Contextuality for Quantum Advantage Tutorial

Eric R. Anschuetz[1,*]

[1] *MIT Center for Theoretical Physics, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

I. QUANTUM MACHINE LEARNING

As noisy intermediate-scale quantum (NISQ) devices are starting to be built, there are many questions as to what kinds of algorithms can be run on these devices. One of the most popular proposals are *quantum machine learning* algorithms. There are many proposals for this, but we are only concerned about one:

	Classical problems	Quantum problems
Classical model	Machine learning	Classical shadows, only study constant-sized reduced density matrices
Quantum model	Variational quantum algorithms	Quantum random access memory (QRAM) reliant

TABLE I. Types of quantum machine learning algorithms. We are here mainly considered with variational quantum algorithms here.

We will call these *variational quantum algorithms* (VQAs) [1] or *quantum machine learning* (QML) algorithms for the remainder of this talk.

II. VARIATIONAL QUANTUM ALGORITHMS

These VQAs use some parameterized quantum state $|\boldsymbol{\theta}\rangle$ prepared on a quantum computer to minimize some loss function $f(\boldsymbol{\theta})$. Some examples of loss functions include:

$$f_{\text{VQE}}(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle, \tag{1}$$

$$f_{\text{CE}}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{v}} \log\left(\langle \boldsymbol{\theta}| \boldsymbol{\Pi}_{\boldsymbol{v}} |\boldsymbol{\theta}\rangle\right). \tag{2}$$

Many problems can be phrased as minimization problems. Combinatorial optimization problems, chemistry problems, machine learning problems, and so on. Of course, the performance of such an algorithm depends completely on the ease of optimizing loss functions of these forms. What is known about this?

It is apparent from these results that generally, variational quantum algorithms are untrainable. To get an idea of how bad this is, look at a picture of a loss landscape for a quantum convolutional neural network (QCNN) in the slides. So, what can we do to fix this?

The answer is to study *restricted* classes of variational quantum algorithms, as the above untrainability results were all constructed from looking at generic variational quantum algorithms. Part of the motivation for considering generic variational quantum algorithms in the first place is we could appeal to complexity theory to talk about their power [8]. Thus, when considering restricted classes of quantum algorithms, the important question is: do restricted classes exist that still maintain any advantage over classical machine learning models?

It turns out, the answer is yes! And, as a side-effect of not being able to appeal to complexity theory, we will be able to point to the exact quantum phenomenon that yields the separation—in other words, our proofs are completely constructive. A negative of not being able to appeal to complexity theory is that we are only able to explicitly show a polynomial separation, but we also show that with minimal additional resources one is able to recover the exponential separation (assuming standard complexity theory assumptions).

* eans@mit.edu

III. QUANTUM CONTEXTUALITY

The quantum resource that we will use to prove these separations is called *quantum contextuality*. If you have not heard of this before, it is essentially a generalization of nonlocality as a quantum phenomenon. Here, we will be focused on so-called "state-independent" contextuality. Let us give an example in terms of Pauli operators. Consider the operators in the following Table, called a Mermin–Peres magic square [9]:

X_1	X_2	X_1X_2	$= 1$
X_1Z_2	Z_1X_2	$-X_1Z_1X_2Z_2$	$= 1$
Z_2	Z_1	Z_1Z_2	$= 1$
$= 1$	$= 1$	$= -1$	

TABLE III. An example of quantum contextuality using a Mermin–Peres magic square [9]. All operators in each row and column commute. Additionally, the product of each row and column is the identity operator, except for the final column, which gives -1. Thus, definite classical values cannot be assigned to each operator without yielding a contradiction.

Let us try and assign definite classical values to the these operators. Consider the following assignment:

1	1	1	$= 1$
1	1	1	$= 1$
1	1	1	$= 1$
$= 1$	$= 1$	$= 1$	

TABLE IV. An example of an incorrect classical assignment.

Can we fix this assignment? Let us try flipping one of the signs:

1	1	1	$= 1$
1	1	1	$= 1$
1	1	-1	$= -1$
$= 1$	$= 1$	$= -1$	

TABLE V. An example of an incorrect classical assignment.

Whoops, another error. Let us try and fix it again:

1	1	1	$= 1$
1	1	1	$= 1$
1	-1	-1	$= 1$
$= 1$	$= -1$	$= -1$	

TABLE VI. An example of an incorrect classical assignment.

It is easy to see that any classical attempt at assigning values to these operators is an exercise in futility. In some sense, the "correct" way to do this classically would be to e.g. assign the value 1 to Z_1Z_2 when measuring it with Z_2 and Z_1, but then assign it -1 when measuring it with X_1X_2 and $-X_1Z_1X_2Z_2$. This is why this phenomenon is called *contextuality*—the measurement result of a given observable depends on other observables that were previously measured. For concreteness, here where we say "measured," we mean via performing phase estimation of this operator applied to some given input state.

IV. QUANTUM CONTEXTUALITY GIVES MEMORY BOUNDS ON CLASSICAL SIMULATION OF QUANTUM PROCESSES

We can immediately see how this can potentially give a memory separation between classical and quantum processes. Let us attempt to classically simulate a process given by measuring Pauli operators on two qubits when beginning in a stabilizer state. We do not require this simulation to be "faithful" (i.e. do not require "strong" quantum simulation), just that the measurement results are consistent with quantum mechanics (this is weaker than even what most people call "weak" quantum simulation); we will call this "very weak" simulation. A quantum model can of course do this with just two qubits of memory: just perform sequential measurements of the operators on the two qubit quantum state. Observe that the 1-qubit version of this problem can be done with 1 bit of memory; 0 can represent $|+\rangle, |+\mathrm{i}\rangle, |0\rangle$, and 1 can represent $|-\rangle, |-\mathrm{i}\rangle, |1\rangle$ in our "very weak" simulation model.

First, we note that the even among stabilizer states, $|{+}{+}\rangle, |{+}{-}\rangle, |{-}{+}\rangle, |{-}{-}\rangle$ are perfectly distinguishable (just measure X_i for $i = 1, 2$), and thus getting these measurements correct already requires two bits of memory. One can even "cheat" and reuse storage. For example, let:

- $|{+}{+}\rangle, |{+}\mathrm{i}+\mathrm{i}\rangle, |00\rangle$ have classical representation 00,
- $|{+}{-}\rangle, |{+}\mathrm{i}-\mathrm{i}\rangle, |01\rangle$ have classical representation 01,
- $|{-}{+}\rangle, |{-}\mathrm{i}+\mathrm{i}\rangle, |10\rangle$ have classical representation 10,
- $|{-}{-}\rangle, |{-}\mathrm{i}-\mathrm{i}\rangle, |11\rangle$ have classical representation 11.

Then, upon measuring X_i, Y_i, Z_i just output the measurement result given $1 - 2c_i$, where c_i is the classical bit in that position. As there is nonzero probability for all of these measurement results, this simulation is "correct" under the very weak error model we require.

However, let us now consider attempting to double up on memory with the state $\frac{1}{2}(|00\rangle + |01\rangle + |10\rangle - |11\rangle)$, by e.g. attempting to represent this state with the classical representation 00. Now consider the measurement of Z_1Z_2; as $|00\rangle$ is represented by the bit string 00, this measurement result must be 1. The post-measurement state of $\frac{1}{2}(|00\rangle + |01\rangle + |10\rangle - |11\rangle)$ is then $\frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$, and of $|{+}{+}\rangle$ is $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. These are obviously distinguishable states, and it is then easy to see that upon measuring e.g. X_1X_2 one gets distinct measurement results with probability 1. Something similar holds for all of the other classical representations, and it is easy to see at least one more bit is needed to represent $\frac{1}{2}(|00\rangle + |01\rangle + |10\rangle - |11\rangle)$.

How did we find this example seemingly out of nowhere? Well, note that $\frac{1}{2}(|00\rangle + |01\rangle + |10\rangle - |11\rangle)$ is the state stabilized by the operators of the second row of Table III, $|{+}{+}\rangle$ by the first row, and $|00\rangle$ the third row. It turns out that generally, this single-shot distinguishability through a measurement sequence follows from quantum contextuality.

Lemma 1 (Quantum contextuality implies single-shot distinguishability)**.** *Consider three states $|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle$ with stabilizers given by the first three rows of a Mermin–Peres magic square, respectively. Assume we are given a single copy of a state $|\phi\rangle$, and want to determine which of $|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle$ this state is* not *with certainty. This can be done.*

Proof. Consider first measuring $-s_1s_1's_2s_2'$ as in Table VII in the given state $|\phi\rangle$. If this measurement result is -1, then the state is not $|\psi_3\rangle$. If this measurement result is $+1$, then measure s_1s_1' as in Table VII. As the post-measurement state of $|\psi_2\rangle$ is stabilized by s_2s_2' and $-s_1s_1's_2s_2'$, it is also stabilized by $-s_1s_1'$, whereas $|\psi_1\rangle$ is stabilized by s_1s_1'; thus, with certainty these give distinct measurement results upon measuring s_1s_1', and can be distinguished with certainty.

s_1	s_1'	s_1s_1'	$= 1$
s_2	s_2'	s_2s_2'	$= 1$
s_1s_2	$s_1's_2'$	$-s_1s_1's_2s_2'$	$= 1$
$= 1$	$= 1$	-1	

TABLE VII. A generic Mermin–Peres magic square.

□

One consequence of Lemma 1 is if three states share contextual stabilizers, even very weak simulation of Pauli measurements on these three states is impossible if they share the same representation in the classical memory of the

simulator. In other words, by counting how often magic squares occur in the stabilizers of a collection of states, we will be able to lower-bound the memory required for classical simulation of Pauli measurements on stabilizer states!

So, what are a collection of stabilizer states with lots of magic squares? To narrow down the problem a bit, let us assume we use Clifford transformations to switch Pauli frames such that the third row of Table VII is always a computational basis state. What classes of states, when their stabilizers are taken for the first two rows of the table, give a magic square?

It turns out we were on the right track when studying the example in Table III—the answer is graph states! If you are unfamiliar with graph states, they are each associated with an unweighted graph with no self-loops. If the adjacency matrix of the graph is $\boldsymbol{A}$, then the stabilizers of the graph state are $X_i \otimes \bigotimes_{j \neq i} Z_j^{A_{ij}}$. Consider two distinct graph states; by definition, they must differ in an (at least one) edge. Let $|\psi_1\rangle$ not have this edge, and $|\psi_2\rangle$ have it, WLOG. Label the edge where they differ $(1, 2)$ WLOG, and let $\boldsymbol{Z_{kl}}$ be the Z_j on other edges emanating from vertex l for the state $|\psi_k\rangle$. Then we have the magic square:

$X_1\boldsymbol{Z_{11}}$	$X_2\boldsymbol{Z_{12}}$	$X_1X_2\boldsymbol{Z_{11}Z_{12}}$	$=1$
$X_1Z_2\boldsymbol{Z_{21}}$	$Z_1X_2\boldsymbol{Z_{22}}$	$-X_1X_2Z_1Z_2\boldsymbol{Z_{21}Z_{22}}$	$=1$
$Z_2\boldsymbol{Z_{11}Z_{21}}$	$Z_1\boldsymbol{Z_{12}Z_{22}}$	$Z_1Z_2\boldsymbol{Z_{11}Z_{12}Z_{21}Z_{22}}$	$=1$
$=1$	$=1$	$=-1$	

TABLE VIII. An example of a Mermin–Peres magic square associated with graph states.

How many graph states are there? Well, there are $\frac{n(n-1)}{2}$ possible edges, and thus there are $2^{\frac{n(n-1)}{2}}$ graph states. There are only $2^{\frac{n^2}{2}+O(n)}$ stabilizer states [10], so this is about as good as we can expect to do.

V. MACHINE LEARNING INTERPRETATIONS OF MEMORY LOWER BOUNDS

We have shown that quantum contextuality can yield memory lower bounds on the very weak simulation of quantum processes. How can this observation be interpreted as a statement in machine learning? So far, we have only talked about qubits, so let us first consider the discrete version of this statement.

Let us assume we are interested in *translation tasks*; that is, we have an input sequence $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_m})$, and we want to translate this sequence into a "correct" translation $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_m})$ (where we have assumed the input and output sequence lengths are the same for simplicity; this can be achieved with sufficient padding, for instance). Inspired by our previous discussion, let us consider the translation task that is as follows: given a sequence of classical descriptions of Pauli operators, output a sequence of measurement outcomes consistent with quantum mechanics.

Focusing first on qubits, we will consider discrete machine learning models—namely, *Bayesian networks*. These are given by directed acyclic graphs, where the graph defines the conditional relationships of the probability distribution via Bayes' rule. The most famous example of a Bayesian network is probably the *hidden Markov model* (HMM), where a latent variable λ_i exists at each time step, from which there exists an input distribution, output distribution, and a transfer distribution. This is essentially as general a representation of a classical sequence of inputs and measurements on some state space that one can have. Here is an example:

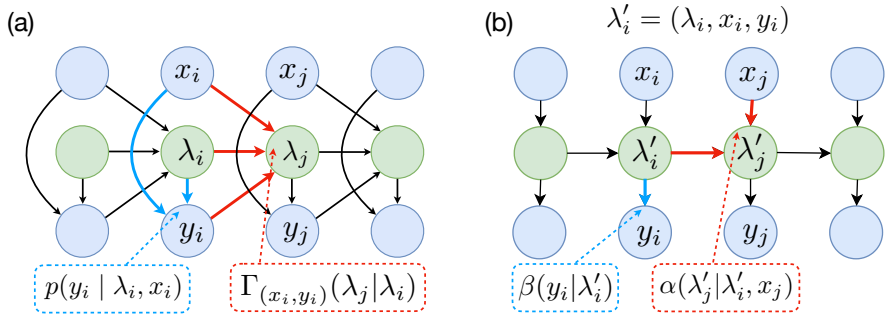

FIG. 1. Diagrams represent hidden Markov models (HMMs).

Note that this essentially automatically gives us a theorem separating the expressive power of basis-enhanced HMMs and HMMs.

Theorem 1 (Hidden Markov models and quantum contextuality)**.** *There exists a family of basis-enhanced 2-gram models with a state space of dimensionality D, that cannot be approximated to finite KL divergence by any classical hidden Markov models in the translation form with a number of hidden units fewer than $D^{\Omega(\log D)}$.*

This is actually a superpolynomial separation in the number of hidden states—where this comes from is the n versus $\sim n^2$ memory separation in terms of (qu)bits, and the fact that the hidden space scales exponentially with these quantities.

We can also state a similar separation for neural networks. There are a lot more things to track because everything becomes continuous rather than discrete, but via similar methods we prove the following theorem:

Theorem 2 (Online stabilizer measurement translation memory lower bound, informal)**.** *There is an n versus $\frac{n(n-3)}{2}$ expressivity separation in the quantum and classical neural network models.*

We prove this theorem by using the continuous-variable quantum mechanical version of Clifford circuits—namely, *Gaussian circuits* (on GKP states). In other words, we measure things like position and momentum operators modulo 2π on an initial infinitely squeezed state. A surprising corollary of this result then is that the addition of the vacuum state makes this model *universal*, and practical implementations of such a network are probably more powerful than the theoretical construction we consider here.

[1] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, Nat. Commun. **5**, 4213 (2014).
[2] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Nat. Commun. **9**, 4812 (2018).
[3] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Nat. Commun. **12**, 1791 (2021).
[4] J. Napp, Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze (2022), arXiv:2203.06174 [quant-ph].
[5] E. R. Anschuetz, in *International Conference on Learning Representations* (2022).
[6] X. You and X. Wu, in *International Conference on Machine Learning* (PMLR, 2021) pp. 12144–12155.
[7] E. R. Anschuetz and B. T. Kiani, Beyond barren plateaus: Quantum variational algorithms are swamped with traps (2022), arXiv:2205.05786 [quant-ph].
[8] Y. Liu, S. Arunachalam, and K. Temme, Nat. Phys. **17**, 1013 (2021).
[9] N. D. Mermin, Phys. Rev. Lett. **65**, 3373 (1990).
[10] S. Aaronson and D. Gottesman, Phys. Rev. A **70**, 052328 (2004).