QMA **and** QCMA *Instructor: Anand Natarajan, Anurag Anshu*

Lecture 8: Oracle Separations Between QMA and QCMA

September 30, 2025 *Scribe: Andrew Huang*

1 In the real world?

What is at stake here:

Do all useful quantum proof states have efficient classical descriptions?

We believe that $\mathsf{QCMA} \subsetneq \mathsf{QMA}$. However, observe that

$$\mathsf{P} \subseteq \mathsf{QCMA} \subseteq \mathsf{QMA} \subseteq \mathsf{PSPACE}.$$

As complexity theory is not currently capable of even proving $\mathsf{P} \subsetneq \mathsf{PSPACE}$, showing an outright separation is unlikely.

2 Oracle separations: what they mean

We are therefore motivated to consider whether we can find partial evidence for such a separation. One such stepping stone would be to ask whether we separate these classes relative to an oracle. What this means is: suppose we give both a QMA machine and a QCMA machine access to some powerful machine $\mathcal{O}$ computing some function, such that each query to $\mathcal{O}$ counts as a unit-cost operation, regardless of the complexity of the function that $\mathcal{O}$ is computing. Can we then prove that there is a problem—however contrived—that a QMA machine can solve with respect to $\mathcal{O}$ but a QCMA machine cannot?

As a warmup, we consider an oracle separation between two classical complexity classes:

Theorem 2.1 ([BGS75])**.** *There are classical oracles $\mathcal{A}$ and $\mathcal{B}$ such that $\mathsf{P}^{\mathcal{A}} = \mathsf{NP}^{\mathcal{A}}$ but $\mathsf{P}^{\mathcal{B}} \neq \mathsf{NP}^{\mathcal{B}}$.*

Proof sketch. For $\mathcal{A}$, it suffices to take an oracle for a PSPACE-complete problem. We then observe that

$$\mathsf{P}^{\mathsf{PSPACE}} = \mathsf{PSPACE} = \mathsf{NPSPACE} = \mathsf{NP}^{\mathsf{PSPACE}}.$$

To construct $\mathcal{B}$, we consider the problem of deciding whether an oracle $\mathcal{O}$, restricted to inputs of length n, always outputs 0 (which is our NO case), or outputs 1 on a single input x (which is our YES case). It is easy to see that this problem is in $\mathsf{NP}^{\mathcal{O}}$ since the witness is simply the input x. The intuition behind why this problem is not in $\mathsf{P}^{\mathcal{O}}$ is that any deterministic machine must look at all inputs of length n to detect a single 1 (which would take exponential time). This idea can be made formal via a standard diagonalization argument. □

Thus, oracle separations don't always say much about the actual underlying complexity classes. However, they *do* tell us something about the techniques needed to analyze the classes. For example, if you wanted to show $\mathsf{QMA} = \mathsf{QCMA}$ and there was a classical oracle separation, then any technique you had for simulating QMA protocols in QCMA must break down in the presence of oracles (it cannot be too "black-box"). In other words, any proof of $\mathsf{QMA} = \mathsf{QCMA}$ would need to be non-relativizing.[1]

One can also interpret an oracle separation in terms of query complexity: think of $\mathcal{O}$ as an exponentially large string (identify its truth table with a binary string). The problem you want to solve has two inputs: an "explicit" input and an "oracle" input. It is often simpler to take the explicit input to just be 1^n, i.e. an encoding of n in unary, so that all of the complexity gets pushed to the implicit input. Our machines are then allowed not regular input access, but only "oracle" or "index" access to locations of the oracle input.

In the classical setting, the query model is pretty straightforward: for an oracle $\mathcal{O}$, on input x, it should return $\mathcal{O}(x)$. What is the right query model in the quantum setting?

In the real world, if you had the source code for $\mathcal{O}$, you could always run it in superposition. So a realistic model should allow for superposition queries to $\mathcal{O}$. This is usually modeled as a unitary which acts on the computational basis states:

$$|x, y\rangle \to |x, y \oplus \mathcal{O}(x)\rangle ,$$

called the standard oracle, or

$$|x, y\rangle \to (-1)^{y \cdot \mathcal{O}(x)} |x, y\rangle$$

called the phase oracle. These two models are equivalent, as one can switch between them by applying Hadamards before and after each query.

As an example of a (classical) oracle separation of quantum complexity classes, the Grover/OR problem (essentially the same one that we saw earlier) gives $\mathsf{BQP}^{\mathcal{O}} \subsetneq \mathsf{NP}^{\mathcal{O}}$.

3 Quantum oracles

The question of whether QMA and QCMA are different was first posed by Aharonov and Naveh:

Question 3.1 ([AN02])**.** Is $\mathsf{QCMA} = \mathsf{QMA}$?

Unfortunately, we don't even know an *oracle* separation between QMA and QCMA. Aaronson and Kuperberg initiated the study of this problem and gave us a *quantum oracle* separation.

Theorem 3.2 ([AK07])**.** *There is a quantum oracle U such that $\mathsf{QMA}^U \neq \mathsf{QCMA}^U$.*

Like a classical oracle, a quantum oracle can be thought of as a unit-cost gate; the difference is that it can implement any unitary. We'll now take a look at how Aaronson and Kuperberg showed their separation.

The AK oracle is a unitary version of Grover; the intuition is that there will be a single hidden marked state embedded into the unitary. To be more specific, the problem is to decide if a unitary U is equal to the identity I (the NO case) or if $U = I - 2|\psi\rangle\langle\psi|$ (the YES case) for some unknown state $|\psi\rangle$.

The QMA upper bound is easy provided you can apply controlled U; the proof is simply $|\psi\rangle$. In particular, this QMA proof system has perfect completeness and soundness.

[1] Many, but not all, of our most frequently used techniques in complexity theory are relativizing, including diagonalization.

The QCMA lower bound is roughly as follows: fix the most popular witness and apply Grover-style techniques. We say σ is p-uniform if for all x, $\sigma(x) \leq \mu_{uniform}(x)/p$. Equivalently: σ is obtained from $\mu_{uniform}$ by conditioning on an event which occurs with probability p.

To prove the lower bound, we need the following geometric lemma from [AK07] about p-uniform distributions:

Lemma 3.3 ([AK07])**.** *Let σ be a p-uniform distribution over dimension-d states, and let ρ be any fixed state. Then*

$$\mathbb{E}_{\psi\sim\sigma}[\langle\psi|\,\rho\,|\psi\rangle] \leq O\left(\frac{1+\log 1/p}{d}\right).$$

With this lemma, we are ready to prove Theorem 3.2:

Proof of Theorem 3.2. Let L be a uniformly random unary language (i.e. for each n, $1^n \in L$ with probability 1/2). For each n, a Haar-random n-qubit state $|\psi_n\rangle$ is sampled. The oracle $U = \{U_n\}_n$ is defined as follows: if $1^n \in L$, then $U_n = I - 2\,|\psi_n\rangle\,\langle\psi_n|$. Otherwise, if $1^n \notin L$, then $U_n = I$.

Note that for all choices of L, $L^U \in \mathsf{QMA}^U$. The proof is simply $|\psi_n\rangle$; upon receiving a state $|\psi\rangle$ for the instance 1^n, Arthur will prepare the state $|+\rangle\,|\psi\rangle$ and apply U controlled on the first register being a 1. He will then apply a Hadamard gate to the first register, measure it, and accept iff he observes a 1. It is not hard to see that this QMA proof system has perfect completeness and soundness for all languages L^U.

Now fix any QCMA machine M (which fixes a polynomial length witness $|w| = p(n)$), and define $S_M(n)$ to be the event that $M^U(1^n, w)$ succeeds: that is, either $1^n \in L$ and there exists a string w such that $\Pr[M^U(1^n, w) = 1] \geq 2/3$, or $1^n \notin L$ and $\Pr[M^U(1^n, w) = 1] \leq 1/3$ for all w.

Claim 3.4. For sufficiently large n, $\Pr[S_M(n) \mid S_M(1), \ldots, S_M(n-1)] \leq 2/3$.

Proof. For each n-qubit pure state $|\psi_n\rangle$, fix the classical witness $w \in \{0,1\}^m$ that maximizes the probability that M accepts, given $U_{|\psi_n\rangle} = I - 2\,|\psi_n\rangle\,\langle\psi_n|$ as an oracle. Let $S(w)$ be the set of states associated with a given witness w. Note that the $S(w)$'s form a partition of all possible n-qubit states, and so there exists a witness w^* such that

$$\Pr[|\psi_n\rangle \in S(w^*)] \geq \frac{1}{2^m}.$$

We now consider the probability that M can distinguisher the unitary $U = I$ from $U_{|\psi_n\rangle}$ for a random $|\psi_n\rangle \in S(w^*)$. Suppose that M makes T queries to the oracle; for $0 \leq t \leq T$, let $|\Phi_t\rangle$ be the final state of M (after all T queries) when the first t queries made were to $U = I$ and the last $T - t$ queries made were to $U = I - 2\,|\psi_n\rangle\,\langle\psi_n|$.

Let ρ_t be the marginal state of the query register just before the t'th query (where the first t queries have been made to $U = I$), and write its decomposition into pure states as $\rho_t = \sum_i p_i\,|\phi_i\rangle\,\langle\phi_i|$. Then we have that

$$\begin{aligned}\|\,|\Phi_t\rangle - |\Phi_{t-1}\rangle\,\|_2 &\leq \sum_i p_i \cdot 2|\langle\phi_i|\psi_n\rangle| \leq \sum_i p_i \cdot 2\sqrt{\langle\psi_n|\phi_i\rangle\langle\phi_i|\psi_n\rangle} \\ &\leq 2\sqrt{\sum_i p_i\langle\psi_n|\phi_i\rangle\langle\phi_i|\psi_n\rangle} = 2\sqrt{\langle\psi_n|\,\rho_t\,|\psi_n\rangle},\end{aligned}$$

by Cauchy-Schwarz and the fact that $U_{|\psi_n\rangle}$ and I behave identically on the components of ρ_t which are orthogonal to $|\psi_n\rangle$.

If we take $|\psi_n\rangle$ over the uniform distribution σ over $S(w^*)$, Lemma 3.3 and Jensen's inequality imply that

$$\mathbb{E}_{|\psi_n\rangle\sim\sigma}[\| \, |\Phi_t\rangle - |\Phi_{t-1}\rangle \, \|_2] \le \mathbb{E}_{|\psi_n\rangle\sim\sigma}[2\sqrt{\langle\psi_n| \, \rho_t \, |\psi_n\rangle}] \le 2\sqrt{\mathbb{E}_{|\psi_n\rangle\sim\sigma}[\langle\psi_n| \, \rho_t \, |\psi_n\rangle]} \le O\left(\sqrt{\frac{m+1}{2^n}}\right).$$

By the triangle inequality, this implies that

$$\mathbb{E}_{|\psi_n\rangle\sim\sigma}[\| \, |\Phi_T\rangle - |\Phi_0\rangle \, \|_2] \le \sum_{t=1}^{T} \mathbb{E}_{|\psi_n\rangle\sim\sigma}[\| \, |\Phi_t\rangle - |\Phi_{t-1}\rangle \, \|_2] \le O\left(T\sqrt{\frac{m+1}{2^n}}\right).$$

Since T and m are both polynomial, this means that for sufficiently large n, M cannot distinguish $|\Phi_T\rangle$ (which corresponds to the YES case) from $|\Phi_0\rangle$ (which corresponds to the NO case) with constant advantage, as desired. □

We thus conclude that

$$\Pr_{L,U}[S_M(1) \wedge S_M(2) \ldots] = 0.$$

Since the number of QCMA machines is countably infinite (by the Solovay-Kitaev theorem), the union bound implies that

$$\Pr_{L,U}[\exists M : S_M(1) \wedge S_M(2) \ldots] = 0.$$

Therefore, $L^U \in \mathsf{QMA}^U \setminus \mathsf{QCMA}^U$ with probability 1. □

4 Towards a classical oracle

We now briefly discuss a series of follow-up works which tried to move closer to a classical oracle separation.

4.1 Fefferman-Kimmel [FK15]

Here, the oracle is an "in-place permutation". That is,

$$\mathcal{O} \, |x\rangle = |\pi(x)\rangle,$$

and you don't have access to an inverse.

The language L is defined as follows: given a permutation over $[N^2] = \{0,1\}^{2n}$, decide whether the preimage of $[N]$ is 2/3 even (the YES case) or 2/3 odd (the NO case), provided one of the two cases is true.

It is easy to see that there is a QMA protocol: the witness is simply the uniform superposition over the preimages of $[N]$. The verifier will test with probability 1/2 each 1) whether the received state is actually the preimage state by applying $\mathcal{O}$ and projecting against the $|[N]\rangle$ state or 2) whether a measured element in the received state is odd and a valid preimage.

Showing that there is no QCMA protocol outright is not known, so Fefferman and Kimmel resort to randomization: when the prover is coming up with its witness, we restrict it to knowing only the set $\pi^{-1}([N])$. The oracle gets to choose the permutation π (consistent with this preimage set) randomly at runtime *after* Merlin has sent his witness; thus, an oracle is fully specified only by a *set* S with size $|S| = N$. With this slightly non-standard restriction, [FK15] is able to rule out any QCMA protocols by using combinatorial sunflowers and the adversary bound.

4.2 Natarajan-Nirkhe [NN24]

Here, the oracle computes the adjacency matrix of a graph, and the problem is to decide if the graph is disconnected (the YES case) or spectrally expanding (the NO case). This problem is in QMA because of a straightforward random walk algorithm: first, one checks that the random walk preserves the witness, and secondly that this witness is orthogonal to the uniform superposition state, which together imply the existence of at least two connected components. Like [FK15], this oracle is "randomized" in some sense: there is a YES and NO distribution of graphs, and the witness in both the quantum and classical setting is only allowed to depend on the distribution (rather than the specific graph instance sampled).

Despite your intuition from stoquastic Hamiltonians, this problem is not in MA. This is since our question is about the first excited state, not the ground state. This problem is also not in BQP, as shown by Ambainis, Childs, and Liu.

To show that this problem is not in QCMA, the intution is similar to the [FK15] setting: fix an optimal witness and find a sunflower (a collection of YES instances with perfect overlap in the core, and very little overlap outside the core).[2] We can then hardwire the witness and use the verifier to show this sunflower is indistinguishable from an "ideal sunflower". Finally, one can show that the ACL bound holds on the ideal sunflower (because of symmetry of this distribution). One can get the core size to be $\text{poly}(n)$, and each non-core point to have probability $1/N^{0.99}$.

4.3 Liu-Mutreja-Yuen [LMY25]

[LMY25] tried to push as far as possible to eliminate randomization from the oracle. They rely on the following conjecture: quantum algorithms cannot distinguish pseudorandom distributions ($x \in \{0,1\}^n$ is δ-dense if for all subsets $S \subseteq n$ of coordinates, $H_{min}(x|_S) \geq (1-\delta)|S|$) from uniformly random.

This conjecture is related to the long-standing Aaronson-Ambainis conjecture (that quantum and classical query complexities are polynomially related for random distributions). If this conjecture holds, then [LMY25] shows that the [NN24] oracle directly gives a separation (without the witness assumption).

4.4 Zhandry [Zha24]

The recent work of [Zha24] takes a different oracle construction and lower-bound approach.

The oracle at hand is a membership oracle for two sets S and U. In the YES case, there is a state $|\psi\rangle$ supported on S, such that $|\hat{\psi}\rangle = \mathcal{F}\,|\psi\rangle$ has moderately high overlap with U:

$$\langle\hat{\psi}|\,\Pi_U\,|\hat{\psi}\rangle \geq \frac{1}{2} + \underbrace{\gamma(n)}_{1/\text{poly}(n)} .$$

[2] To fix the issue of the QMA witness depending on the graph, one supplies a state which depends only on the core of the sunflower (which is close enough to the ideal witness).

In the NO case, the overlap with all $|\psi\rangle$ supported on S is below $1/2 + \gamma(n)/2$.

[Zha24] relies on the following conjecture: one can pick S to have (fractional) size smaller than any inverse-poly, and fixing S, one can pick U at random such that

- U is k-wise uniform for super-polynomial k, and
- (S, U) is a YES instance with $\gamma = \frac{1}{\mathrm{poly}(n)}$, with probability very close to 1 over choice over U.

This follows from a conjecture which Zhandry proposes about rounding approximately k-wise uniform distributions to exactly k-wise uniform.

Suppose this problem is in QCMA. Then we will construct a NO instance S', U that is also accepted by the verifier. The idea is to fix a witness for a YES instance and run the QCMA verifier a super-polynomial number of times. This can be used to sample "heavy" query points from S. Let S' be the small set you sample this way. Using a Grover-type bound, one can show that you cannot sample many more points this way than the size of the witness. Also, this procedure hits all the points that the verifier was likely to query, so the verifier cannot distinguish between the new small set S' and the original set S. However, S' is so small that the problem is forced to be a NO instance (as its Fourier transform is very close to uniform).

In this argument, one really needed U to not leak info about S. This is what the exact k-wise uniformity (from the conjecture) gives us.

References

[AK07] Scott Aaronson and Greg Kuperberg. Quantum versus classical proofs and advice. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 115–128. IEEE, 2007.

[AN02] Dorit Aharonov and Tomer Naveh. Quantum np-a survey. *arXiv preprint quant-ph/0210077*, 2002.

[BGS75] Theodore Baker, John Gill, and Robert Solovay. Relativizations of the p=?np question. *SIAM Journal on computing*, 4(4):431–442, 1975.

[FK15] Bill Fefferman and Shelby Kimmel. Quantum vs classical proofs and subset verification. *arXiv preprint arXiv:1510.06750*, 2015.

[LMY25] Jiahui Liu, Saachi Mutreja, and Henry Yuen. Qma vs qcma and pseudorandomness. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1520–1531, 2025.

[NN24] Anand Natarajan and Chinmay Nirkhe. A distribution testing oracle separation between qma and qcma. *Quantum*, 8:1377, 2024.

[Zha24] Mark Zhandry. Toward separating QMA from QCMA with a classical oracle. *arXiv preprint arXiv:2411.01718*, 2024.