$$

Proof. The algorithm A is as follows. Flip a fair coin. If heads, output 1 with probability $1 - \frac{1}{2^{m+1}}$ and 0 otherwise. If tails, run Q using the maximally mixed state in the witness register (and output the measurement in the standard 0/1 basis).

If $f(x) = 1$, then

$$\text{Prob}[A(x) = 1] \geq \frac{1}{2}\left(1 - \frac{1}{2^{m+1}}\right) + \frac{1}{2}\left(\frac{1}{2^m} \cdot (1 - 2^{-(m+2)})\right) > \frac{1}{2},$$

where the tails probability comes from the fact that the promised "good witness" occurs with probability 2^{-m} in the maximally mixed state.

If $f(x) = 0$, then

$$\text{Prob}[A(x) = 1] \leq \frac{1}{2}\left(1 - \frac{1}{2^{m+1}}\right) + \frac{1}{2}\left(\frac{1}{2^{m+2}}\right) < \frac{1}{2},$$

□

Remark 2.2. The one-copy of the witness soundness amplification of Marriott-Watrous was crucial to the above proof. We needed completeness and soundness error less than 2^{-m}, where m is the number of witness qubits. Naive boosting, with multiple copies of the same witness, would not have sufficed!

We leave as an exercise to the reader details showing that there is also a randomized polynomial time algorithm A' that shows $\{I, \{0,1\}, f\} \in \mathsf{PP}$. The sketch is as follows. Polynomial time quantum algorithm A can be simulated in exponential time and space by maintaining the full joint state of all qubits at each timestep. Using depth-first traversals, you can modify this simulation to use polynomial space (and exponential time). Finally, you can pick two random branches of this simulation to get a biased bit in the direction of whether quantum algorithm A is more likely to accept or more likely to reject. More details can be found on pages 11 to 14 of Ryan O'Donnell's notes from Lecture 24 of the CMU Quantum Information Theory course.

3 Local Hamiltonians

We will now introduce Local Hamiltonians and the Local Hamiltonian Problem. This problem is very natural and it is complete for QMA.

Definition 3.1. Recall the Pauli operator set $\{X, Y, Z, I\}$. An n-qubit Pauli operator is the n-dimensional tensor of (1-qubit) Pauli operators. We define P to be a *weight k (n-qubit) Pauli operator* if $P = U_1 \otimes \ldots \otimes U_n$ such that $|\{i : U_i \in \{X, Y, Z\}| = k$.

Definition 3.2. A k-local Hamiltonian is any Hermitian matrix H_n which can be written as the sum of weight k Pauli matrices, i.e. $H_n = \sum b_\alpha P_\alpha$.

Definition 3.3. For a Hamiltonian H and quantum state ρ (density operator), the *energy of the state* is just the value $E(\rho) = \mathrm{Tr}(\rho H)$. For pure states $\rho = |\psi\rangle\langle\psi|$, $E(\rho) = \langle\psi| H |\psi\rangle$. The notation $E_0(H) := \min_\rho\{E(\rho H)\}$ denotes the *ground energy* of the (physical system associated with) Hamiltonian H. This minimizer is called the ground state.

Fact 3.4. The ground energy of Hamiltonian H is its smallest eigenvalue.

Definition 3.5. The *k-local Hamiltonian Problem* (k-LHP) is a promise decision problem parameterized by $k \in \mathbb{N}$ and two functions $a, b : \mathbb{N} \to \mathbb{N}$ satisfying $a(n) < b(n)$ for all n. The inputs are k-local n-qubit Hamiltonians whose smallest eigenvalue is either less than $a(n)$ or more than $b(n)$. Formally,

$$\mathcal{I} = \{H : E_0(H) \notin [a(n), b(n)]\},$$

and

$$f(H) = \begin{cases} 1 & E_0(H) \leq a(n) \\ 0 & E_0(H) \geq b(n) \end{cases}.$$

Detour One might ask why this problem arises naturally in computer science. In the physical world, particles interact only locally, and Hamiltonians capture the time evolution of such systems, so it is reasonable to expect them to reflect locality as well. As we will explore in this lecture and the next, there is a deep relationship between the locality of quantum circuits and that of Hamiltonians.

3.1 Using Quantum Phase Estimation for Ground Energy

Lemma 3.6. *k-$\mathsf{LHP}(a, b)$ with $b(n) - a(n) = \Omega(\frac{1}{\mathsf{poly}(n)})$ is in QMA.*

Quick Summary. We give a sketch of the proof idea. The witness is the ground state (read: eigenstate) associated with the ground energy (read: lowest eigenvalue) of the inputted Hamiltonian. The verifier circuit performs quantum phase estimation, which has some error, but this is why we have the non-negligible gap in the a, b promise parameters.

More Background. Quantum Phase Estimation (QPE) is a procedure for extracting the eigenphase[1] of a unitary operator. Given a unitary U with eigenvector $|v\rangle$ and eigenvalue $e^{2\pi i\theta}$ for some $\theta \in [0, 1)$, QPE maps

$$|0\rangle^{\otimes n} \otimes |v\rangle \mapsto \left|\tilde{\theta}\right\rangle \otimes |v\rangle ,$$

where the first register encodes an estimate $\tilde{\theta}$ of θ using n ancilla qubits.

Precision. If $2^n\theta$ is an integer, then after applying the inverse quantum Fourier transform (QFT) the measurement register collapses to the computational basis state

$$|2^n\theta\rangle .$$

If $2^n\theta \notin \mathbb{Z}$, the register instead becomes a superposition

$$\sum_{k=0}^{2^n-1} \alpha_k |k\rangle , \qquad \alpha_k = \frac{1}{2^n} \frac{1 - e^{2\pi i(2^n\theta - k)}}{1 - e^{2\pi i(\theta - k/2^n)}},$$

whose probability distribution is sharply peaked around $k \approx 2^n\theta$. Thus, increasing n narrows the distribution and improves the precision of the phase estimate.

Hamiltonians and Unitaries. For a Hermitian matrix (Hamiltonian) H with eigenvector $|v\rangle$ and eigenvalue λ,

$$H |v\rangle = \lambda |v\rangle ,$$

define the unitary

$$U = e^{iHt}.$$

Then

$$U |v\rangle = e^{i\lambda t} |v\rangle ,$$

so the QPE phase is

$$\theta = \frac{\lambda t}{2\pi} \mod 1.$$

Ground Energy Estimation. By preparing a state with nonzero overlap on the ground eigenvector $|v_{\min}\rangle$, QPE applied to $U = e^{iHt}$ yields an estimate of the corresponding eigenphase $\theta_{\min}$, from which one recovers the ground energy as

$$\lambda_{\min} \approx \frac{2\pi}{t} \theta_{\min}.$$

Thus, QPE provides a direct route from time-evolution unitaries to eigenvalue (energy) estimation of Hamiltonians.

Remark 3.7. This discussion included time parameter t, which can be set arbirarily (e.g. $t = 1$).

[1] Eigenphase is the phase associated with an eigenvalue of a unitary operator. Unitary operators have eigenvalues of the form $e^{2\pi i\theta}$ for $\theta \in [0, 1)$ because they are norm preserving.

4 Circuit to Hamiltonian

We now describe a powerful technique mapping computation to local Hamiltonians.

4.1 Cook-Levin Recap: Mapping classical computation to CSPs

Recall that the Cook-Levin theorem showed that SAT is NP-complete. We should interpret this result as showing that *dynamic* computations can be mapped to *static* constraint satisfaction problems (CSPs) where the output of the computation is read from a bit of a satisfying assignment. As an example, consider the following classical (reversible) circuit in Figure 1.

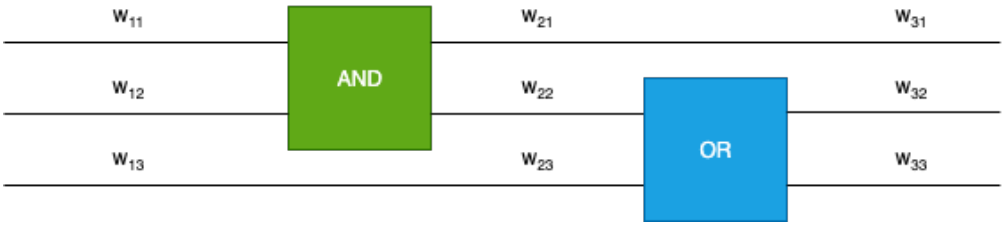

Figure 1: A nine wire two-gate reversible circuit.

The idea of Cook-Levin is to use a computation "tableau". We can think of $w_{t,i}$ as the value of the i-th wire at time t, where time is measured in number of gates applied. We then apply constraints that relate at most $2k$ of the total wire set, where k is the maximum arity of any gate in the original computation. See Figure 2 below.

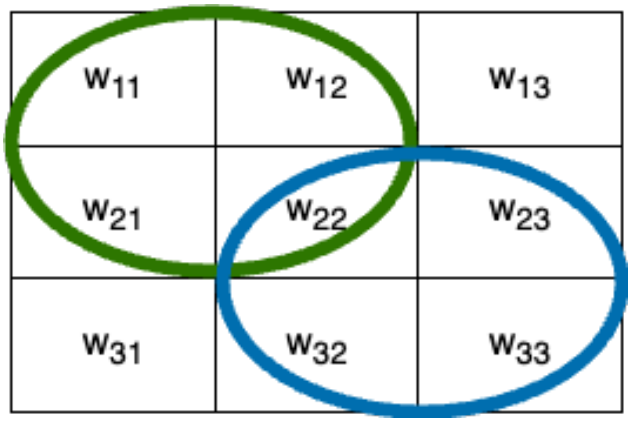

Figure 2: The green constraint ensures that $\mathsf{AND}(w_{1,1}, w_{1,2}) = (w_{2,1}, w_{2,2})$. The blue constraint ensures that $\mathsf{OR}(w_{2,2}, w_{2,3}) = (w_{3,2}, w_{3,3})$. There are also implicit "copy constraints" to capture the time evolution of wires that were not impacted at some timestep (e.g. ensure that $w_{1,3} = w_{2,3}$).

Remark 4.1. This is sometimes called a *gadget reduction* in the literature.

Remark 4.2. A small technical note about the presentation of the gadget reduction above. We described a mapping whereby the CSP is always satisfiable, and you interpret the computation's output by reading a bit of the satisfying assignment. This was useful because we could think about

the *satisfying assignment as an encoding of the computation trace*. However, sometimes it is more useful to relate the CSP's satisfiability to the existence of an input that yields output 1. This change is simple; we add a single extra constraint on the output bit. We will switch between these pespectives when convenient.

4.2 Mapping CSPs to Local Hamiltonians

We claim that a constraint satisfaction problem (CSP) can be mapped to a local Hamiltonian H. This mapping can be seen as a gadget reduction: introduce one qubit for each variable, and translate each constraint into a local Hamiltonian term.[2] The overall Hamiltonian is then

$$H = \sum_i H_i.$$

Example. Consider a CNF clause $C_i := x_1 \vee x_2 \vee \neg x_3$. The associated local Hamiltonian term should penalize any state that has amplitude on $|001\rangle_{1,2,3}$. Formally, we define

$$H_i := |001\rangle\langle 001|_{1,2,3} \otimes I^{\otimes(n-3)}.$$

This operator assigns an energy penalty of 1 to any computational basis state with $(x_1, x_2, x_3) = (0, 0, 1)$. Thus the ground space of H_i corresponds exactly to the satisfying assignments of clause C_i.

Remark 4.3. This example easily generalizes to any width-k CNF clause. The associated Hamiltonian term will have width k. Indeed, the terms (and the overall Hamiltonian) will be diagonal in matrix representation.

Combined with the tableau reduction from classical computation to CSPs, we have argued that there is an $O(1)$-local Hamiltonian that has ground energy 0 iff some input yields an output of 1. Moreover, each gate can be encapsulated by a local Hamiltonian term. Can there be a similar mapping from *quantum* computation to local Hamiltonians?

5 Quantum Computation to Local Hamiltonian

5.1 Local Indistinguishability

We now argue that a direct translation of the classical technique is destined to fail. Indeed, due to entanglement, even state transition of a 1-qubit gate cannot be represented in a local Hamiltonian.

Consider the following quantum states:

$$|\mathsf{CAT}_+\rangle := \frac{1}{\sqrt{2}}\left(|00\cdots0\rangle + |11\cdots1\rangle\right)$$
$$|\mathsf{CAT}_-\rangle := \frac{1}{\sqrt{2}}\left(|00\cdots0\rangle - |11\cdots1\rangle\right)$$

These quantum states can be prepared by circuits and have the following properties:

[2] In fact, these Hamiltonian terms will be diagonal in the computational basis.

- We can switch from $|\mathsf{CAT}_+\rangle$ to $|\mathsf{CAT}_-\rangle$ by applying a Z gate to any qubit.
- If we trace out[3] any one qubit, the two states are the same. Without loss of generality, tracing out qubit 0,

$$\begin{aligned}
&\mathrm{Tr}_i\left(|\mathsf{CAT}_+\rangle\langle\mathsf{CAT}_+|\right)\\
&= \mathrm{Tr}_i\left(\frac{1}{2}|00\cdots0\rangle\langle 00\cdots0| + \frac{1}{2}|11\cdots1\rangle\langle 11\cdots1|\right)\\
&= \mathrm{Tr}_i\left(|\mathsf{CAT}_-\rangle\langle\mathsf{CAT}_-|\right)
\end{aligned}$$

This phenomenon is called *local indistinguishability*. There is no *local* way to distinguish between the two states, and hence the action of the Z gate, if we insist on a local Hamiltonian term acting on the qubits in the states.

5.2 Next Time: Feynmann-Kitaev clock construction

Next class, we will discuss the Feynmann-Kitaev clock construction and how it addresses the local indistinguishability issue.

[3]Recall that tracing out a qubit or any subsystem requires thinking about the quantum state as a density operator. If we have a pure state $|\psi\rangle$, the corresponding density operator is $\rho := |\psi\rangle\langle\psi|$.