OVERPARAMETRIZATION IN QAOA

Stephen Zhang*
Department of Mathematics
University of Toronto
stephenn.zhang@mail.utoronto.ca

Wentao Cui*
Department of Physics
University of Toronto
wentao.cui@mail.utoronto.ca

ABSTRACT

We set to study overparmetrization in the quantum alternating operator ansatz (QAOA) as exhibited in Wiersema et al. (2020) and Kiani et al. (2020). We tackle these problems from two fronts, one from a quantum optimal control theory approach and one from a classical deep learning approach. In particular, we try to apply results from optimal control theory to determine the optimization landscape of learning unitaries and we try to adapt the results concerning overparametrization in classical deep learning theory.

1 INTRODUCTION AND PRELIMINARIES

The quantum alternating operator ansatz (QAOA) has found many applications to various problems including some recent work that has involved using QAOA (or some variant of it) as a means to generate unitaries or approximate the ground state energy of a Hamiltonian (Wiersema et al., 2020; Kiani et al., 2020). The QAOA was introduced as an extension of the quantum approximate optimization algorithm, which is also abbreviated as QAOA[1] (Hadfield et al., 2019).

In the paper of Kiani et al. (2020), they set out to determine the convergence of gradient descent to a d-dimensional target unitary using an alternating ansatz of the form

$$e^{-iAt_K}e^{-iB\tau_K}\cdots e^{-iAt_1}e^{-iB\tau_1}$$

where A, B are sampled from a Gaussian Unitary Ensemble (GUE). Using the Frobenius norm as their loss, their results showed that gradient descent typically converges once the the number of parameters is at least d^2. Interestingly enough, this dependence on the number of parameters did not depend on whether or not the target unitary was Haar random or of the form

$$e^{-iAt_N}e^{-iB\tau_N}\cdots e^{-iAt_1}e^{-iB\tau_1}$$

where $2N << d^2$. Therefore, the problem requires at least d^2 parameters in order for gradient descent to learn.

In Wiersema et al. (2020), a similar behavior was observed. Their paper was focused on the problem of approximating the ground state energy of a Hamiltonian H and while they do not use the QAOA, they use the Hamiltonian Variational Ansatz (HVA) which takes on a similar alternating form. They observed the following from experiments on the Transverse-Field Ising Model and the XXZ-model:

- The overparametrized regime is at most $poly(N)$ where N is the system size
- If the parameters are randomly initialized, in the underparametrized regime, while it is still possible to converge to a global minimum, it also might not
- If the parameters are randomly initialized, in the overparametrized regime, it will *always* converge to a global minimum.

* Equal Contribution

[1] For this report, QAOA will refer to the quantum alternation operator ansatz

We would like to demystify the dependence that the optimization process has on the number of parameters.

As the problem of generating unitaries has previously been analyzed from a quantum optimal control setting (Ho et al., 2009), from one front, we set out to determine whether we can apply quantum optimal control techniques to mathematically justify the d^2 threshold exhibited in Kiani et al. (2020). Furthermore, as Kiani et al. (2020) uses the QAOA, we will take a deeper look into the optimality of QAOA in the different context of approximating ground state energy. Section 2 will introduce quantum optimal control, sufficient conditions that guarantee a trap-free optimization landscape, and the optimality of the QAOA and how some results from classical optimal control theory can carry over. Lastly, by running numerical experiments, we will investigate whether or not we can simplify the problem of generating unitaries by considering only traceless A, B.

From the other side, section 3 will focus on casting this problem into the context of deep learning theory. Seeking to minimize the l^2 norm between our learned unitary and the target unitary, our problem has many similarities with training deep linear neural networks–a longstanding field where convergence is well-studied. By using Lie theory and the exponential map to view our problem from the perspective of the Lie algebra $\mathfrak{u}(n)$, we attempt to make the desired relation for A and B restricted to particular commutation relations. Finally, also motivated by deep learning theory we study the effect on convergence rate by instead using natural gradient descent.

2 Quantum Optimal Control Theory

There has been extensive work done in the field quantum optimal control theory to study the existence of local traps in the optimization landscape (Russell et al., 2017; Riviello et al., 2017; Ho et al., 2009). We will dedicate this section to explaining how the problem of learning unitaries with alternating unitaries can be rephrased as a quantum optimal control problem and introduce some of the existing theorems in quantum optimal control theory.

2.1 Introduction to Quantum Optimal Control

We will begin by introducing the general set up of a quantum optimal control problem. Let $\mathcal{H}$ denote a Hilbert space of dimension $N < \infty$. While it is possible to consider the case where $\mathcal{H}$ is infinite dimensional, for a number of problems of interest (including our problem of learning unitaries), it suffices to only work with a finite dimensional Hilbert space (De Fouquieres & Schirmer, 2013). The evolution of this quantum system will be dictated by a unitary operator $U_f(t) \in \mathbf{U}(N)$ satisfying the Schrodinger equation

$$\dot{U}_f(t) = -iH_f(t)U_f(t), \quad U_f(0) = I \tag{1}$$

where H_f is the Hamiltonian of the system that will depend on a *control* f. Typically, these functions f will be taken from $L^2[0,T]$, the space of square integrable functions on the interval $[0,T]$ where $T > 0$ (Ho et al., 2009; De Fouquieres & Schirmer, 2013). Another simplification that is made is that the Hamiltonian $H_f(t)$ will only depend on the control functions in the following manner:

$$H_f(t) = H_0 + \sum_{i=1}^{m} f_i(t)H_i \tag{2}$$

where H_0 is called the drift of the system and $\{H_i\}_{i=1}^m$ are the control Hamiltonians (De Fouquieres & Schirmer, 2013; Morales et al., 2020). We will make the same simplification for the remainder of this report. The objective of a quantum optimal control problem is to minimize

$$\min_{f\in L^2[0,T]} J(V_T(f)) \tag{3}$$

where $J : G \to \mathbb{R}$ is your cost or your fidelity function and $V_T : L^2[0,T] \to G$ is the map $f \mapsto U_f(T)$. G is typically $\mathbf{U}(N)$ or $SU(N)$ depending on the properties of the time evolution $U_f(T)$.

2.1.1 Example: Generating Unitary Transformations

Suppose you are given a target unitary $\mathcal{W}$. The problem of generating unitary transformations can be phrased as the following optimal control problem:

$$\min_{f \in L^2[0,T]} ||\mathcal{W} - V_T(f)||_F^2$$

where $|| \cdot ||_F$ denotes the Frobenius norm. For a more detailed analysis of this problem, (Ho et al., 2009) identify the critical points and analyze the Hessian at those critical points for this problem.

2.2 Optimization Landscapes

In practise, finding a closed-form solution is difficult or impossible so quantum control problems are typically solved by resorting to classical optimization methods such as gradient descent or the downhill simplex method (Moore et al., 2008). Furthermore, oftentimes the space of control functions will need to be constrained to a subspace, which we will denote as $\mathcal{F}$, that is easier to work with (e.g. piecewise constant functions). For this reason, the question of interest is under which conditions can we ensure that the constrained optimization landscape does not contain any *traps*.

Definition 2.1. *(Traps) A trap for a given quantum optimal control problem*

$$\min_{f \in L^2[0,T]} J(V_T(f))$$

is a control function $f \in L^2[0,T]$ such that f is a local optima of $J(V_T(f))$ but not global.

Notice that saddle points are not considered to be traps. The reasoning for this is that the optimization process is rarely attracted to saddle points and this behaviour does not increase with the number of saddle points (Riviello et al., 2017). We will now go over three conditions that when in conjunction, will ensure that the optimization landscape does not contain any local traps (Riviello et al., 2015; 2014).

2.2.1 Controllability

The first of the three conditions is that the global optima is reachable for some time $0 \leq T < \infty$. For example, in the generating unitary transformations problem, if we are given a target unitary $\mathcal{W}$, controllability is the condition that there exists $T < \infty$ such that for any $\epsilon > 0$

$$\min_{f \in L^2[0,T]} ||\mathcal{W} - V_T(f)|| < \epsilon$$

The controllability of quantum control problems has been studied extensively where the techniques typically depend on results in Lie theory (Morales et al., 2020; Altafini, 2002; Wu et al., 2011). Moreover, controllability is likely to be a necessary condition for the optimization landscape to not contain local traps. In particular, Wu et al. (2011) constructed an example of the gate synthesis control problem without the controllability assumption where there will be a guaranteed local sub-optima in the optimization landscape. Depending on the quantum control problem, there are many mathematical definitions of controllability (Albertini & D'Alessandro, 2001). We will only cover one that is most relevant to us called *operator-controllability*.

Definition 2.2. *(Reachable Set) Given a quantum system, a drift Hamiltonian H_0, and a collection of control Hamiltonians $\{H_i\}_{i=1}^m$, the reachable set at time $T > 0$ is defined as*

$$\mathcal{R}(T) = \{\mathcal{W} \in G : \exists f \in \mathcal{F}, \exists U_f(t) \text{ solution of } (1) \text{ and } U_f(T) = \mathcal{W}\}$$

The reachable set is then defined to be $\mathcal{R} = \overline{\cup_{T>0}\mathcal{R}(T)}$.

Definition 2.3. *(Operator-Controllability) We say the quantum system is operator controllable if $\mathcal{R}$ contains every desired unitary (or special unitary) operator.*

For other notions of controllability such as *pure-state-controllability* or *density-matrix-controllability* we refer the reader to the paper by Albertini & D'Alessandro (2001). The most common approach to check for operator-controllability is to verify the *Lie rank algebra condition* (Albertini & D'Alessandro, 2001; Morales et al., 2020; De Fouquieres & Schirmer, 2013). The theorem as presented in (Albertini & D'Alessandro, 2001) is as follows:

Theorem 2.4. *The reachable set $\mathcal{R}$ attainable from identity for the quantum system described by (2) is given by the connected Lie subgroup $e^{\mathcal{L}}$ corresponding to the Lie algebra $\mathcal{L}$ generated by $\{iH_0, iH_1, ..., iH_m\}$.*

Proof. See Albertini & D'Alessandro (2001) Theorem 1. □

While this is one approach of proving controllability, other approaches have been discovered including graph theoretic approaches (Altafini, 2002; Schirmer et al., 2003).

2.2.2 Functional Derivative of V_T is Full Rank

The second condition that is required to hold is that the functional derivative of $V_T : \mathcal{F} \to G$ is of full rank. To provide some insight as to why this condition is required, by the chain rule, we have that

$$\frac{\delta J}{\delta f(t)} = \left\langle \nabla J(V_T), \frac{\delta V_T}{\delta f(t)} \right\rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the Hilbert-Schmidt inner product. Thus, if we assume $\frac{\delta V_T}{\delta f(t)}$ is non-singular

$$\frac{\delta J}{\delta f(t)} = 0 \Leftrightarrow \nabla J(V_T) = 0$$

If $\nabla J(V_T) = 0$, we call the critical point a *kinematic critical point*. At a critical point, if $\frac{\delta V_T}{\delta f(t)}$ is of full rank, we say the critical point is *regular*. Otherwise, we say it is *singular*. While this condition is necessary for theoretical results (Brif et al., 2010), it was shown experimentally that for other optimal control problems, regular critical points dominate the optimization landscape and even if singular local suboptima might exist, they rarely impede the optimization process (Riviello et al., 2014; Wu et al., 2012).

For the case where $G = SU(N)$, Russell et al. (2017) were able to show that a different condition is sufficient. Their key theorem is as follows:

Theorem 2.5. *(Russell et al., 2017) Consider a parameterized family of time dependent Hamiltonians $H[E_n, \lambda_k]$ that are traceless and satisfy equation 2. The condition that V_T is smooth, globally surjective, and transverse to the level sets of J is sufficient to replace the full rank condition of V_T for almost all λ_k (up to a measure zero set of λ_k).*

As remarked in Russell et al. (2017), transversality is generally easier to satisfy than the full rank condition as transversality only requires that the functional derivative of V_T has rank 1 and that ∇J is contained within its range.

2.2.3 Enough Control Resources

The third and final condition is that there are enough control resources for V_T to be globally surjective (Russell et al., 2017). Often times, this condition will be further simplified to just being that the control field is unconstrained. In practise, this condition is very rarely satisfied due to experimental limitations. There are a variety of constraints that you can impose that will impact the optimization landscape. Riviello et al. (2015) ran a number of experiments where they looked at the impacts of restricting the number of control variables, the control period T, and more. The experiments of Riviello et al. (2015) (Fig. 3) showed that there is a fairly sharp threshold value for T for which the optimization fails. Determining this threshold value has been studied for specific problems and is usually called time optimal control (Schulte-Herbrüggen et al., 2005; Khaneja et al., 2001).

Definition 2.6. *(Infimizing Time) Recall the definition of reachable set at time T, $\mathcal{R}(T)$. Given $\mathcal{W} \in G$, the infimizing time of $\mathcal{W}$ is defined as*

$$t^*(\mathcal{W}) = \inf\{T \geq 0 : \mathcal{W} \in \overline{\mathcal{R}(T)}\}$$

For the case where $G = SU(N)$, Khaneja et al. (2001) developed methods to compute the infimizing time for a number of control problems by using results from Lie theory and analyzing something they call the *adjoint control system*.

From the perspective of QAOA and optimal control theory, another way to restrict our control space is by restricting the number of jumps (or bangs) that our control functions can have (Yang et al., 2017; Brady et al., 2020). To the best of our knowledge, determining the optimal number of bangs that a control function should have has yet to be determined and we shall discuss this further in the next section.

2.3 Optimality of the QAOA

We will dedicate this section to explaining the motivation behind the QAOA from an optimal control perspective. While this section will pertain to a different quantum optimal control problem, it will provides some good analysis and insight into the structure of the QAOA. Suppose we are given two Hamiltonians A, B and a state $|x_0\rangle$ that is initially a ground state of A. The objective is to minimize

$$\min_{f \in L^2[0,T]} \langle x_T | B | x_T \rangle$$

where $|x_T\rangle = V_T(f)|x_0\rangle = U_f(T)|x_0\rangle$. Furthermore, equation 2 will be of the form

$$H_f(t) = f(t)A + (1 - f(t))B$$

where $0 \leq f \leq 1$. It was first proposed by Yang et al. (2017) that QAOA is optimal by Pontryagin's minimum principle. We will now summarize their results and their reasoning as shown in Brady et al. (2020). Since equation 1 must hold, we can apply a Lagrange multiplier, $|\lambda(t)\rangle$, to impose the Schrodinger equation constraint. This gives us the function:

$$\phi(f) = \langle x_T | B | x_T \rangle + \int_0^T dt \langle \lambda(t) | \left(-\frac{d}{dt} - iH_f(t) \right) U_f(t) | x_0 \rangle + c.c.$$

where $c.c.$ means the complex conjugate of the previous term. What was shown in Yang et al. (2017) and Brady et al. (2020) is that a number of necessary conditions will be derived by the Lagrange multiplier but most importantly, an application of Pontryagin's minimum principle will give the necessary condition that

$$\frac{\delta \phi}{\delta f(t)} \delta f(t) \geq 0$$

where we only consider the allowed variations of $\delta f(t)$. With this necessary condition, we can determine the behaviour of $f(t)$ by looking at the sign of $\frac{\delta \phi}{\delta f(t)}$

- If $\frac{\delta \phi}{\delta f(t)} = 0$, then we have no constraint on $f(t)$
- If $\frac{\delta \phi}{\delta f(t)} < 0$, then we know we must have $\delta f(t) < 0$ which implies that $f(t) = 1$.
- If $\frac{\delta \phi}{\delta f(t)} > 0$, then we know we must have $\delta f(t) > 0$ which implies that $f(t) = 0$.

Yang et al. (2017) claimed that the case where $\frac{\delta \phi}{\delta f(t)} = 0$ is non-generic in the models that they were considering (SK spin-glass models) so they concluded that the optimal protocol is this bang-bang protocol (f jumps between 0 and 1) which is the form of the QAOA.

On the other hand, Brady et al. (2020) showed that this may not be the case in general. In particular, they showed numerically that for an extended period time in the middle of the time interval $(a, b) \subseteq [0, T]$, you will have that for $t \in (a, b)$, $\frac{\delta \phi}{\delta f(t)} = 0$. In other words, there are regions where the optimal control function may not follow this bang-bang behaviour. Instead Brady et al. (2020) proposed that a better ansatz would be one that is of the form bang-anneal-bang.

2.3.1 Bang-Anneal-Bang Protocal

Brady et al. (2020) proposed that the optimal protocol for the quantum approximate optimization algorithm would be of the form bang-anneal-bang where the *annealing* would occur in the period of time where $\frac{\delta\phi}{\delta f(t)} = 0$. They showed mathematically that for the case of a bounded time T, the optimal control will always start and end with a bang. What this annealing would look like is essentially a smooth transition between A, B instead of these sharp jumps. Furthermore, in the case where the total time T is constrained, QAOA attempts to approximate this annealing section by exhibiting a sequence of much shorter bangs. They backed up their claims numerically with experiments done on transverse-field Ising models where the optimal control functions found using gradient descent did follow this bang-anneal-bang behaviour (see Fig. 1 in Brady et al. (2020)).

2.4 Quantum Optimal Control and Learning Unitaries

We will dedicate this subsection to expressing the problem of learning unitaries by gradient descent into a quantum optimal control problem. Observe that Kiani et al. (2020) are applying QAOA to the optimal control problem of generating unitaries. Of note is that the QAOA applied by Kiani et al. (2020) is not motivated in the same sense as Brady et al. (2020); Yang et al. (2017) as the former allows the parameters τ_i, t_i to be negative. Due to the dependence on requiring at least d^2 parameters, it suggests that at least one of the three conditions mentioned in the previous section is being violated until we reach this d^2 threshold. As mentioned in Kiani et al. (2020), controllability is already satisfied almost surely. Thus, the only conditions of interest are whether the functional derivative of V_T is of full rank and whether we have sufficient control resources. Since restricting the number of parameters is a constraint on the space of controls, it is likely the condition that there are sufficient control resources that is being violated. Specifically, by restricting the number of parameters, in the context of what was shown in 2.3, we are restricting the number of jumps (or bangs) that our control functions can have. Furthermore, we are not constraining the total time T as in the experiments of Brady et al. (2020).

2.4.1 Motivation of Experiments

To verify that the full rank condition (2.2.2) is satisfied, we would like to apply the theorem proven by (Russell et al., 2017). However, their theorem is dependent on $G = SU(N)$ or that the control Hamiltonians are traceless. Thus, it will be interesting to see whether the results still hold if we only consider traceless matrices A, B and whether we can simplify the problem.

2.4.2 Numerical Results

Motivated by the reason above, our numerical experiment aims to answer the following question

- Does a similar dependence on the number of parameters still hold if the control Hamiltonians are taken to be traceless?

2.4.3 Learning with Traceless Hamiltonians

We use the same experimental set up as Kiani et al. (2020) [2] where we add an extreme case of overparametrization, $10.00d^2$. We will make matrices A, B traceless by simply considering the matrices $A' = A - \frac{\mathrm{Tr}(A)}{d} I$ and $B' = B - \frac{\mathrm{Tr}(B)}{d} I$. We ran 10 experiments on 5 qubit systems (i.e. $d = 32$) for each model size with 10,000 steps of vanilla gradient descent and a learning rate of $0.001/\alpha$ where αd^2 is the number of parameters.

As depicted in Figure 1, we get that for the $2d^2$ and $10d^2$ cases, gradient descent finds a local critical point early on in the training process and stays there. In the underparametrized case, we see that the behavior is similar to that of the non-traceless case. However, while the $2d^2, 10d^2$ cases did find

[2] We would like to thank the authors of Kiani et al. (2020) for providing us with their code

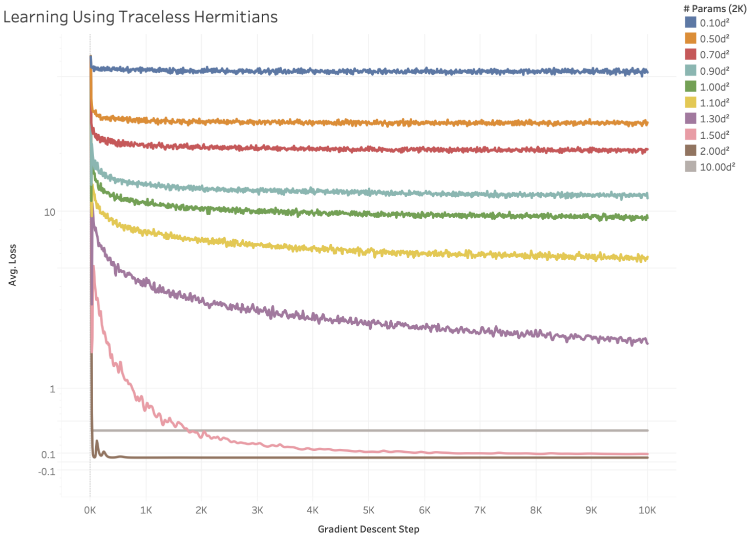

Figure 1: Learning a Haar target unitary using QAOA and traceless control Hamiltonians. The loss axis is logarithmic scale and this is the average loss across all experiments except for the experiment where the $10.00d^2$ converged to a zero-loss critical point.

critical points, only one of these points in the twenty experiments were a zero loss critical point. By considering the extreme case of overparametrization in the $10.00d^2$, it is likely that this is not a problem of the overparameterization threshold being higher. We hypothesize that the reasoning for this is due to the fact that when only considering the randomly sampled traceless matrices, the problem is no longer controllable. This suggests that simplifying this problem in this manner will not work.

3 Overparametrization and Classical Deep Learning Theory

3.1 Introduction to Neural Networks

We first introduce the preliminaries of the theory of neural networks. The context of the classic problem is that we are given a set of training data $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{m} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\}$, and we wish to train a function which can statistically predict these results. This predictor is taken from the family of mappings $\mathcal{H} = \{h_\theta : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\}$, with free parameters $\theta \in \boldsymbol{\Theta}$. For the purposes of our investigation, we will focus on linear neural networks. We define the set of linear neural networks with depth N and widths of hidden layers $d_1, \ldots, d_{N-1}$ to be the class of predictors

$$\mathcal{H} = \{\mathbf{x} \to W_N \cdots W_1 \mathbf{x} \mid W_i \in \mathbb{R}^{d_i \times d_{i-1}}, 1 \leq i \leq N\}$$

This can be thought of as an N stage process where at each stage the output of the previous is multiplied by some matrix of dimensions $d_i \times d_{i-1}$, with it's parameters to be somehow determined.

The free parameters of the predictors are learned through gradient descent (or it's many variations) by minimizing the l^2 loss between the prediction of the training instances and the labels to which they correspond:

$$\min_{\theta \in \Theta} L(\theta) := \frac{1}{2m} \sum_{i=1}^{m} ||h_\theta(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)}||^2$$

By gradient descent, we mean to say that we approach a (local) minimum of the loss function by iteratively and incrementally adjusting the learning parameters in the direction of steepest local descent. Denoting by $W_i(t)$ the state of W_i at iteration t,

$$W_i(t+1) \leftarrow W_i(t) - \eta \frac{\partial L}{\partial W_i}(W_1(t), \ldots, W_N(t))$$

For our linear networks, in the case that the covariance matrix of the dataset $\mathbf{x}$ is equal to the identity (which one can always obtain through a transformation known as whitening Kessy et al. (2018)), we can recast this into the following form Arora et al. (2019)

$$L = \min_{W_i, 1 \leq i \leq N} L(W_1, \ldots, W_N) = \min_{W_i, 1 \leq i \leq N} \frac{1}{2} ||W_N \cdots W_1 - \Phi||_F^2$$

Where F denotes the Frobenius norm between matrices, and $\Phi = \frac{1}{m}\mathbf{y}\mathbf{x}^T \in \mathbb{R}^{d_y \times d_x}$ the cross-covariance matrix. Thus, we have formulated our problem to seek to minimize the Frobenius norm between a target matrix and the product of matrices with learnable weights.

Given a network of this form, there are many results in the literature which guarantee convergence during training with conditions on the weight matrices. We give a particularly cited result below, proved in Arora et al. (2019).

Definition 3.1. *For $\delta \geq 0$, we define that the matrices $W_i, 1 \leq i \leq N$ as given above are δ-balanced provided that*

$$||W_{i+1}^T W_{i+1} - W_i W_i^T||_F \leq \delta \quad \forall 1 \leq i \leq N-1$$

This condition only needs to be true on weight initiation, since provided this it will remain so for all iterations of gradient descent Arora et al. (2018). Furthermore, this is almost surely guaranteed for an initialization via a random Gaussian distribution, as proved in Appendix B of Arora et al. (2019).

Definition 3.2. *Given a target $\Phi \in \mathbb{R}^{d_N \times d_0}$, it's smallest singular value $c_{min}(\Phi)$, and $c > 0$, we define that a matrix $W \in \mathbb{R}^{d_N \times d_0}$ has deficiency margin c with respect to Φ if*

$$||W - \Phi||_F \leq c_{min}(\Phi) - c$$

Intuitively, this can be interpreted as that every matrix at least as close to Φ as W satisfies that all its singular values are at least a distance c from 0.

Finally we have the convergence result:

Theorem 3.3. *If gradient descent on a linear neural network is initiated such that $W_i, 1 \leq i \leq N$ all have deficiency margin $c > 0$ with respect to Φ, and that their initial weights are δ-balanced for some $\delta \propto \frac{c^2}{N^3}$, then a sufficiently small learning rate will guarantee convergence to a global minimum.*

3.2 Neural Networks and QAOA

The above formulation of training classical linear neural networks bears many similarities to the problem of learning unitaries in QAOA, where our goal is to find the times $t_i, \tau_i, 1 \leq i \leq K$ that, for a given target unitary U, will minimize the loss,

$$L = \min_{t_i,\, \tau_i, 1 \leq i \leq K} ||e^{-iAt_K} e^{-iB\tau_K} \cdots e^{-iAt_1} e^{-iB\tau_1} - U||_F^2$$

Since classical deep learning theory is a longstanding field with a variety of powerful pre-proven results, one may be tempted to cast the problem of learning unitaries into a problem of deep learning and apply it's techniques to shed theoretical insight on the former. Ultimately, we seek to explain the overparameterization phenomenon observed in Kiani et al. (2020). However, upon comparison, we immediately see a major problem. Namely, the QAOA is not linear; the resulting unitary is obtained by the successive multiplication of matrix exponentials, where the free parameters to be trained are located in the exponent. Hence, we seek an approach to linearize the problem.

3.2.1 Appealing to $\mathfrak{u}(n)$

The fact that we are grappling with the problem of learning unitaries, for which $U \in U(n)$ opens the door to the rich theory of Lie groups, which we will draw from. Associated to the Lie group $U(n)$, viewed as differentiable manifold, is its tangent space at the identity e. This is known as the Lie algebra $\mathfrak{u}(n)$. As a matrix group (which we will do for the remainder of this section), $\mathfrak{u}(n)$ can be viewed as the set of skew-Hermitian $n \times n$ matrices. That is, $X^\dagger = -X$.

The reason which we appeal to the Lie algebra is due to the exponential map, which maps elements of $\mathfrak{u}(n)$ to elements of $U(n)$. For matrix Lie groups as we have, this reduces to the matrix exponential.

$$\exp(X) = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$$

One can show that this sum is well defined. Furthermore, the fact that $U(n)$ is both compact and connected can be used to show that our exponential map is surjective Djoković (1980). That is, for each $U \in U(n)$. there exists an $X \in \mathfrak{u}(n)$ such that $\exp(X) = U$.

This naturally gives rise to a well-defined matrix logarithm which converges Hall & Hall (2003):

$$\log(U) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} (U - I_n)^k$$

This allows us to recast the problem of learning unitaries in $U(n)$ to a learning problem in $\mathfrak{u}(n)$, as 2 matrices will converge in the Lie group provided their logarithms converge in the Lie algebra. Motivated by Hyland & Rätsch (2017) along with the Baker-Campbell-Hausdoff formula, at this point we consider restricting the form of A and B to satisfy certain commutation relations which will allow the problem to be linearized from the perspective of the Lie algebra. For instance, in the case where A and B are commute, or simultaneously diagonalizable as Hermitian operators from the GUE, the problem can be framed as that of a depth-1 network. Seeing what other commutation relations we can do this for is currently an area of our investigation.

3.3 Using Natural Gradient Descent

In the spirit of neural networks, a natural question is to wonder how the rate of convergence is affected if we change the optimizer. Aside from naive gradient descent on which Kiani et al. (2020) based their analysis, the paper also compared the effects of using the Adam optimizer. They found that although the asymptotic losses were consistent to those obtained through gradient descent, the rate of convergence was much faster.

In a similar vein, we thought it would be interesting to compare these results for those obtained from a natural gradient, motivated by QAOA and VQE-motivated discussions on the quantum natural gradient in Stokes et al. (2020).

In classical deep learning theory, the l^2 geometry given in our case by the Frobenius norm is ill-equipped to weight space due to a general parameter redundancy Neyshabur et al. (2015). However, when viewed as a smooth manifold, the statistical space permits a definition of the Fisher information metric, which is the infinitesimal form of the relative entropy. Natural gradient descent increments the weights in the direction of steepest descent with respect to the geometry defined by this metric (viewing now our space as a Riemannian manifold). One advantage of this approach include that it is invariant under reparameterizations which may distort the space, leading to large plateaus in the l^2 norm and make training very difficult Amari (2000). Hence, when applied to our problem, we expect that the losses to converge faster.

However, upon implementing this optimizer, we found the opposite; the losses took up to 2 times as long to converge when compared to the naive gradient descent optimizer used in Kiani et al. (2020). We suspect this may be due to the selection of learning rate in the natural gradient optimizer. A topic of current investigation involves testing this over a greater range of learning rates spanning from 10^{-1} to 10^{-4}. It is expected that, for the optimal learning rate, the rate of convergence will exceed that of (stochastic) gradient descent. Lastly, just as with the Adam optimizer, the asymptotic losses did not change compared to using gradient descent for given parameter sizes. This can be explained by that the amount of information which can be encoded in the parameters does not change, and although for fewer than $2d^2$ parameters there is not enough freedom to encode the entire unitary, there is still roughly a closest alternative which is found regardless of the optimizer.

4 Conclusion and Future Work

This report attempts to address the dependence on overparametrization that occurs when learning Haar random unitaries with gradient descent. We approached this problem from two sides, the quantum optimal control side and the classical deep learning approach.

From the quantum optimal control approach, we introduced the general setup of quantum optimal control problems as well as three conditions that when all hold true, is sufficient for the optimization landscape to not have any local traps. To close off the quantum optimal control section, we rephrased the problem as a quantum optimal control problem that uses the QAOA and ran some experiments. Our numerical experiment suggests that if the control Hamiltonians A, B are taken to be randomly sampled from a traceless Gaussian unitary ensemble, it is likely that the problem is no longer controllable.

With regards to deep learning theory, we saw explicitly the similarities between minimizing the Frobenius norm between the learned unitary and target unitary in QAOA, compared to training linear neural networks, except weight matrices in the latter are replaced with exponentials in the former. Studying this connection involves linearizing our problem, for which we used Lie theory to view it from the perspective of the Lie algebra for special cases of commutators. Whether this relation holds in general remains to be seen, and is a subject of further study. Finally, we found unituitively that by using natural gradient descent, the loss took more steps to converge.

For future work, from the optimal control side, one approach would be to investigate whether a bang-anneal-bang protocol would also be more suitable for this generating unitaries optimal control problem and whether we can apply Pontryagin's minimum principle to this problem in a similar manner as shown in Section 2.3.

REFERENCES

F. Albertini and D. D'Alessandro. Notions of controllability for quantum mechanical systems. In *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No.01CH37228)*, volume 2, pp. 1589–1594 vol.2, 2001. doi: 10.1109/CDC.2001.981126.

C. Altafini. Controllability of quantum mechanical systems by root space decomposition of su(n). *Journal of Mathematical Physics*, 43:2051–2062, 2002.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10, 11 2000. doi: 10.1162/089976698300017746.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization, 2018.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks, 2019.

Lucas T. Brady, Christopher L. Baldwin, Aniruddha Bapat, Yaroslav Kharkov, and Alexey V. Gorshkov. Optimal protocols in quantum annealing and qaoa problems, 2020.

C. Brif, R. Chakrabarti, and H. Rabitz. Control of quantum phenomena: past, present and future. *New Journal of Physics*, 12:075008, 2010.

Pierre De Fouquieres and Sophie G. Schirmer. A closer look at quantum control landscapes and their implication for control optimization. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 16(03):1350021, 2013. doi: 10.1142/S0219025713500215. URL `https://doi.org/10.1142/S0219025713500215`.

DragomirZ˘ Djoković. On the exponential map in classical lie groups. *Journal of Algebra*, 64(1): 76–88, 1980.

Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor Rieffel, Davide Venturelli, and Rupak Biswas. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms*, 12(2):34, Feb 2019. ISSN 1999-4893. doi: 10.3390/a12020034. URL `http://dx.doi.org/10.3390/a12020034`.

B. Hall and B.C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Graduate Texts in Mathematics. Springer, 2003. ISBN 9780387401225. URL `https://books.google.ca/books?id=m1VQi8HmEwcC`.

Tak-San Ho, Jason Dominy, and Herschel Rabitz. Landscape of unitary transformations in controlled quantum dynamics. *Phys. Rev. A*, 79:013422, Jan 2009. doi: 10.1103/PhysRevA.79.013422. URL `https://link.aps.org/doi/10.1103/PhysRevA.79.013422`.

Stephanie L. Hyland and Gunnar Rätsch. Learning unitary operators with help from u(n), 2017.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, Jan 2018. ISSN 1537-2731. doi: 10.1080/00031305.2016.1277159. URL `http://dx.doi.org/10.1080/00031305.2016.1277159`.

Navin Khaneja, Roger Brockett, and Steffen J. Glaser. Time optimal control in spin systems. *Phys. Rev. A*, 63:032308, Feb 2001. doi: 10.1103/PhysRevA.63.032308. URL `https://link.aps.org/doi/10.1103/PhysRevA.63.032308`.

Bobak Kiani, Seth Lloyd, and Reevu Maity. Learning unitaries by gradient descent. 2020. URL `https://arxiv.org/abs/2001.11897`.

Katharine Moore, Michael Hsieh, and Herschel Rabitz. On the relationship between quantum control landscape structure and optimization complexity. *The Journal of Chemical Physics*, 128(15): 154117, 2008. doi: 10.1063/1.2907740. URL `https://doi.org/10.1063/1.2907740`.

M. E. S. Morales, J. D. Biamonte, and Z. Zimborás. On the universality of the quantum approximate optimization algorithm. *Quantum Information Processing*, 19(9):291, 2020. doi: 10.1007/s11128-020-02748-9. URL https://doi.org/10.1007/s11128-020-02748-9.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *CoRR*, abs/1506.02617, 2015. URL http://arxiv.org/abs/1506.02617.

Gregory Riviello, Constantin Brif, Ruixing Long, Re-Bing Wu, Katharine Moore Tibbetts, Tak-San Ho, and Herschel Rabitz. Searching for quantum optimal control fields in the presence of singular critical points. *Phys. Rev. A*, 90:013404, Jul 2014. doi: 10.1103/PhysRevA.90.013404. URL https://link.aps.org/doi/10.1103/PhysRevA.90.013404.

Gregory Riviello, Katharine Moore Tibbetts, Constantin Brif, Ruixing Long, Re-Bing Wu, Tak-San Ho, and Herschel Rabitz. Searching for quantum optimal controls under severe constraints. *Phys. Rev. A*, 91:043401, Apr 2015. doi: 10.1103/PhysRevA.91.043401. URL https://link.aps.org/doi/10.1103/PhysRevA.91.043401.

Gregory Riviello, Re-Bing Wu, Qiuyang Sun, and Herschel Rabitz. Searching for an optimal control in the presence of saddles on the quantum-mechanical observable landscape. *Phys. Rev. A*, 95: 063418, Jun 2017. doi: 10.1103/PhysRevA.95.063418. URL https://link.aps.org/doi/10.1103/PhysRevA.95.063418.

Benjamin Russell, Herschel Rabitz, and Re-Bing Wu. Control landscapes are almost always trap free: a geometric assessment. *Journal of Physics A: Mathematical and Theoretical*, 50(20): 205302, apr 2017. doi: 10.1088/1751-8121/aa6b77. URL https://doi.org/10.1088%2F1751-8121%2Faa6b77.

Sonia G. Schirmer, Ivan C.H. Pullen, and Allan I. Solomon. Controllability of quantum systems. *IFAC Proceedings Volumes*, 36(2):281 – 286, 2003. ISSN 1474-6670. doi: https://doi.org/10.1016/S1474-6670(17)38905-X. URL http://www.sciencedirect.com/science/article/pii/S147466701738905X. 2nd IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control 2003, Seville, Spain, 3-5 April 2003.

T. Schulte-Herbrüggen, A. Spörl, N. Khaneja, and S. J. Glaser. Optimal control-based efficient synthesis of building blocks of quantum algorithms: A perspective from network complexity towards time complexity. *Phys. Rev. A*, 72:042331, Oct 2005. doi: 10.1103/PhysRevA.72.042331. URL https://link.aps.org/doi/10.1103/PhysRevA.72.042331.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, May 2020. ISSN 2521-327X. doi: 10.22331/q-2020-05-25-269. URL http://dx.doi.org/10.22331/q-2020-05-25-269.

Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. Exploring entanglement and optimization within the hamiltonian variational ansatz. *CoRR*, abs/2008.02941, 2020. URL https://arxiv.org/abs/2008.02941.

Re-Bing Wu, Michael A. Hsieh, and Herschel Rabitz. Role of controllability in optimizing quantum dynamics. *Phys. Rev. A*, 83:062306, Jun 2011. doi: 10.1103/PhysRevA.83.062306. URL https://link.aps.org/doi/10.1103/PhysRevA.83.062306.

Re-Bing Wu, Ruixing Long, Jason Dominy, Tak-San Ho, and Herschel Rabitz. Singularities of quantum control landscapes. *Phys. Rev. A*, 86:013405, Jul 2012. doi: 10.1103/PhysRevA.86.013405. URL https://link.aps.org/doi/10.1103/PhysRevA.86.013405.

Zhi-Cheng Yang, Armin Rahmani, Alireza Shabani, Hartmut Neven, and Claudio Chamon. Optimizing variational quantum algorithms using pontryagin's minimum principle. *Phys. Rev. X*, 7:021027, May 2017. doi: 10.1103/PhysRevX.7.021027. URL https://link.aps.org/doi/10.1103/PhysRevX.7.021027.