Quantum Natural Gradient

James Stokes
Simons Foundation
Flatiron Institute
New York, NY 10010 USA

Josh Izaac
Xanadu
777 Bay Street, Toronto, Canada

Nathan Killoran
Xanadu
777 Bay Street, Toronto, Canada

Giuseppe Carleo
Simons Foundation
Flatiron Institute
New York, NY 10010 USA

Abstract

A quantum generalization of Natural Gradient Descent is presented as part of a general-purpose optimization framework for variational quantum circuits. The optimization dynamics is interpreted as moving in the steepest descent direction with respect to the Quantum Information Geometry, corresponding to the real part of the Quantum Geometric Tensor (QGT), also known as the Fubini-Study metric tensor. An efficient algorithm is presented for computing a block-diagonal approximation to the Fubini-Study metric tensor for parametrized quantum circuits, which may be of independent interest.

1 Introduction

Variational optimization of parametrized quantum circuits is an integral component for many hybrid quantum-classical algorithms which are arguably the most promising applications of Noisy Intermediate-Scale Quantum (NISQ) computers [21]. Applications include the Variational Quantum Eigensolver (VQE) [20], Quantum Approximate Optimization Algorithm (QAOA) [5] and Quantum Neural Networks (QNNs) [6, 10, 23].

All the above are examples of stochastic optimization problems whereby one minimizes the expected value of a random cost function over a set of variational parameters, using noisy estimates of the cost and/or its gradient. In the quantum setting these estimates are obtained by repeated measurements of some Hermitian observables for a quantum state which depends on the variational parameters.

A variety of optimization methods have been proposed in the variational quantum circuit literature for determining optimal variational parameters including derivative-free (zeroth-order) methods such as Nelder-Mead, finite-differencing [8] or SPSA [25]. Recently the possibility of exploiting direct access to first-order gradient information has been explored. Indeed quantum circuits have been designed to estimate such gradients with minimal overhead compared to objective function evaluations [22].

One motivation for exploiting first-order gradients is theoretical: in the convex case, the expected error in the objective function using the best known zeroth-order stochastic optimization algorithm scales polynomially with the dimension d of the parameter space, whereas Stochastic Gradient Descent (SGD) converges independently of d. Another motivation stems from the empirical success of stochastic gradient methods in training deep neural networks, which involve minimizing non-convex objective functions over high-dimensional parameter spaces.

The application of SGD to deep learning suffers from the caveat that successful optimization hinges on careful hyper-parameter tuning of the learning rate (step size) and other hyper-parameters such

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

as Momentum. Indeed a vast literature has developed devoted to step size selection (see e.g. [11]). The difficulty of choosing a step size can be understood intuitively in the simple quadratic bowl approximation, where the optimal step size depends on the maximum eigenvalue of the Hessian, a quantity which is difficult to calculate in high dimensions. In practical applications the step size selection problem is overcome by using adaptive methods of stochastic optimization such as Adam [14] which have enjoyed wide adoption because of their ability to dynamically select a step size by maintaining a history of past gradients.

Independently of the improvements arising from historical averaging as in Momentum and Adam, it is natural to ask if the geometry of quantum states favors a particular optimization strategy. Indeed, it is well-known that the choice of optimization is intimately linked to the choice of geometry on the parameter space [19]. In the most well-known case of vanilla gradient descent, the relevant geometry corresponds to the l_2 geometry as can be seen by rewriting the iterative update rule as

$$\theta_{t+1} := \theta_t - \eta \nabla L(\theta_t) = \arg\min_{\theta \in \mathbb{R}^d} \left[[\langle \theta - \theta_t, \nabla L(\theta_t) \rangle + \frac{1}{2\eta} \|\theta - \theta_t\|_2^2 \right] \quad , \tag{1}$$

where L is the loss as a function of the variational parameters $\theta \in \mathbb{R}^d$ and η is the step size. Thus, vanilla gradient descent moves in the steepest descent direction with respect to the l_2 geometry.

In the deep learning literature, it has been argued that the l_2 geometry is poorly adapted to the space of weights of deep networks, due to their intrinsic parameter redundancy [19]. The Natural Gradient [1], in contrast, moves in the steepest descent direction with respect to the Information Geometry. This natural gradient descent is advantageous compared to the vanilla gradient because it is invariant under arbitrary re-parametrizations [1] and moreover possesses an approximate invariance with respect to over-parametrizations [17], which are typical for deep neural networks.

In a similar spirit, the quantum circuit literature has investigated the impact of geometry on dynamics of variational algorithms. In particular, it was shown that under the assumption of strong convexity, the l_2 geometry is sub-optimal in some situations compared to the l_1 geometry [9]. The intuitive argument put forth favoring the l_1 geometry is that some quantum state ansatze can be physically interpreted as a sequence of pulses of Hamiltonian evolution, starting from a fixed reference state. In this particular parametrization, each variational parameter can be interpreted as the duration of the corresponding pulse. This is not the only useful parametrization of quantum states, however, and it is thus desirable to find a descent direction which is not tied to any particular parametrization.

Ref. [9] leaves open the problem of finding the relevant geometry for general-purpose variational quantum algorithms and this paper seeks to fill that void. The contributions of this papers are as follows:

- We point out that the demand of invariance with respect to arbitrary reparametrizations can be naturally fulfilled by introducing a Riemannian metric tensor on the space of quantum states, and that the implied descent direction is invariant with respect to reparametrizations by construction.
- We note that the space of quantum states is naturally equipped with a Riemannian metric, which differs from l_2 and l_1 geometries explored previously. In fact, in the absence of noise, the space of quantum states is a complex projective space, which possesses a unique unitarily-invariant metric tensor called the Fubini-Study metric tensor. When restricted to the submanifold of quantum states defining the parametric family, the Fubini-Study metric tensor emerges as the real part of a more general geometric quantity called the Quantum Geometric Tensor (QGT).
- We show that the resulting gradient descent algorithm is a direct quantum analogue of the Natural Gradient in statistics literature, and reduces to it in a certain limit.
- We present quantum circuit construction which computes a block-diagonal approximation to the Quantum Geometric Tensor and show that a simple diagonal preconditioning scheme outperforms vanilla gradient descent in terms of number of iterates required to achieve convergence

2 Theory

Consider the set of probability distributions on N elements; that is, the set of positive vectors $p \in \mathbb{R}^N$, $p \succeq 0$ which are normalized in the 1-norm $\|p\|_1 = 1$. The function $d(p, q) = \arccos(\langle \sqrt{p}, \sqrt{q} \rangle)$ is easily shown to be metric (Fisher-Rao metric) on the probability simplex Δ^{N-1}, where $\sqrt{p}$ and $\sqrt{q}$ denote the elementwise square root of the probability vectors in the probability simplex $p, q \in \Delta^{N-1}$.

Now consider a parametric family of strictly positive probability distributions $p_\theta \succ 0$ indexed by real parameters $\theta \in \mathbb{R}^d$. It can be shown that the infinitesimal squared line element between two members of the parametric family is given by $d^2(p_\theta, p_{\theta+\mathrm{d}\theta}) = \frac{1}{4}\sum_{(i,j)\in[d]^2} I_{ij}(\theta)\mathrm{d}\theta^i\mathrm{d}\theta^j$, where $I_{ij}(\theta)$ are the components of a Riemannian metric tensor (with possible degeneracies) called the Fisher Information Matrix. Letting $p_\theta(x)$ denote the component of the probability vector p_θ corresponding to $x \in [N]$ we have,

$$I_{ij}(\theta) = \sum_{x\in[N]} p_\theta(x) \frac{\partial \log p_\theta(x)}{\partial \theta^i} \frac{\partial \log p_\theta(x)}{\partial \theta^j} \quad . \tag{2}$$

Now consider a N-dimensional complex Hilbert space $\mathbb{C}^N$. Given a vector $\psi \in \mathbb{C}^N$ which is normalized in the 2-norm $\|\psi\|_2 = 1$, a pure quantum state is defined as the projection $P_\psi = |\psi\rangle\langle\psi|$ onto the one-dimensional subspace spanned by the unit vector ψ. In direct analogy with the simplex, the function $d(P_\psi, P_\phi) = \arccos(|\langle\psi, \phi\rangle|)$ is easily shown to be a metric (Fubini-Study metric) on the space of pure states where $\psi, \phi \in \mathbb{C}^N$ are unit vectors. Letting ψ_θ denote a parametric family of unit vectors, the infinitesimal squared line element between two states defined by the parametric family is given by $d^2(P_{\psi_\theta}, P_{\psi_{\theta+\mathrm{d}\theta}}) = \sum_{(i,j)\in[d]^2} g_{ij}(\theta)\,\mathrm{d}\theta^i\mathrm{d}\theta^j$, where $g_{ij}(\theta) = \mathrm{Re}[G_{ij}(\theta)]$ is the Fubini-Study metric tensor, which can be expressed in terms of the following Quantum Geometric Tensor (see [26] for a review),

$$G_{ij}(\theta) = \langle\partial_i\psi_\theta, \partial_j\psi_\theta\rangle - \langle\partial_i\psi_\theta, \psi_\theta\rangle\,\langle\psi_\theta, \partial_j\psi_\theta\rangle \quad . \tag{3}$$

Consider a parametric family of unitary operators $U(\theta) \in U(N)$ which are indexed by real parameters $\theta \in \mathbb{R}^d$. Given a fixed reference unit vector $|0\rangle \in \mathbb{C}^N$ and a Hermitian operator $H = H^\dagger$ acting on $\mathbb{C}^N$, we consider the optimization problem $\min_{\theta\in\mathbb{R}^d} L(\theta)$, where $L(\theta) = \frac{1}{2}\langle\psi_\theta, H\psi_\theta\rangle$ and $\psi_\theta = U_\theta|0\rangle$. Global optimization of the nonconvex objective function $L(\theta)$ is impractical, so we instead propose to search for local optima by iterating the discrete-time dynamical system,

$$\theta_{t+1} = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \left[\langle\theta - \theta_t, \nabla L(\theta_t)\rangle + \frac{1}{2\eta}\|\theta - \theta_t\|^2_{g(\theta_t)}\right] \quad , \tag{4}$$

where $g(\theta_t)$ is the symmetric matrix with (i, j) component $\mathrm{Re}[G_{ij}(\theta_t)]$. The minimizer is given in terms of the pseudo-inverse $g^+(\theta_t)$ as follows, $\theta_{t+1} = \theta_t - \eta\, g^+(\theta_t)\nabla L(\theta_t)$. In the continuous-time limit corresponding to vanishing step size $\eta \to 0$, the dynamics (2) is equivalent to imaginary-time evolution within the variational subspace according to the Hamiltonian H.

In a digital quantum computer the Hilbert space dimension $N = 2^n$ is exponential in the number of qubits $n \in \mathbb{N}$ and the Hilbert space has a natural tensor product decomposition into two-dimensional factors $\mathbb{C}^N = \mathbb{C}^{2^n} = (\mathbb{C}^2)^{\otimes n}$. A parametric family of unitaries relevant to variational quantum algorithms consists of decompositions into a products of $L \geq 1$ non-commuting layers of unitaries. Specifically, assume that the variational parameter vector is of the form $\theta = \boldsymbol{\theta}_1 \oplus \cdots \oplus \boldsymbol{\theta}_L \in \mathbb{R}^d$ where $\oplus$ denotes the direct sum (concatenation) and consider a unitary operator acting on n qubits of the form $U_L(\theta) := V_L(\boldsymbol{\theta}_L)W_L \cdots V_1(\boldsymbol{\theta}_1)W_1$, where $V_l(\boldsymbol{\theta}_l)$ and W_l are parametric and non-parametric unitary operators, respectively. For later convenience, we introduce the notation $U_{[l_1:l_2]} := V_{l_2}W_{l_2} \cdots V_{l_1}W_{l_1}$ for representing subcircuits between layers $l_1 \leq l_2$ Moreover, we define the convenience state $\psi_l := U_{[1:l]}|0\rangle$ for each layer $l \in [L]$,

Computing the QGT corresponding to a parametrized quantum circuit is a challenging task. We will show, nevertheless, that block-diagonal components of the tensor can be efficiently computed on a quantum computer. Consider the lth layer of the circuit parametrized by $\boldsymbol{\theta}_l$ and let ∂_i and ∂_j denote the partial derivative operators acting with respect to any pair of components of $\boldsymbol{\theta}_l$ (not necessarily distinct). For each layer $l \in [L]$ there exist Hermitian generator matrices K_i and K_j such that, $\partial_i V_l(\boldsymbol{\theta}_l) = -\mathrm{i}K_i V_l(\boldsymbol{\theta}_l)$ and $\partial_j V_l(\boldsymbol{\theta}_l) = -\mathrm{i}K_j V_l(\boldsymbol{\theta}_l)$ where for simplicity we have dropped the layer index l from the Hermitian generator K_j. The commutativity of the partial derivative operators combined with unitarity of $V_l(\boldsymbol{\theta}_l)$ implies that $[K_i, K_j] = 0$. It follows from unitarity of the subcircuit $U_{(l:L]}$ and Hermiticity of the generator K_i that $\langle\partial_i\psi_\theta|\partial_j\psi_\theta\rangle = \langle\psi_l|K_iK_j|\psi_l\rangle$. Similarly, the so-called Berry connection is given by $\mathrm{i}\langle\psi_\theta|\partial_j\psi_\theta\rangle = \langle\psi_l|K_j|\psi_l\rangle$. Combining these expressions we obtain the following form for the lth block of the QGT, $G^{(l)}_{ij} = \langle\psi_l|K_iK_j|\psi_l\rangle - \langle\psi_l|K_i|\psi_l\rangle\langle\psi_l|K_j|\psi_l\rangle$ In fact, since K_iK_j is Hermitian, the block-diagonal of the Fubini-Study metric tensor coincides with the block-diagonal of the QGT. The preceding calculation demonstrates the following key facts:

1. The lth block of the Fubini-Study metric tensor can be evaluated in terms of quantum expectation values of Hermitian observables.

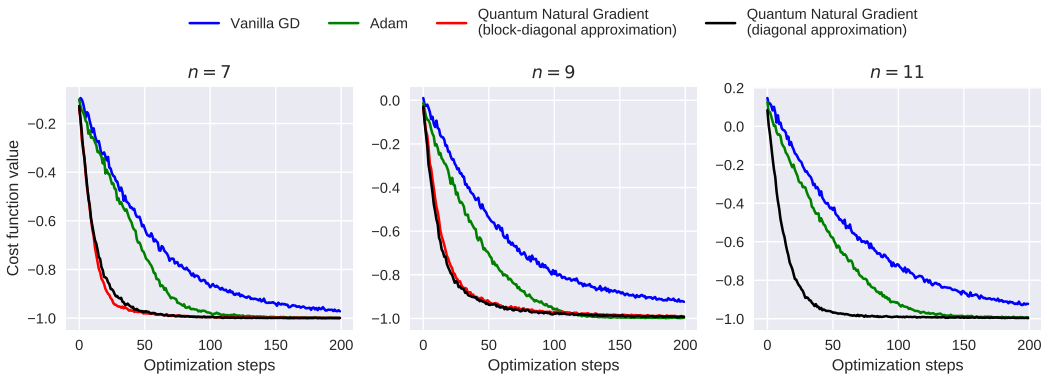

Figure 1: The cost function value for $n = 7, 9, 11$ qubits and $l = 5$ layers as function of training iteration for four different optimization dynamics. 8192 shots (samples) are used per required expectation value during optimization.

2. The states ψ_l defining the quantum expectation values are prepared by subcircuits of the full quantum circuit and are thus experimentally realizable.

3 Numerical Experiments

In order to validate the choice of geometry, numerical experiments were conducted to compare the Quantum Natural Gradient approach against vanilla gradient descent and the Adam optimizer. These experiments were performed with the open-source quantum machine learning software library PennyLane [2, 22]. New functionality was added for efficiently computing the block-diagonal $g_{ij}^{(l)}$ and diagonal g_{ii} approximations of the Fubini-Study metric tensor for arbitrary n-qubit parametrized quantum circuits on quantum hardware.

For numerical verification, we considered the circuit of [20], which consists of an initial fixed layer of $R_y(\pi/4)$ gates acting on n qubits, followed by L layers of parametrized Pauli rotations interwoven with 1D ladders of controlled-Z gates, and target Hermitian observable chosen to be the same two-Pauli operator $Z_1 Z_2$ acting on the first and second qubit which has a ground state energy of -1. Starting from the same random initialization of Ref. [20], we optimize the parametrized Pauli rotation gates using vanilla gradient descent, the Adam optimzer, and the Quantum Natural Gradient optimizer, with both the block-diagonal and diagonal approximations. The results are shown in Fig. 1 for $n = 7, 9, 11$ qubits, $L = 5$ layers, and with the optimization performed using 8192 samples per expectation value. In all cases the vanilla gradient descent fails to find the minimum of the objective function, while the Quantum Natural Gradient descent finds the minimum in a small number of iterations, in both block-diagonal and strictly diagonal approximation. In addition, we present a comparison with the Adam optimizer which is a non-local averaging method. In this particular circuit, Adam is capable of finding the minimum but requires a larger number of iterations than the Quantum Natural Gradient. Furthermore, the improvement afforded by the Quantum Natural Gradient optimizer appears more significant with increasing qubit number. Note that for $n = 11$, we do not include the block-diagonal approximation, due to the increased classical overhead associated with numerically computing the shared eigenbasis for each parametrized layer. However, this overhead can likely be negated by implementing the techniques of [4] and [7]. To investigate the effects of variable circuit depth, the numerical experiment was repeated with $n = 9$ qubits, and parametric quantum circuits with $L = 3, 4, 5, 6$ layers. The results are shown in Fig. 2, highlighting that the Quantum Natural Gradient optimizer retains its advantage with increasing circuit depth.

4 Relationship with existing work

It is easy to see that the Quantum Natural Gradient subsumes the Natural Gradient as a special case. Indeed if $\{|x\rangle : x \in [N]\}$ denotes an orthonormal basis for $\mathbb{C}^N$ then one can easily verify that for the family of unit vectors defined by $\psi_\theta = \sum_{x\in[N]} \sqrt{p_\theta(x)}\,|x\rangle$ we have $G_{ij}(\theta) = \frac{1}{4} I_{ij}(\theta)$. In contrast

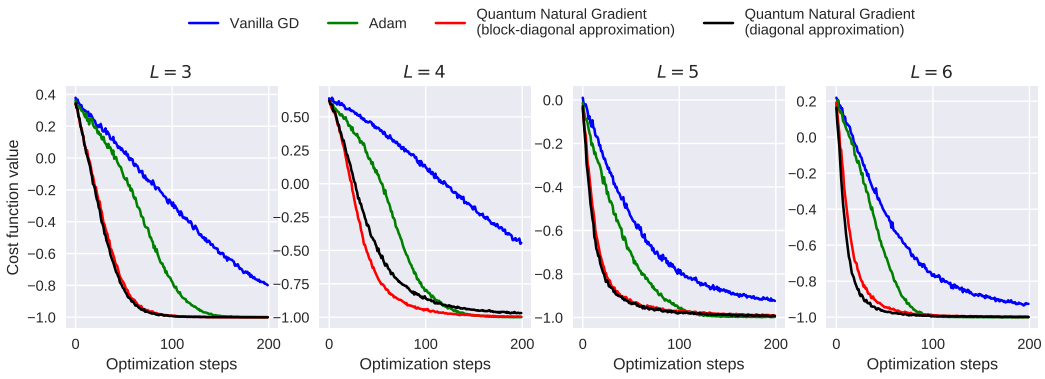

Figure 2: The cost function value for $n = 9$ qubits and $l = 3, 4, 5, 6$ layers as function of training iteration for four different optimization dynamics. 8192 shots (samples) are used per required expectation value during optimization.

to classical statistical learning, however, there is no direct relationship between the quantum Fisher Information and the curvature of the objective. In the Variational Quantum Monte Carlo literature, the Stochastic Reconfiguration algorithm [24] has been developed which produces a stochastic estimate of (2) by classical sampling from the Born probability distribution corresponding to ψ_θ. An associated real-time evolution algorithm, which exploits the the imaginary part $\mathrm{Im}[G_{ij}(\theta)]$ of the Quantum Geometric Tensor (3) has been developed in [16] and subsequently demonstrated on quantum hardware in [3]. For details on the geometry of the time-dependent variational principle we refer the reader to [15, Proposition 2.4]. Variational imaginary-time evolution on hybrid quantum-classical devices has been previously investigated in [18, 12, 13]. In these works, the choice of optimization geometry can be shown to correspond to the unit sphere $\mathbb{S}^{N-1} = \{\psi \in \mathbb{C}^N : \|\psi\|_2 = 1\}$, rather than the complex projective space $\mathbb{CP}^{N-1}$ utilized in this paper. Recently, Ref. [27] appeared which considers general evolution of variational density matrices in both real and imaginary time, from a different perspective. By restricting their proposal to pure state projectors (elements of $\mathbb{CP}^{N-1}$) they find an algorithm equivalent to ours.

References

[1] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, and Nathan Killoran. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

[3] Ming-Cheng Chen, Ming Gong, Xiao-Si Xu, Xiao Yuan, Jian-Wen Wang, Can Wang, Chong Ying, Jin Lin, Yu Xu, Yulin Wu, et al. Demonstration of adiabatic variational quantum computing with a superconducting quantum coprocessor. *arXiv preprint arXiv:1905.03150*, 2019.

[4] Ophelia Crawford, Barnaby van Straaten, Daochen Wang, Thomas Parks, Earl Campbell, and Stephen Brierley. Efficient quantum measurement of pauli operators. *arXiv preprint arXiv:1908.06942*, 2019.

[5] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

[6] Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018.

[7] Pranav Gokhale, Olivia Angiuli, Yongshan Ding, Kaiwen Gui, Teague Tomesh, Martin Suchara, Margaret Martonosi, and Frederic T Chong. Minimizing state preparations in variational quantum eigensolver by partitioning into commuting families. *arXiv preprint arXiv:1907.13623*, 2019.

[8] Gian Giacomo Guerreschi and Mikhail Smelyanskiy. Practical optimization for hybrid quantum-classical algorithms. *arXiv preprint arXiv:1701.01450*, 2017.

[9] Aram Harrow and John Napp. Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms. *arXiv preprint arXiv:1901.05374*, 2019.

[10] William James Huggins, Piyush Patil, Bradley Mitchell, K Birgitta Whaley, and Miles Stoudenmire. Towards quantum machine learning with tensor networks. *Quantum Science and Technology*, 4:024001, 2018.

[11] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[12] Tyson Jones and Simon C Benjamin. Quantum compilation and circuit optimisation via energy dissipation. *arXiv preprint arXiv:1811.03147*, 2018.

[13] Tyson Jones, Suguru Endo, Sam McArdle, Xiao Yuan, and Simon C Benjamin. Variational quantum algorithms for discovering hamiltonian spectra. *Physical Review A*, 99(6):062304, 2019.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] PH Kramer and Marcos Saraceno. *Geometry of the time-dependent variational principle in quantum mechanics*. Springer, 1981.

[16] Ying Li and Simon C Benjamin. Efficient variational quantum simulator incorporating active error minimization. *Physical Review X*, 7(2):021050, 2017.

[17] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

[18] Sam McArdle, Tyson Jones, Suguru Endo, Ying Li, Simon C Benjamin, and Xiao Yuan. Variational ansatz-based quantum simulation of imaginary time evolution. *npj Quantum Information*, 5(1):1–6, 2019.

[19] Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

[20] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5:4213, 2014.

[21] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018.

[22] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

[23] Maria Schuld, Alex Bocharov, Krysta Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *arXiv preprint arXiv:1804.00633*, 2018.

[24] Sandro Sorella, Michele Casula, and Dario Rocca. Weak binding between two aromatic rings: Feeling the van der waals attraction by quantum monte carlo methods. *The Journal of Chemical Physics*, 127(1):014105, 2007.

[25] James C Spall et al. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[26] F Wilczek and A Shapere. Geometric phases in physics. *Geometric Phases In Physics. Series: Advanced Series in Mathematical Physics, ISBN: 978-9971-5-0621-6. WORLD SCIENTIFIC, Edited by F Wilczek and A Shapere, vol. 5*, 5, 1989.

[27] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C Benjamin. Theory of variational quantum simulation. *Quantum*, 3:191, 2019.