QTML 2021

Quantum Techniques in Machine Learning

Booklet of Extended Abstracts

Monday November 8 2021

Quantum Algorithms for Reinforcement Learning with a Generative Model*

Daochen Wang† Aarthi Sundaram‡ Robin Kothari‡
Ashish Kapoor‡ Martin Roetteler‡

Abstract

Reinforcement learning studies how an agent should interact with an environment to maximize its cumulative reward. A standard way to study this question abstractly is to ask how many samples an agent needs from the environment to learn an optimal policy for a γ-discounted Markov decision process (MDP). For such an MDP, we design quantum algorithms that approximate an optimal policy (π^*), the optimal value function (v^*), and the optimal Q-function (q^*).

Problem Setup

We study an *infinite-horizon discounted MDP*, M, with a finite set, $\mathcal{S}$, of *states*, where at each state an agent can choose to take an action from a finite set, $\mathcal{A}$, of *actions*. Upon taking an action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, the agent receives *reward* $r[s,a] \in [0,1]$ and transitions to a state $s' \in \mathcal{S}$ with some probability $p(s'|s,a)$. The last parameter needed to specify M is the *discount factor* $\gamma \in [0,1)$, which discounts the reward the agent receives at later time steps t by a factor of γ^t. Hence M is conveniently summarized by a 5-tuple, $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$. For convenience, we write $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$, the cardinalities of $\mathcal{S}$ and $\mathcal{A}$ respectively, and $\Gamma := (1-\gamma)^{-1}$.

Given such an MDP, the agent's goal is to choose actions to maximize its expected sum of γ-discounted rewards over infinitely many time steps. Following standard practice, we assume the agent has full knowledge of $\mathcal{S}$, $\mathcal{A}$, r, and γ, but not p at the outset. A primary objective is to compute a deterministic *policy* $\pi : \mathcal{S} \to \mathcal{A}$ for the agent that specifies the action $a = \pi(s)$ it should take at $s \in \mathcal{S}$ to best achieve its goal with high probability.

For a given policy $\pi : \mathcal{S} \to \mathcal{A}$, the *value-function* (or simply *value*) of π, $v^\pi : \mathcal{S} \to [0,\Gamma]$, and the *Q-function* of π, $q^\pi : \mathcal{S} \times \mathcal{A} \to [0,\Gamma]$, are defined by

$$\begin{aligned} v^\pi[s] &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r[s_t,a_t] \;\middle|\; s_0 = s, \;\; \forall i \geq 0 \,:\, a_i = \pi[s_i]\right], \text{ and} \\ q^\pi[s,a] &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r[s_t,a_t] \;\middle|\; s_0 = s, \; a_0 = a, \;\; \forall i \geq 1 \,:\, a_i = \pi[s_i]\right], \end{aligned} \tag{1}$$

where the expectations are over the probabilistic state transitions, i.e., for all $i \geq 0$, s_{i+1} is sampled from the distribution $p(\cdot|s_i,a_i)$. It is known that any such MDP admits an optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$, in the strong sense that $v^{\pi^*}[s] \geq v^\pi[s]$ and $q^{\pi^*}[s,a] \geq q^\pi[s,a]$ for all $\pi \in \Pi$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$, where Π is the space of all policies. It is common to write $v^* := v^{\pi^*}$ and $q^* := q^{\pi^*}$.

*Full paper presented at ICML 2021 and is available at http://proceedings.mlr.press/v139/wang21w.html
†University of Maryland
‡Microsoft Quantum and Microsoft Research

Goal: Output an ϵ-accurate estimate of	Classical sample complexity	Quantum sample complexity	
	Upper and lower bound	Upper bound	Lower bound
q^*	$\frac{SA\Gamma^3}{\epsilon^2}$	$\frac{SA\Gamma^{1.5}}{\epsilon}$	$\frac{SA\Gamma^{1.5}}{\epsilon}$
v^*, π^*	$\frac{SA\Gamma^3}{\epsilon^2}$	$\min\{\frac{SA\Gamma^{1.5}}{\epsilon}, \frac{S\sqrt{A}\Gamma^3}{\epsilon}\}$	$\frac{S\sqrt{A}\Gamma^{1.5}}{\epsilon}$

Table 1: Quantum computing allows for speedups in terms of the parameters ϵ, $\Gamma := (1-\gamma)^{-1}$, and A, but not S. All bounds are for maximum failure probability δ constant. There are constraints on the ranges of ϵ for which these results are valid, for details, please refer to our full paper.

The goal of our work is to design algorithms that compute q^*, v^*, and π^* using as few resources as possible. By resource, we refer to the sample complexity, that is, the number of calls to the unitary matrix $\mathcal{G} : \mathbb{C}^S \otimes \mathbb{C}^A \otimes \mathbb{C}^S \otimes \mathbb{C}^J \to \mathbb{C}^S \otimes \mathbb{C}^A \otimes \mathbb{C}^S \otimes \mathbb{C}^J$ with

$$\mathcal{G} : |s\rangle \otimes |a\rangle \otimes |0\rangle \otimes |0\rangle \mapsto |s\rangle \otimes |a\rangle \otimes \Big(\sum_{s' \in \mathcal{S}} \sqrt{p(s'|s,a)}\ |s'\rangle \otimes |v_{s'}\rangle \Big), \tag{2}$$

where $0 \leq J \in \mathbb{Z}$ is arbitrary and $|v_{s'}\rangle \in \mathbb{C}^J$ are arbitrary. $\mathcal{G}$ can be instantiated whenever we have a (classical) computer simulator of the environment, for example, in the case of computer games.

Main Results

Table 1 summarizes our main results. The classical sample complexities have only recently been completely characterized for all three quantities [LWC+20] for the full range of $\epsilon \in (0, \Gamma]$. As the table shows, for computing q^*, we construct a quantum algorithm that offers a quadratic speedup in terms of Γ and ϵ. For computing v^* and π^*, we construct a second quantum algorithm that offers an additional quadratic speedup in terms of A at the expense of Γ. Our quantum algorithms can be thought of as quantizations of (a combination of) the classical algorithms of [SWWY18, SWW+18]. We use quantum minimum finding [DH96] and quantum mean estimation [Mon15] to perform the quantization which requires a delicate analysis. Conversely, we also prove quantum lower bounds for computing q^*, v^*, and π^*. Our lower bounds show that our q^* algorithm is optimal, that we have optimal algorithms for v^* and π^* provided one of Γ or A is constant, but that there may still be a faster quantum algorithm for v^* and π^*. Our proof technique in fact allows us to reprove the *classical* lower bounds in a qualitatively stronger way than existing bounds.

We remark that the time complexities of our quantum algorithms are the same as their sample complexities up to log factors assuming that the generative model can be called in constant time and that we have access to quantum random access memory (QRAM) [GLM08]. This is because the classical algorithms of [SWWY18, SWW+18] that we quantize satisfies this property and the quantum subroutines we use to quantize it also satisfy this property.

Quantum reinforcement learning is not new [DCLT08, DTB16, PDM+14, DTB17, JTPN+21]. However, ours is the first work to *rigorously characterize* the quantum complexity of the main tasks in reinforcement learning, that of computing q^*, v^*, and π^*.

References

[DCLT08] D. Dong, C. Chen, H. Li, and T. Tarn. Quantum Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(5):1207–1220, 2008.

[DH96] Christoph Dürr and Peter Høyer. A Quantum Algorithm for Finding the Minimum, 1996. arXiv:quant-ph/9607014

[DTB16] Vedran Dunjko, Jacob M. Taylor, and Hans J. Briegel. Quantum-Enhanced Machine Learning. *Physical Review Letters*, 117(13), 2016.

[DTB17] V. Dunjko, J. M. Taylor, and H. J. Briegel. Advances in quantum reinforcement learning. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 282–287, 2017.

[GLM08] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum Random Access Memory. *Physical Review Letters*, 100:160501, 2008.

[JTPN+21] Sofiene Jerbi, Lea M. Trenkwalder, Hendrik Poulsen Nautrup, Hans J. Briegel, and Vedran Dunjko. Quantum Enhancements for Deep Reinforcement Learning in Large Spaces. *PRX Quantum*, 2:010328, 2021.

[LWC+20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, pages 12861–12872. Curran Associates, Inc., 2020.

[Mon15] Ashley Montanaro. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181):20150301, 2015.

[PDM+14] Giuseppe Davide Paparo, Vedran Dunjko, Adi Makmal, Miguel Angel Martin-Delgado, and Hans J. Briegel. Quantum Speedup for Active Learning Agents. *Physical Review X*, 4(3):031002, 2014.

[SWW+18] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 5186–5196, 2018.

[SWWY18] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance Reduced Value Iteration and Faster Algorithms for Solving Markov Decision Processes. In *Proceedings of the 29th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 770–787, USA, 2018.

Universal Compiling and (No-)Free-Lunch Theorems for Continuous Variable Quantum Learning

Tyler Volkoff, Zoë Holmes, Andrew Sornborger
Los Alamos National Laboratory

Extended abstract

Quantum compiling, where a parameterized quantum circuit $V(\theta)$ is trained to learn a target unitary U, is a fundamental primitive for quantum computing. It can be used to find optimal (i.e. short-depth and noise-resistant) circuits to aid the implementation of larger algorithms. Or as a tool, analogous to a quantum sensing protocol, to learn the dynamics of an unknown experimental system (see Fig. 1a). In this presentation, we will introduce algorithms for continuous-variable (CV) variational quantum compiling which are motivated by extending the "no-free-lunch" theorems of supervised learning theory to the quantum CV setting. These algorithms utilize readily available Gaussian resources, such as coherent states of varying intensities and two-mode squeezed states with varying entanglement. We further prove that our algorithms are trainable, thereby providing a workaround to obstructions to scalability such as the barren plateau phenomenon that plagues the finite dimensional setting [MBS$^+$18].

To illustrate the connection between variational quantum compiling and quantum no-free-lunch theorems, consider the task of learning a unitary U on a d-dimensional Hilbert space. By considering a set S of training states of the form $U \otimes \mathbb{I}_{\mathcal{R}}|\psi^{\mathfrak{r}}_{\text{TMSS}}(r,\phi_j)\rangle$, where $|\psi^{\mathfrak{r}}_{\text{TMSS}}(r,\phi_j)\rangle \propto \sum_{n=0}^{\mathfrak{r}-1}(\tanh r)^n e^{in\phi_j}|n\rangle \otimes |n\rangle$ are truncated two-mode squeezed states with random phase ϕ_j and Schmidt rank $\mathfrak{r} < d$, the following quantum no-free-lunch theorem applies [SCH$^+$20]:

$$E_S(E_U(R_U(V_S))) = 1 - \frac{\mathfrak{r}^2|S|^2 + d + 1}{d(d+1)}\,. \tag{1}$$

Here V_S is the approximation to U output by the learning algorithm when trained on S, and $R_U(V_S)$ is the risk function. We chose the truncated two mode squeezed states so that training on Gaussian states is obtained as $\mathfrak{r} \to \infty$. If $\mathfrak{r} = cd$ for any $c < 1$, then a training set S of cardinality 1 allows to get the expected risk to $1 - c$ in the $d \to \infty$ limit. That is, a single training state with maximum entanglement dimension can be used to perfectly learn an unknown unitary U.

This fact motivates the introduction of the following faithful cost function for CV variational compiling,

$$C_{\text{LE-TMSS}_r}(V(\theta), U) := 1 - |\langle\psi^m_{\text{TMSS}}(r)|UV(\theta)^\dagger \otimes \mathbb{I}_B|\psi^m_{\text{TMSS}}(r)\rangle|^2\,. \tag{2}$$

Here the $2m$-mode two-mode squeezed state on register AB is defined by

$$|\psi^m_{\text{TMSS}}(r)\rangle := \lim_{\mathfrak{r}\to\infty} \bigotimes_{j=1}^{m} |\psi^{\mathfrak{r}}_{\text{TMSS}}(r)\rangle_{A_jB_j}. \tag{3}$$

The cost function (3) can be computed with a simple, depth 6 CV circuit that utilizes only photon number detection (Fig. 1b). We further introduce a local version of (3) for which the gradient goes to zero subexponentially in the number m of modes and therefore does not exhibit a "barren plateau landscape"(3).

This presentation will also detail how an analogous, but fully infinite dimensional, chain of reasoning can produce an alternative algorithm for CV compiling that, in principle, does not require

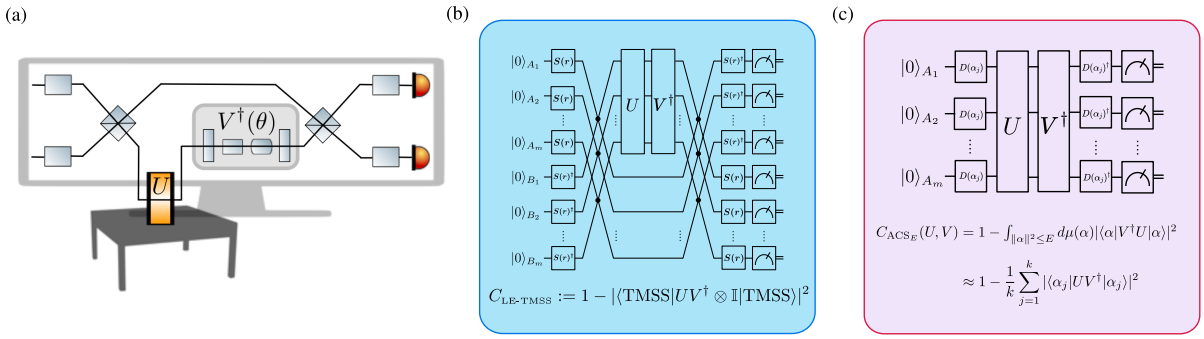

Figure 1: Here we sketch (a) an experimental circuit to learn the unitary U implemented by a novel optical material (shown in orange) by training a parameterised quantum circuit $V^\dagger(\theta)$ implemented on an optical quantum computer (shown in grey). The specific circuits to compute our proposed costs Eq (2) and Eq. (5) are shown in (b) and (c) respectively.

entanglement. Specifically, we develop a CV no-free-lunch theorem for learning m-mode linear optical operations via learning their corresponding orthogonal matrix O. We derive the equality

$$E_S(E_O(R_O(T_S))) = \frac{1}{2} - \frac{|S|}{4m}\,, \tag{4}$$

which gives the expected risk (averaged over all possible orthogonal matrices and training sets consisting of $|S|$ coherent state training pairs) in terms of the number of modes m of the CV system and amount of training data $|S|$.

This result motivates an alternative faithful cost in terms of the fidelity between a coherent state acted on by U and a coherent state acted on by V averaged over $|S|$ coherent states with energy less that E, i.e. the cost

$$C_{\text{ACS}_E}(V,U) = 1 - \frac{1}{|S|}\sum_{j=1}^{|S|} \left|\langle \boldsymbol{\alpha}_j|V^\dagger U|\boldsymbol{\alpha}_j\rangle\right|^2 \tag{5}$$

where $\|\boldsymbol{\alpha}_j\|^2 < E$ for all j. This can be computed using the non-entangling circuit shown in Fig. 1(c). Eq. (4) implies that computing this cost function for $|S| = 2m$ coherent states is sufficient to faithfully compile m-mode linear optical unitaries, which is consistent with the intuition from the fact that an orthogonal phase space transformation has rank $2m$.

Thus our results show how theorems from statistical learning theory can be used to motivate near-term CV quantum compiling algorithms. We illustrate the wide applicability of our cost functions for CV quantum compiling by numerically demonstrating efficient learning of arbitrary single-mode Gaussian unitaries, the generalized beamsplitter operation, and Kerr non-linearities. We expect our algorithms to find applications in a broad range of areas including the characterization of nonlinear optical media, entanglement spectroscopy, and optimal CV circuit design.

References

[MBS+18] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nat. Comm.*, 9(1):4812, 2018.

[SCH+20] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles. Reformulation of the no-free-lunch theorem for entangled data sets. *arXiv preprint arXiv:2007.04900*, 2020.

Enhancing Combinatorial Optimization with Quantum Generative Models

Javier Alcazar[1] and Alejandro Perdomo-Ortiz[1,*]

[1]*Zapata Computing Canada Inc., 325 Front St W, Toronto, ON, M5V 2Y1*

Abstract: Combinatorial optimization is one of the key candidates in the race for practical quantum advantage. In this work, we introduce a new family of quantum-enhanced optimizers and demonstrate how quantum machine learning models known as quantum generative models can find lower minima than those found by means of state-of-the-art classical solvers. **Preprint:** arXiv:2101.06250

We present two new quantum-enhanced optimization strategies. The first scheme leverages data points evaluated during the optimization search from any quantum or classical optimizer. In this scheme, we show how our quantum generative model boosts the performance of classical solvers in hard-to-solve instances where the classical solver is not capable of making progress as a stand-alone strategy. The second quantum optimization strategy works as a stand-alone solver. Here we show its superior performance when the goal is to find the best minimum within the least number of cost function evaluations. Under this setting, we benchmark our quantum-enhanced optimization strategies against several solvers, including Bayesian optimizers which are known to be one of the best competing solvers in such tasks. To illustrate our findings, these benchmarks are performed in the context of the portfolio optimization problem by constructing instances from the S&P 500 stock market index. We show that our quantum-inspired generative models based on tensor networks generalize to unseen candidates with lower cost function values than any of the candidates seen by the classical solvers. This is the first demonstration of the generalization capabilities of quantum generative models that brings real value in the context of an industrial-scale application.

Along with machine learning and the simulation of materials, combinatorial optimization is one of top candidates for practical quantum advantage. That is, the moment where a quantum-assisted algorithm outperforms the best classical algorithms in the context of a real-world application with a commercial or scientific value. There is an ongoing portfolio of techniques to tackle optimization problems with quantum subroutines, ranging from algorithms tailored for quantum annealers (e.g., Refs. [1, 2]), gate-based quantum computers (e.g., Refs. [3, 4]) and quantum-inspired (QI) models based on tensor networks (e.g., Ref. [5]).

Regardless of the quantum optimization approach proposed to date, there is a need to translate the real-world problem into a polynomial unconstrained binary optimization (PUBO) expression – a task which is not necessarily straightforward and that usually results in an overhead in terms of the number of variables. Specific real-world use cases illustrating these PUBO mappings are depicted in Refs. [6] and [7]. Therefore, to achieve practical quantum advantage in the near-term, it would be ideal to find a quantum optimization strategy that can work on arbitrary objective functions, bypassing the translation and overhead limitations raised here. In our work, we offer a solution to these challenges by proposing a family of quantum enhanced optimizers (QEOs) which can scale to large problems where combinatorial problems become intractable in real-world settings. Since our solver does not rely on the details of the objective function to be minimized it is categorized in the family of the so-called *black-box solvers*. Another highlight of our approach is that it can utilize available observations obtained from attempts to solve the optimization problem. These initial evaluations can come from any source, from random search trials to tailored state-of-the-art classical (or quantum) optimizers for the specific problem at hand.

Our QEO strategy is based on two key ideas. First, our model relies on a probabilistic component which aims to capture the correlations in previously observed data (step 0-3 in Fig. 1). In the proposal presented here, our QEOs leverage the probabilistic modeling framework of generative models. Second, the (quantum) generative models need to be capable of generating new "unseen" solution candidates which have the potential to have a lower value for the objective function than those already "seen" and used as the training set (step 4-6 in Fig. 1). This is the fundamental concept of generalization: the most desirable and important feature of any practical ML model. Finally, the new set is merged with the seed data set (step 7 in Fig. 1) to form an updated seed data set (step 8 in Fig. 1) which is to be used in the next iteration of the algorithm.

Our work elaborate on these components and demonstrate these two properties in the context of the tensor-network-based generative models and its application to a non-deterministic polynomial-time hard (NP-hard) version of the portfolio optimization in finance, in particular, adding a cardinality constraint in the number of assets in the portfolio. The selection of optimal investment on a specific set of assets, or *portfolios*, is a problem of great interest in the area of quantitative finance. The goal of this optimization task, introduced by Markowitz [8], is to generate a set of portfolios that offers either the highest expected return (profit) for a defined level of risk or the lowest risk for a given level of expected return. In this work, we focus in the combinatorial optimization problem of choosing portfolios which minimizes its volatility or risk given a specific target return. A brief summary results for both strategies are presented in figure 2. The *relative quantum enhancement*, η, in figure 2 (a) is computed as $\eta = \frac{C^{\rm cl}_{\rm min} - C^{\rm QEO}_{\rm min}}{C^{\rm cl}_{\rm min}} \times 100\%$., where $C^{\rm cl}_{\rm min}$ is the lowest minimum value found by modes 1 or 2, while $C^{\rm QEO}_{\rm min}$ corresponds to the lowest value found with the quantum-enhaced approach.

* alejandro@zapatacomputing.com

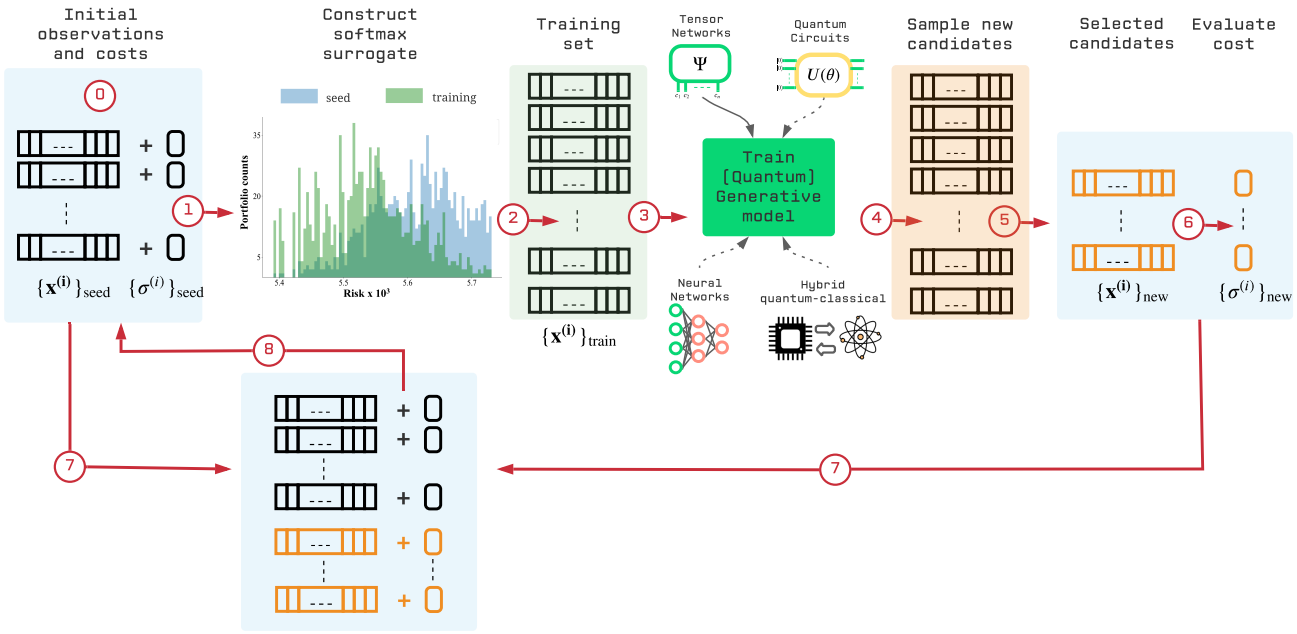

FIG. 1. **Scheme for our Quantum-Enhanced Optimizer (QEO).** The QEO framework leverages generative models to utilize previous samples coming from any classical or quantum solver to propose candidate solutions which might be out of reach for conventional solvers.

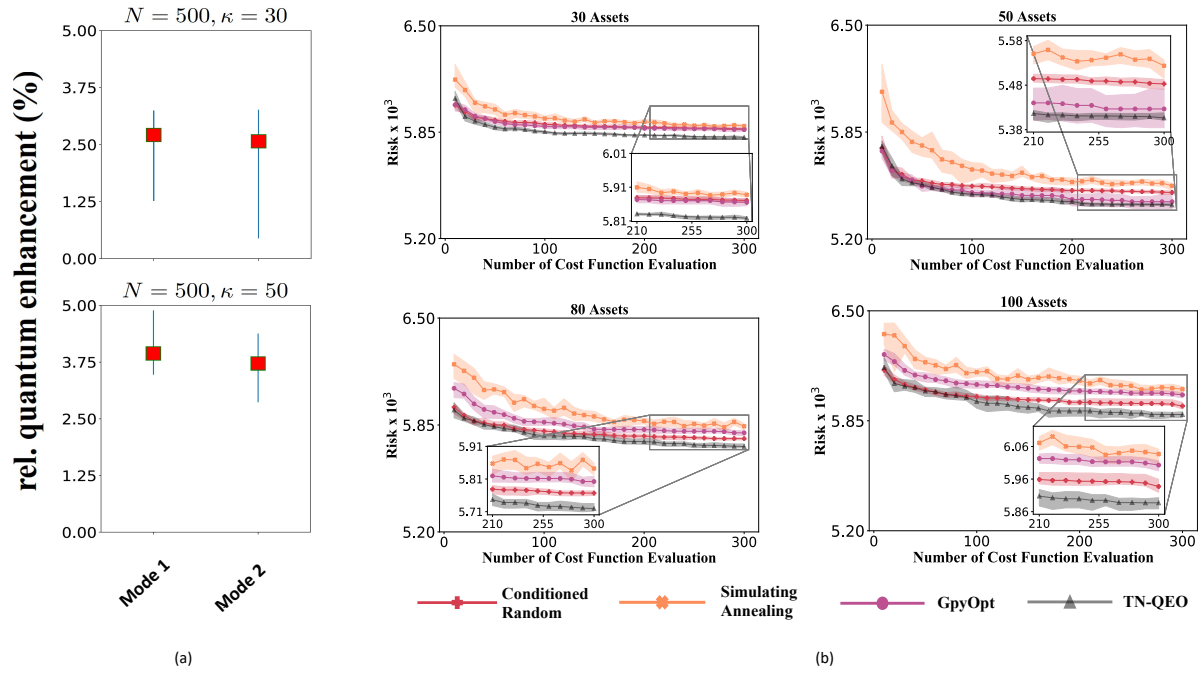

FIG. 2. (a) TN-QEO as a *booster*. Results showing the relative quantum enhancement from TN-QEO over simulated annealing (SA) under two different models corresponding to different annealing strategies. The comparison is such that both modes of SA and QEO have equal computational time. Positive values of η indicate runs where TN-QEO outperformed the respective classical strategies. (b) TN-QEO as a *stand-alone* solver. This is a comparison of TN-QEO against three classical competing algorithms.

[1] Tadashi Kadowaki and Hidetoshi Nishimori, "Quantum annealing in the transverse ising model," Phys. Rev. E. **58**, 5355 (1998).
[2] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, Joshua Lapan, Andrew Lundgren, and Daniel Preda, "A quantum adiabatic evolution algorithm applied to random instances of an NP-Complete problem," Science **292**, 472–475 (2001).
[3] Sam Gutmann Edward Farhi, Jeffrey Goldstone, "A quantum approximate optimization algorithm," arXiv:1411.4028 (2014).
[4] Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G Rieffel, Davide Venturelli, and Rupak Biswas, "From the quantum approximate optimization algorithm to a quantum alternating operator ansatz," Algorithms **12**, 34 (2019).
[5] Samuel Mugel, Carlos Kuchkovsky, Escolastico Sanchez, Samuel Fernandez-Lorenzo, Jorge Luis-Hita, Enrique Lizaso, and Roman Orus, "Dynamic portfolio optimization with real datasets using quantum processors and quantum-inspired tensor networks," (2020), arXiv:2007.00017 [quant-ph].
[6] A. Perdomo-Ortiz, N. Dickson, M. Drew-Brook, G. Rose, and A. Aspuru-Guzik, "Finding low-energy conformations of lattice protein models by quantum annealing," Sci. Rep. **2**, 571 (2012).
[7] Alejandro Perdomo-Ortiz, Alexander Feldman, Asier Ozaeta, Sergei V. Isakov, Zheng Zhu, Bryan O'Gorman, Helmut G. Katzgraber, Alexander Diedrich, Hartmut Neven, Johan de Kleer, Brad Lackey, and Rupak Biswas, "Readiness of quantum optimization machines for industrial applications," Phys. Rev. Applied **12**, 014004 (2019).
[8] Harry Markowitz, "Portfolio selection," The Journal of Finance **7**, 77–91 (1952).

Diagnosing barren plateaus with tools from quantum optimal control

Martín Larocca,[1,2] Piotr Czarnik,[2] Kunal Sharma,[3,2] Gopikrishnan Muraleedharan,[2] Patrick J. Coles,[2] and M. Cerezo[2,4]

[1] *Departamento de Física "J. J. Giambiagi" and IFIBA, FCEyN, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina*
[2] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3] *Hearne Institute for Theoretical Physics and Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA USA*
[4] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Variational Quantum Algorithms (VQAs) have received considerable attention due to their potential for achieving near-term quantum advantage. However, the observation of barren plateaus, a phenomenon by which the landscape becomes exponentially flat in the number of qubits, has raised uncertainty around their scalability. In this work we employ tools from quantum optimal control to develop a framework that can diagnose the presence, or absence, of barren plateaus for a certain class of periodic ansatzes.

A. Introduction

Quantum computers hold the promise to achieve computational speed-ups over classical supercomputers for certain tasks [1–4]. However, despite recent tremendous progress in quantum technologies, present-day quantum devices (known as Noisy Intermediate-Scale Quantum (NISQ) devices) are constrained by the limited number of qubits, connectivity, and by the presence of quantum noise [5]. One of the most promising computational models for making use of near-term quantum computers are Variational Quantum Algorithms (VQAs) [6]. Despite the wide application of VQAs, their widespread use is still limited by several challenges that can hinder their success. One of the main threats to VQA scalability is the so-called barren plateau phenomenon, where the cost function becomes untrainable due to gradients that vanish, on average, exponentially with the system size [7–18].

In this talk we will discuss the results in [19], i.e. how certain tools from quantum optimal control theory, in particular, the notions of controllability and dynamical lie algebra, can be leveraged to diagnose the presence (or absence) of barren plateaus in certain families of VQAs. Let us note that the framework developed here can be readily leveraged by the field of quantum machine learning, e.g. to study the trainability of certain classes of quantum neural networks, since these can be regarded as special cases of VQAs.

B. General framework

We consider an optimization task where the goal is to minimize a cost function of the form $C(\boldsymbol{\theta}) = \mathrm{Tr}[OU(\boldsymbol{\theta})\rho U^\dagger(\boldsymbol{\theta})]$. Here, ρ is an input state on n qubits in a d-dimensional Hilbert space with $d = 2^n$, $U(\boldsymbol{\theta})$ a parametrized quantum circuit (PQC), and O is a Hermitian operator that defines the task at hand. Moreover, we analyze layered parametrized quantum circuits that, as shown in Fig. 1.(i), have a periodic structure of the form

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\boldsymbol{\theta}_l), \quad U_l(\boldsymbol{\theta}_l) = \prod_{k=0}^{K} e^{-iH_k\theta_{lk}} \,. \tag{B1}$$

Here, the index l indicates the layer, $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lK})$ contains the parameters of such layer (such that $\boldsymbol{\theta} = \{\boldsymbol{\theta}_l\}_{l=1}^L$) and H_j are Hermitian traceless operators that generate the unitaries in the ansatz. Let us define $\mathcal{G} = \{H_k\}_{k=0}^K$ as the set of generators of the PQC. In what follows we refer to this type of ansatz as a Periodic Structure Ansatz (PSA). A standard result in quantum optimal control (QOC) relates the unitaries that can be generated upon variation of the parameters in the ansatz, to the so-called Dynamical Lie Algebra (DLA), defined as:

Definition 1 (Dynamical Lie Algebra)**.** *The Dynamical Lie Algebra (DLA) is the Lie Algebra generated by repeated nested commutators of the operators in set of generators $\mathcal{G}$, i.e.,*

$$\mathfrak{g} = \mathrm{span}\,\langle iH_0, \ldots, iH_K\rangle_{Lie} \,, \tag{B2}$$

where $\langle\mathcal{S}\rangle_{Lie}$ denotes the Lie closure, i.e., the set obtained by repeatedly taking the commutator of the elements in $\mathcal{S}$.

In particular, the expressible unitaries $U(\boldsymbol{\theta})$ belong to the so-called Dynamical Lie Group $\mathbb{G} = e^{\mathfrak{g}}$.

C. Results

The main result in [19] is that the dimension of the DLA plays a crucial role in determining if the cost function will, or will not, exhibit a barren plateau. First, we consider the case when the cases when the ansatz is *controllable*, that is, when its dynamical Lie algebra is full rank. Here, if the system has some a symmetry, so that the Hilbert space is $\mathcal{H} = \bigoplus_j \mathcal{H}_j$ with each $\mathcal{H}_j$ invariant under $\mathbb{G}$, then we can show that

$$\mathrm{Var}_{\boldsymbol{\theta}}[\partial_\mu C(\boldsymbol{\theta})] = \frac{2d_k}{(d_k^2-1)^2}\Delta(H_\mu^{(k)})\Delta(O^{(k)})\Delta(\rho^{(k)})\,. \tag{C1}$$

Here $\Delta(A) = D_{HS}\left(A, \mathrm{Tr}[A]\frac{\mathbb{1}_d}{d}\right)$, with $D_{HS}(A,B) = \mathrm{Tr}[(A-B)^2]$ the Hilbert-Schmidt distance, and where we defined $A^{(k)}$ as the reduction of operator A onto the subspace of $\mathcal{H}_k$. Hence, if ρ belongs to an invariant subspace where the system is controllable, then the scaling of the cost function partial derivative variance is determined by the dimension of the invariant subspace rather than by the dimension $d = 2^n$ of the Hilbert space. That is, the variance of the cost function partial derivatives will be exponentially vanishing in exponentially large subspaces and can be polynomially vanishing in polynomially large subspaces.

Here we recall that when analyzing barren plateaus, one usually studies the scaling of the variance of the cost function partial derivatives. This is due to the fact that an exponentially vanishing $\mathrm{Var}_{\boldsymbol{\theta}}[\partial_\mu C(\boldsymbol{\theta})]$ implies that the cost derivatives exponentially concentrate around its average of zero. That is, the landscape becomes exponentially flat.

Furthermore, we extend our results to the cases where the system is not controllable and Eq. (C1) does not hold. Here, we have numerical evidence pointing towards a connection between the dimension of the algebra and the vanishing of the variance, a notion that we formalize in the following observation

Observation 1. *Let the state ρ belong to a subspace $\mathcal{H}_k$ associated with a DLA $\mathfrak{g}_k$ and allow the depth of the PSA to be such that the distribution of unitaries generated by $U(\boldsymbol{\theta})$ has converged to the Haar measure in the Lie group $\mathbb{G}$. Then, the scaling of the variance of the cost function partial derivative is inversely proportional to the scaling of the dimension of the DLA as*

$$\mathrm{Var}_{\boldsymbol{\theta}}[\partial_\mu C(\boldsymbol{\theta})] \in \mathcal{O}\left(\frac{1}{\mathrm{poly}(\dim(\mathfrak{g}_k))}\right)\,. \tag{C2}$$

The implications of this observation are as follows. First, it shows that systems with a subspace DLA $\mathfrak{g}_k$ (i.e., the DLA restricted to subspace $\mathcal{H}_k$) that is polynomially growing with the system size can exhibit gradients that vanish polynomially with the system size, and hence may not exhibit a barren plateau. Conversely, systems with a subspace DLA that is exponentially growing with the system size can exhibit gradients that vanish exponentially with the system size, and hence could exhibit a barren plateau as the circuit depth increases. Crucially, Observation 1, can be used to diagnose the gradient scaling of the cost function by determining the size of the algebra generated by the set of generators of $U(\boldsymbol{\theta})$.

To further support the claim in Observation 1, we performed numerical simulations of VQAs with ansatzes with DLAs with different scalings (which we show in Fig. 1.(iii)). In particular we choose a Transverse Field Ising Model (TFIM) ansatz, defined by $\mathcal{G}_{\mathrm{TFIM}} = \{\sum_{i=1}^{n_f} Z_iZ_{i+1}, \sum_{i=1}^{n} X_i\}$ (see panel (a), $n_f = n-1$ and $n_f = n$ correspond to *open* and *closed* boundary conditions, respectively) and find that the variance of the cost function vanishes polynomially (see panel (b)), as predicted by the Observation and the fact that the underlying DLA, $\mathfrak{g}_{\mathrm{TFIM}}$, is polynomial. Interestingly, we find that a slight change in the set of generators $\mathcal{G}_{\mathrm{LTFIM}} = \mathcal{G}_{\mathrm{TFIM}} \bigcup \{\sum_{i=1}^{n} Z_i\}$ (the addition of a single extra generator) can lead to an exponential algebra and thus, as predicted by our Observation and confirmed by simulations (see, again, panel (b)), exhibit exponentially vanishing gradients.

D. Outlook

The results presented in [19] constitute a basic framework for diagnosing the presence of barren plateaus in VQAs using tools from QOC. The tools introduced here can be actively used to design ansatzes, as one could potentially predict if an ansatz, or a modification to the ansatz, will lead to the cost function exhibiting a barren plateau. Hence, our work can be considered as paving the way towards trainability-aware ansatz design.

While here we mainly focus on the trainability of ansatzes for near-term quantum computing, our results should also be considered as useful in the broader context of QOC. Moreover, we expect our results to impact the quantum machine learning community, as they could help in the desing of barren-plateau-free quantum neural networks.

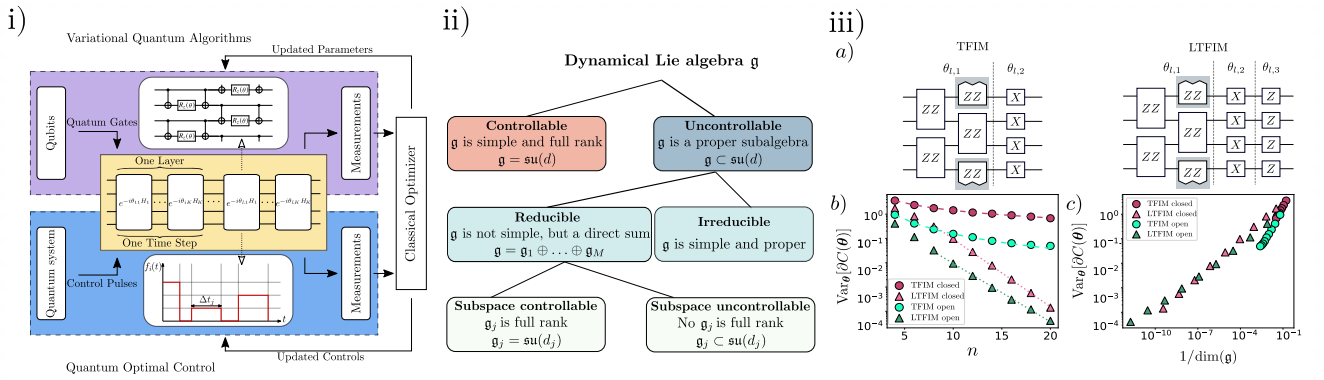

FIG. 1. **iii) Numerical results for the TFIM and LTFIM models.** a) Schematic illustration of a single layer of the ansates. b) Variance of the cost function partial derivative of the cost function in (**??**) versus the number of qubits n for each ansatz. The dashed (dotted) lines indicate the best polynomial (exponential) fit. The plot is shown in a log-linear scale. c) Variance of the cost function partial derivative versus $1/\dim(\mathfrak{g})$. The plot is shown in a log-log scale.

[1] Peter W Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th annual symposium on foundations of computer science* (Ieee, 1994) pp. 124–134.

[2] Aram W Harrow, Avinatan Hassidim, and Seth Lloyd, "Quantum algorithm for linear systems of equations," Physical Review Letters **103**, 150502 (2009).

[3] Dominic W Berry, Andrew M Childs, Richard Cleve, Robin Kothari, and Rolando D Somma, "Simulating hamiltonian dynamics with a truncated taylor series," Physical Review Letters **114**, 090502 (2015).

[4] Iulia M Georgescu, Sahel Ashhab, and Franco Nori, "Quantum simulation," Reviews of Modern Physics **86**, 153 (2014).

[5] John Preskill, "Quantum computing in the nisq era and beyond," Quantum **2**, 79 (2018).

[6] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles, "Variational quantum algorithms," Nature Reviews Physics **1**, 19–40 (2021).

[7] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven, "Barren plateaus in quantum neural network training landscapes," Nature communications **9**, 1–6 (2018).

[8] M Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles, "Cost function dependent barren plateaus in shallow parametrized quantum circuits," Nature communications **12**, 1–12 (2021).

[9] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles, "Noise-induced barren plateaus in variational quantum algorithms," arXiv preprint arXiv:2007.14384 (2020).

[10] Marco Cerezo and Patrick J Coles, "Higher order derivatives of quantum neural networks with barren plateaus," Quantum Science and Technology **6**, 035006 (2021).

[11] Kunal Sharma, M. Cerezo, Lukasz Cincio, and Patrick J Coles, "Trainability of dissipative perceptron-based quantum neural networks," arXiv preprint arXiv:2005.12458 (2020).

[12] Andrew Arrasmith, M. Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J Coles, "Effect of barren plateaus on gradient-free optimization," arXiv preprint arXiv:2011.12245 (2020).

[13] Zoë Holmes, Andrew Arrasmith, Bin Yan, Patrick J Coles, Andreas Albrecht, and Andrew T Sornborger, "Barren plateaus preclude learning scramblers," arXiv preprint arXiv:2009.14808 (2020).

[14] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe, "Entanglement induced barren plateaus," arXiv preprint arXiv:2010.15968 (2020).

[15] Taylor L Patti, Khadijeh Najafi, Xun Gao, and Susanne F Yelin, "Entanglement devised barren plateau mitigation," arXiv preprint arXiv:2012.12658 (2020).

[16] Arthur Pesah, M. Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles, "Absence of barren plateaus in quantum convolutional neural networks," arXiv preprint arXiv:2011.02966 (2020).

[17] Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J Coles, "Connecting ansatz expressibility to gradient magnitudes and barren plateaus," arXiv preprint arXiv:2101.02138 (2021).

[18] Andrew Arrasmith, Zoë Holmes, M Cerezo, and Patrick J Coles, "Equivalence of quantum barren plateaus to cost concentration and narrow gorges," arXiv preprint arXiv:2104.05868 (2021).

[19] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo, "Diagnosing barren plateaus with tools from quantum optimal control," arXiv preprint arXiv:2105.14377 (2021).

Adaptive shot allocation for fast convergence in variational quantum algorithms

Andrew Arrasmith,[1] Andi Gu,[1,2] Angus Lowe,[3] Pavel A. Dub,[4] and Patrick J. Coles[1]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[2] *Department of Physics, University of California, Berkeley, CA 94720, USA*
[3] *Department of Combinatorics and Optimization and Institute for Quantum Computing, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*
[4] *Chemistry Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Variational Quantum Algorithms (VQAs) are a promising approach for practical applications on near-term quantum computers. Here we present a new optimization method, called the global Coupled Adaptive Number of Shots (gCANS) method, which is efficient in both the number of iterations shots required for a full VQA optimization.

I. INTRODUCTION

Quantum computing may unlock previously intractable computations for a variety of applications in the physical sciences, industry, and beyond. However, the quantum computers that will be available in the near term are limited by having few qubits as well as hardware noise that limits the number of operations that can be performed before the information being manipulated degrades. Variational quantum algorithms (VQAs) [1, 2] are a promising approach to near-term quantum computing as they typically require many fewer qubits as well as much shorter run times than traditional quantum algorithms.

The reduced need for quantum resources allowed by VQAs comes at the cost of needing to classically optimize the control sequence, or ansatz, used to prepare the quantum state. This means that the efficiency of the method is largely determined by the computational expense of performing this optimization, which can often be non-trivial. The theoretical run-time (and monetary cost on some platforms) primarily depends on the number of different circuits run and the total number of shots. In order to realize time efficiency and affordability for VQAs, one therefore needs an optimizer that uses few iterations and few shots without requiring costly hyperparameter tuning.

In this work, we introduce a new optimization method that adaptively allocates shots for the measurement of each gradient component at each iteration. This optimizer, which we call the global-Coupled Adaptive Number of Shots (gCANS) method, uses a criterion for allocating shots that incorporates information about the overall scale of the shot cost for the iteration.

II. GCANS SHOT ALLOCATION RULE

We propose a new approach to SGD that, like iCANS [3], allows the number of shots per gradient component to vary. However, rather than allowing them to vary independently, we now optimize our expected gain globally over the entire gradient vector. That is, rather than taking an individual efficiency as in (**??**), our figure of merit is now:

Using $\mathcal{G}$ to denote the lower bound on the gain in stepping from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$) (i.e. $f(\boldsymbol{\theta}^{(t)}) - f(\boldsymbol{\theta}^{(t+1)})$), we are typically interested in its expectation:

$$\mathbb{E}[\mathcal{G}] = \left(\alpha - \frac{L\alpha^2}{2}\right)\left\|\boldsymbol{\nabla} f(\boldsymbol{\theta}^{(t)})\right\|^2 - \frac{L\alpha^2}{2}\sum_i \frac{\sigma_i^2}{s_i}\,. \tag{1}$$

where σ_i is the standard deviation of X_i.

$$\gamma = \frac{\mathbb{E}[\mathcal{G}]}{\sum_{k=1}^{d} s_k} \tag{2}$$

Using the first order optimality condition $\boldsymbol{\nabla}_{\boldsymbol{s}}\gamma = 0$ (further details provided in Appendix **??**), we obtain the rule:

$$s_i = \frac{2L\alpha}{(2 - L\alpha)}\frac{\sigma_i \sum_{k=1}^{d}\sigma_k}{\left\|\boldsymbol{\nabla} f(\boldsymbol{\theta})\right\|^2} \tag{3}$$

This results from a *global* metric for efficiency, hence we term this shot count prescription global coupled adaptive number of shots (gCANS). In addition to being shot frugal, we have also proven that this update step results in a geometric convergence rate for convex optimization.

Problem	Optimizer	Iterations (K)	Shots (S)	Cost (thousands of USD)	Estimated Time (hours)
He_2^+	iCANS	3015	4.6×10^7	127	12.86
	gCANS	**353**	$\mathbf{1.4 \times 10^7}$	**18**	**1.98**
	SGD-DS	853	3.5×10^7	44	4.86
	ADAM	1450	8.7×10^7	84	9.79
NH_3	iCANS	8301	5.1×10^8	5968	564.44
	gCANS	**1243**	$\mathbf{9.8 \times 10^7}$	**901**	**85.72**
	SGD-DS	1395	1.8×10^8	1036	100.10
	ADAM	2740	5.5×10^8	2104	207.51

TABLE I. We list the number of iterations and number of shots (averaged over 10 random starts) required to reach chemical accuracy for the four optimizers we compare for the equilibrium configurations for He_2^+ and NH_3. We also include the corresponding hypothetical cost, in USD, if these optimizers were to be used for VQE on Amazon Braket. The costs are computed with $C = 0.3PK + 0.00035S$ dollars, where K is the number of iterations, P is the number of Pauli terms in the expansion of $\hat{H}$, and S is the total number of shots used [5]. We estimate the wall clock time by assuming shots are taken with a frequency of 5 kHz (the sampling rate of [6]) and that it takes 0.1 seconds to change the circuit being run (following [7]). This gives a time estimate $T = 0.1PK + 0.0002S$ seconds. We neglect latency or time spent on classical update steps. Neither the cost or the time estimates accounts for the burden of the hyperparameter tuning, which would likely be substantial for performing SGD-DS.

III. NUMERICAL INVESTIGATION

We compare four optimizers: gCANS, iCANS[3], ADAM[4], and stochastic gradient descent with a shot budget increases geometrically with each iteration (hereafter called SGD-DS). We find that gCANS is largely insensitive to hyperparameter choice while performing as well or better than the other methods for VQE applied to He_2^+ as well as NH_3. See Table I for a comparison of the resource requirements to reach chemical accuracy for each of these methods.

IV. CONLCUSION

We contend gCANS should be considered the efficient optimization method of choice for variational algorithms and quantum machine learning as:

i) gCANS consistently outperforms each of the optimizers we test, achieving chemical accuracy with fewer shots and fewer circuit compilations. This translates to faster and cheaper (see Table I) experiments, bringing us closer to a practical implementation of VQE on near-term quantum computers.

ii) Similar to iCANS, gCANS is extremely robust to changes in its hyperparameters (see Appendix **??**), unlike optimizers such as SGD-DS. This robustness reduces the resources required to identify the appropriate settings of these hyperparameters.

iii) Unlike iCANS, which typically requires many iterations, gCANS has a proven geometric convergence rate.

[1] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 1(1):19–40, 2021.

[2] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. Noisy intermediate-scale quantum (nisq) algorithms. *arXiv preprint arXiv:2101.08448*, 2021.

[3] Jonas M Kübler, Andrew Arrasmith, Lukasz Cincio, and Patrick J Coles. An adaptive optimizer for measurement-frugal variational algorithms. *Quantum*, 4:263, 2020.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[5] Judd Winick, Steve Hamaker, and Morten Hansen. Amazon braket pricing, 2018. Accessed: 8/18/2021.

[6] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574:505–510, 10 2019.

[7] Kevin J Sung, Jiahao Yao, Matthew P Harrigan, Nicholas C Rubin, Zhang Jiang, Lin Lin, Ryan Babbush, and Jarrod R McClean. Using models to improve optimizers for variational quantum algorithms. *Quantum Science and Technology*, 5(4):044008, 2020.

Exploring Quantum Perceptron and Quantum Neural Network structures with a teacher-student scheme

Aikaterini Gratsea[1] and Patrick Huembeli[2]

[1]*ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels (Barcelona), Spain**

[2]*Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Abstract We introduce a teacher-student scheme that could systematically compare any QNN architectures and evaluate their relative expressive power. In this work, we focus particularly on two recent proposals. We also discuss alterations of these models to better understand the role of hidden units and the non-linearities in these architectures.

The full article can be found here: https://arxiv.org/abs/2105.01477. Code to reproduce the results and explore further settings can be found in the following Github repository: https://github.com/KaterinaGratsea/Teacher-student_scheme.

1. INTRODUCTION

Machine learning (ML) combined with quantum information processing gave rise to quantum machine learning (QML), which entails ML tasks performed (at least partially) on quantum computers. Recent results suggest possible advantages for generative QML applications compared to their classical analogs [1], but these models are unlikely to run on current quantum devices since they would require a fault tolerant universal quantum computer. In theory, to build a specific QML architectures on a quantum device, e.g. a quantum neural network (QNN), one could always make the classical logic of a neural network (NN) reversible and implement it via unitaries on a quantum computer [2]. This approach would require many qubits and error correction which is unrealistic for current quantum devices either. Furthermore, it is not clear if such an approach would yield any quantum advantage. In recent years, with the advent of Noisy Intermediate-Scale Quantum (NISQ) devices [3] near-term QML applications, such as variational quantum circuits (VQCs) have attracted increasing attention. VQCs are strong candidates for many general classical and quantum optimization applications [4–6], but also for the contruction of quantum perceptrons or neurons [2, 7, 8]. Following the classical analog of NN, QNNs can be built by casting together several of these quantum perceptrons [9, 10] to build deep structures with hidden layers. In general though, the role of these hidden units and non-linearities in QNNs that are built from VQCs is not as straightforward as for their classical counter part.

The aim of this works is to compare the relative expressive power of two conceptually different QML architectures. The first one is the quantum perceptron (QP) as described in [7], which can be cast together with other quantum perceptrons to build deep QNN structures [9]. The non-linear activations are introduced with dissipative units where ancillary qubits form the output layer, while the qubits from the input layers are discarded [11]. The second one is re-uploading architecture (RU) [12] which introduces the non-linear behaviour by repeatedly re-uploading the input data after each trainable unitary. We compare these two architectures to better understand the dissipative nature of the deferred measurement and the re-uploading of the data. To have a fair comparison, we use the same data encoding architecture in both models and to avoid the effect of data selection we deploy a so-called teacher-student scheme, where each architecture will play once the role of the teacher and once the student. Even more, we use different realizations of the teachers and obtain the average performances for the students. This scheme offers a systematic comparison which aims to understand the role of the hidden units and non-linearities in quantum models.

2. TEACHER-STUDENT SCHEME

Here, we introduce the teacher-student scheme that aims to to benchmark different realizations of QP and QNN against each other. We systematically compare the two aforementioned architectures (QP and RU), but the scheme could be used for any circuits. The main idea is that one architecture (for example the QP) will play the role of the teacher and generate the labels that will be used to train the student (for example the re-uploading quantum model). Thus, we avoid to generate artificial data sets, such as e.g the circle data set in [12], that could possibly favour one of the architectures. With the teacher-student scheme, we directly see the data structures that each architecture can generate and how well other architectures could learn them.

* gratsea.katerina@gmail.com

a. Notion of teacher The teacher generates the labels for a fixed set of inputs $\mathcal{D} = \{\boldsymbol{x}^k\}$ with 2 dimensional input vectors $\boldsymbol{x} = \{x_1, x_2\}$ on a grid $x_i \in [-\pi, \pi]$. For a fixed teacher architecture we choose several random initializations for the parameters $\boldsymbol{w}$ of the processing gates. We use the measurement outcomes of the ancilla qubit as the model predictions y^k of the input data. This way we can generate several data sets for different random initializations. The predictions have continuous values $y^k \in [-1, 1]$ given by the outcome of the measurement $\langle\psi^k| Z |\psi^k\rangle$, but we also generate binary valued labels by choosing $y^k_{\text{binary}} = \text{sign}(y^k)$. The teachers with binary labels focus more on the basic characteristics of the data structures, while the ones with continuous labels also care for the details. We can visualize the data structures with *prediction maps*, which are the density plots of the model predictions and labels y^k for the input data $\boldsymbol{x}^k$.

b. Notion of student We train the students with the labeled data generated by their teachers to learn those data structures. It is not obvious how to define a good/bad student, since different tools can be used to characterize their performance. The *prediction maps* of the students are best for visualizing the similarity of the student's and teacher's predictions of the label y^k to gain qualitative results. For a more rigorous quantitative comparison we compute the *relative entropy* between the student's and teacher's outputs y^k. Specifically, we use the information divergence (Kullback–Leibler divergence or relative entropy) which defines a distinguishable measure between two probability distributions P and Q [13]:

$$S(P\|Q) = \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i}. \tag{1}$$

When the two distributions are similar, the value of the relative entropy is close to zero. To interpret the predicted labels y^k as probabilities, we offset and re-normalize them ($y^k > 0$, $\sum_{\boldsymbol{x}^k \in \mathcal{D}} y^k = 1$). Then, to compare two prediction maps, the information divergence is calculated by summing over the whole input space. Here, we are interested in the average relative entropy of all teacher-student pairs. Another qualitative metric is the *loss function* which determines the success of the training. When the student is trained with the binary valued labels the percentage of the correctly predicted labels can be computed. We refer to this as the *accuracy score* which gives an overall performance of the student. To identify a good/bad student all these tools should be taken into account.

3. CONCLUSIONS

Inspired by the recent works [7, 12], we explored the expressive power of QPs, their formation to QNNs and the RU models implemented on NISQ devices. In order to systematically compare the architectures, we introduced a so-called teacher-student scheme, where the studied models are introduced once as a teacher and once as a student. This way we can avoid to generate synthetic data sets that might give an advantage to certain architectures and it creates a more fair framework for comparing any quantum models.

Specifically, we showed that the deep structures that can be built with QPs only increase the expressivity of a model if the data are uploaded several times. It is not sufficient to use deferred measurements to generate hidden non-linearities similar to classical NNs if the output of a QP is reused. We explored several different ways of how to leverage deferred measurements to generate hidden non-linearities, but the expressive power of QNNs only improved when additional data-uploadings were added. This suggests that the non-linear behaviour induced by a measurement of a single QP cannot be generalized to deep QNNs if the single QPs are cast together in a coherent way. Therefore, it is still an open question how to build deep QNNs in a coherent way, where measurements only occur at the end of the computation. Thus, one should not expect a one-to-one mapping of quantum and classical NNs.

These results are in accordance with the recent work [14], which shows that the number of times that the data are encoded determines the functions that can be approximated. The needed non-linearities in a quantum model can be generated (apart from the measurement) from the encoding gates that are non-linear functions of the input data. Performing PCA on the probability vectors, we showed that given the encoding, the data can already be separated without further processing. Therefore, the performance of a QP is strongly affected by the encoding and the dataset itself. Apart from the encoding, the processing plays an important role as well. The universal approximation capability of different quantum models has been discussed extensively in [12, 15–17], but it does not provide any information about how well the circuit could perform or how many parameters it needs to approximate a function within a certain error. The calculation of the average relative entropy showed that the trainable part of the circuit affects the functions that can be approximated. This effect will be further explored in subsequent research.

For future work, it will be of great interest to explore different perceptron models and compare their performance with the teacher-student scheme. Then, the question arises which perceptron model will be the ideal building block of QNN architectures and how quantum perceptrons could be combined to form a deep QNN. Another research direction is to explore other quantum models with no direct analog with classical NN, like the re-uploading model or quantum kernels in general. Finally, it would be of great importance to further explore the role of entanglement and encoding in QPs, in their formation to QNNs and in other quantum models.

REFERENCES

[1] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, "On the quantum versus classical learnability of discrete distributions," (2020), arXiv:2007.14451 [quant-ph].
[2] K. H. Wan, O. Dahlsten, H. Kristjánsson, R. Gardner, and M. S. Kim, npj Quantum Information **3** (2017), 10.1038/s41534-017-0032-4.
[3] J. Preskill, Quantum **2**, 79 (2018).
[4] E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm," (2014), arXiv:1411.4028 [quant-ph].
[5] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, Nature Communications **5** (2014), 10.1038/ncomms5213.
[6] E. Farhi and H. Neven, "Classification with quantum neural networks on near term processors," (2018), arXiv:1802.06002 [quant-ph].
[7] F. Tacchino, C. Macchiavello, D. Gerace, and D. Bajoni, npj Quantum Information **5** (2019), 10.1038/s41534-019-0140-4.
[8] S. Yan, H. Qi, and W. Cui, Physical Review A **102** (2020), 10.1103/physreva.102.052421.
[9] F. Tacchino, P. Barkoutsos, C. Macchiavello, I. Tavernelli, D. Gerace, and D. Bajoni, Quantum Science and Technology **5**, 044010 (2020).
[10] E. Torrontegui and J. J. García-Ripoll, EPL (Europhysics Letters) **125**, 30004 (2019).
[11] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, "Trainability of dissipative perceptron-based quantum neural networks," (2020), arXiv:2005.12458 [quant-ph].
[12] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Quantum **4**, 226 (2020).
[13] I. Bengtsson and K. Zyczkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement* (Cambridge University Press, 2006).
[14] M. Schuld and N. Killoran, Physical Review Letters **122** (2019), 10.1103/physrevlett.122.040504.
[15] M. Schuld, R. Sweke, and J. J. Meyer, "The effect of data encoding on the expressive power of variational quantum machine learning models," (2020), arXiv:2008.08605 [quant-ph].
[16] A. Pérez-Salinas, J. Cruz-Martinez, A. A. Alhajri, and S. Carrazza, "Determining the proton content with a quantum computer," (2021), arXiv:2011.13934 [hep-ph].
[17] A. Pérez-Salinas, D. López-Núñez, A. García-Sáez, P. Forn-Díaz, and J. I. Latorre, "One qubit as a universal approximant," (2021), arXiv:2102.04032 [quant-ph].

Towards understanding the power of quantum kernels in the NISQ era

Xinbiao Wang Yuxuan Du * Yong Luo Dacheng Tao

Abstract

In this work, we theoretically and empirically study the power of quantum kernels in the NISQ era. We first prove that the advantage of quantum kernels vanishes for large size of datasets, few number of measurements, and large system noise. With the aim of preserving the superiority of quantum kernels in the NISQ era, we further devise an effective method via indefinite kernel learning. Numerical simulations accord with our theoretical results. Arxiv

1 Introduction

A key problem in the field of quantum computing is understanding whether quantum machine learning (QML) models implemented on noisy intermediate-scale quantum (NISQ) machines can achieve quantum advantages. Recently, Huang et al. [1] partially answered this question by the lens of quantum kernel learning. Namely, they exhibited that quantum kernels can learn specific datasets with lower generalization error over the optimal classical kernel methods. Despite the promising achievements, most of the theoretical results in [1] are established on the ideal setting. In particular, they assumed that the number of measurements is infinite and the exploited quantum system is noiseless, where both of them are impractical for NISQ devices. The quantum kernel returned by NISQ machines, affected by the system noise and a finite number of measurements, may be indefinite and therefore does not obey the results claimed in [1]. Driven by attractive merits comprised by quantum kernel methods and the deficiencies of near-term quantum machines, a crucial question is: *Does the power of quantum kernels still hold in the NISQ era?* A positive affirmation of this question will not only contribute to a wide range of machine learning tasks to gain prediction advantages but can also establish the quantum deep learning theory.

Problem setup. let us first recap the necessary notations and the explicit form of the achieved generalization error bound. Denote $\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^n$ as the training data, $\mathrm{W} \in \mathbb{R}^{n\times n}$ as the ideal quantum kernel whose element equals to $\mathrm{W}_{ij} = \mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})\rho(\boldsymbol{x}^{(j)}))$, $\forall i,j \in [n]$, where $\rho(\boldsymbol{x}^{(i)})$ refers to the density operator of the encoded quantum data with respect to $\boldsymbol{x}^{(i)}$. In the NISQ scenario, the depolarization channel

$$\mathcal{N}_p(\rho) = (1-p)\rho + \frac{p\mathbb{I}_{2^N}}{2^N} \tag{1}$$

with depolarizing rate p and the finite number of measurements m are considered. Thereby, the estimated quantum kernel is denoted by $\widehat{\mathrm{W}}$ whose element yields

$$\widehat{\mathrm{W}}_{ij} = \frac{1}{m}\sum_{k=1}^{m} V_k,\ \forall i,j \in [n], \tag{2}$$

where $V_k \sim \mathrm{Ber}(\widetilde{\mathrm{W}}_{ij})$ is the output of a quantum measurement and $X \sim \mathrm{Ber}(p)$ refers to the Bernoulli distribution with $\Pr(X=0)=p$ and $\Pr(X=1)=1-p$, and $\widetilde{\mathrm{W}}_{ij} = \mathrm{Tr}(\mathcal{N}_p(\rho(\boldsymbol{x}^{(i)})\rho(\boldsymbol{x}^{(j)})))$. The aim of quantum kernels learning in NISQ era is using the quantum kernel $\widehat{\mathrm{W}}$ instead of W to infer a hypothesis.

2 results

Theorem 1. *Let the size of training dataset be n and the number of measurements is m. Define $Y = [y_1, \cdots, y_n]^\top$ as the label vector and $c_{\mathrm{W}} = \| \mathrm{W}^{-1} \|_2$. Suppose the system noise is modeled by $\mathcal{N}_p$ in Eqn. (1). With probability*

*Corresponding author, duyuxuan123@gmail.com

at least $1-\delta$*, the noisy quantum kernel* $\widehat{W}$ *in Eqn. (2) can be used to infer a hypothesis* $h(\boldsymbol{x})$ *with generalization error*

$$\mathbb{E}_{\boldsymbol{x},\widehat{\mathrm{W}}}\left|h(\boldsymbol{x})-y\right| \leq \tilde{O}\left(\sqrt{\frac{c_1}{n}+\sqrt{\frac{1}{c_2}\frac{n}{\sqrt{m}}}}\right) \tag{3}$$

where $c_1 = Y^\top \mathrm{W}^{-1} Y$ *and* $c_2 = \max(c_{\mathrm{W}}^{-2}((\frac{1}{2}\log(\frac{4n^2}{\delta}))^{\frac{1}{2}} + m^{\frac{1}{2}}p\left(1+\frac{1}{2^{N+1}}\right))^{-1} - \frac{n}{\sqrt{m}}c_{\mathrm{W}}^{-1}, 0)$.

Notably, when the depolarization noise is considered, the generalization error of the noisy quantum kernel $\mathbb{E}_{\boldsymbol{x},\widehat{\mathrm{W}}}|h(\boldsymbol{x})-y|$ *will always have a term* $n^{1/4}$.

Theorem [1] provides a central theoretical contribution for this paper that a larger data size n, a higher system noise p, and a fewer number of measurements m will make the generalization advantage of quantum kernels *inconclusive*. This result indicates a negative conclusion of using quantum kernels implemented on NISQ devices to tackle large-scale learning tasks with evident advantages, which is contradicted with the claim of the study [1] such that a larger data size n promises a better generalization error.

Our second contribution is empirically demonstrating that under the NISQ setting , the advantage of quantum kernels may be preserved by suppressing the estimation error of the kernel matrix. Concretely, we adopt the advanced spectral transformation techniques, which are developed in indefinite kernel learning[2, 3, 4, 5], to alleviate the negative effect induced by the system noise and the finite measurements. The corresponding theoretical result is summarized in Lemma 1. Numerical simulation results demonstrate that the performance of noisy quantum kernels can be improved by 14% (see figure 1). Our work opens up a promising avenue to combine classical indefinite kernel learning methods with quantum kernels to attain quantum advantages in the NISQ era.

Lemma 1. *Let* W *and* $\widehat{\mathrm{W}}$ *be the ideal and noisy quantum kernel respectively. Applying the spectral transformation techniques to* $\widehat{\mathrm{W}}$*, the obtained kernel* $\widehat{\mathrm{W}}_\diamond \in \left\{\widehat{\mathrm{W}}_{\mathrm{c}}, \widehat{\mathrm{W}}_{\mathrm{f}}, \widehat{\mathrm{W}}_{\mathrm{s}}\right\}$ *yields*

$$\left\|\mathrm{W}-\widehat{\mathrm{W}}_\diamond\right\|_{\mathrm{F}} \leq \left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_{\mathrm{F}}, \tag{4}$$

where $\|\cdot\|_{\mathrm{F}}$ *refers to the Frobenius norm, and* $\widehat{\mathrm{W}}_{\mathrm{c}}, \widehat{\mathrm{W}}_{\mathrm{f}}, \widehat{\mathrm{W}}_{\mathrm{s}}$ *refer to matrices calibrated by three different spectral transformation methods, i.e., clipping[3], flipping[6], and shifting[7] methods.*

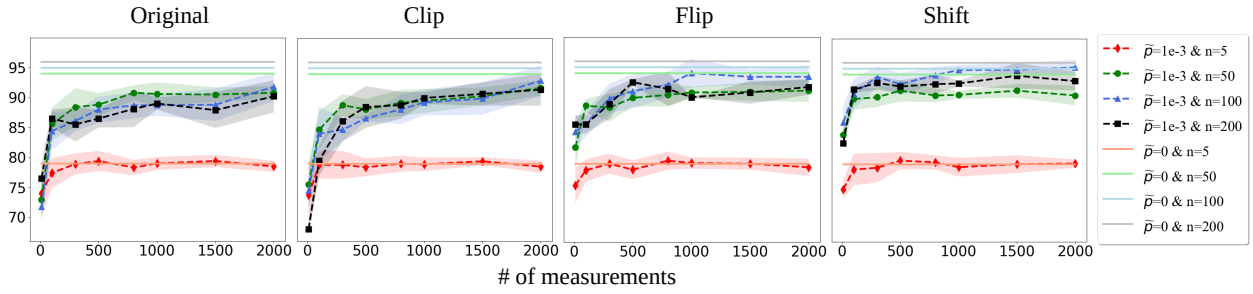

Figure 1: **The comparison of noisy quantum kernels with different calibration methods.** The left subplot depicts the simulation results of noisy quantum kernels calibrated by the nearest projection as labeled by 'Original'. The rest three subplots from left to right illustrates the simulation results calibrated by the spectral transformation methods, i.e., clipping, flipping, and shifting methods.

3 Summary

In this study, we investigate the generalization performance of quantum kernels under the NISQ setting. We theoretically exhibit that a large size of the training dataset, a small number of measurement shots, and a large amount of quantum system noise can destroy the superiority of quantum kernels. To improve performance of quantum kernels in the NISQ era, we further prove that effective spectral transformation techniques have the potential to maintain the advantage of quantum kernels in the NISQ era. Besides the theoretical results, we empirically demonstrate that spectral transformation techniques have the capability of improving performance of noisy quantum kernels for both the depolarization noise and noise extracted from the real quantum-hardware (IBMQ-Melbourne). The achieved results in this study fuel the exploration of quantum kernels assisted by other advanced calibration methods to accomplish practical tasks with advantages in the NISQ era.

References

[1] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature communications*, 12(1):1–9, 2021.

[2] Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J Smola. Learning with non-positive kernels. In *Proceedings of the twenty-first international conference on Machine learning*, page 81, 2004.

[3] Gang Wu, Edward Y Chang, and Zhihua Zhang. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In *Proceedings of the 22nd International Conference on Machine Learning*, volume 8. Citeseer, 2005.

[4] Ronny Luss and Alexandre d'Aspremont. Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, 1(2):97–118, 2009.

[5] Yihua Chen, Eric K Garcia, Maya R Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(3), 2009.

[6] Thore Graepel, Ralf Herbrich, Peter Bollmann-Sdorra, and Klaus Obermayer. Classification on pairwise proximity data. *Advances in neural information processing systems*, pages 438–444, 1999.

[7] Volker Roth, Julian Laub, Motoaki Kawanabe, and Joachim M Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1540–1551, 2003.

Quantum evolution kernel : Machine learning on graphs with programmable arrays of qubits

Louis-Paul Henry,[1] Slimane Thabet,[1,*] Constantin Dalyac,[1,2] and Loïc Henriet[1]

[1] *Pasqal, 2 avenue Augustin Fresnel, 91120 Palaiseau*
[2] *LIP6, CNRS, Sorbonne Université, 4 Place Jussieu, 75005 Paris, France*

(Dated: September 14, 2021)

We introduce a graph kernel based on the time-evolution of a quantum system, whose Hamiltonian encodes the topology of the input graph. We study analytically the procedure and show numerically that it performs well compared to standard kernels. Finally, we characterize the implementation on a neutral-atom quantum processor.

INTRODUCTION

In many fields such as chemistry [1, 2], bioinformatics [3, 4], social network analysis[5], computer vision[6], or natural language processing [7, 8], some observations have an inherent graph structure, requiring the development of specific algorithms to exploit the information contained in their structures. A large body of work exists in the classical machine learning literature, trying to study graphs through the use of *graph kernels* that are measures of similarity between graphs[9–12]. The idea behind the graph kernel approach is very generic, and consists first in finding a way to associate any graph with a *feature vector* encapsulating its relevant characteristics (the *feature map*) and then to compute the similarity between those vectors, in the form of a scalar product in the feature space.

Previous work on quantum graph kernels has been done[13–15], both as generic algorithms introducing novel approaches inspired by quantum mechanics (Quantum graph neural networks[16], Quantum-walk kernels[17, 18]), as well as with a more specific computing platform in mind (*e.g.* using photonic devices[19]).

Here we propose a more versatile and easily scalable graph kernel based on similar ideas. The core principle is to encode the information about a graph in the parameters of a tunable Hamiltonian acting on an array of qubits and to measure a carefully chosen observable after an alternating sequence of *free* evolution (*i.e.* with this Hamiltonian) and/or parametrized pulses, similarly to what is done in the Quantum Approximate Optimization Algorithm (QAOA)[20], or after a continuously parametrization of the Hamiltonian, similarly to what can be done in optimal control[21]. This kernel can be realized with state-of-the-art Quantum Processing Units(QPUs), in particular with Rydberg atom processors [22–24], in which highly tunable Hamiltonians can be realized to encode a wide range of graph topologies with up to hundreds of qubits.

I. QUANTUM EVOLUTION KERNEL

The time-evolution of a quantum state on a graph is a rich source of features for machine learning tasks such as the aforementioned classification and regression. We present here a new Quantum Evolution Kernel (QEK), using the dynamics of an interacting quantum system as a tool to characterize graphs. The approach we propose consists in associating each graph with a probability distribution, obtained through the measurement of an observable on a quantum system

* slimane.thabet@pasqal.io

whose dynamics is driven by the topology of the graph. The QEK between two graphs and is then given as a distance between their respective probability distributions.

More specifically, we consider a system whose time-evolution is governed by the Hamiltonian $\hat{\mathcal{H}}_{\mathcal{G}}$, that can be any parametrized Hamiltonian whose topology of interactions is that of the graph under study. We introduce then another Hamiltonian $\hat{\mathcal{H}}_{\theta}$, parametrized by a set of parameters θ, independent of the graph $\mathcal{G}$ and use it to apply *pulses* to the system (*i.e.* letting it evolve with a coherent single-qubit driving $\hat{\mathcal{H}}_{\theta}$).

The system starts in a predefined state $|\psi_0\rangle$. After an initial pulse with $\hat{\mathcal{H}}_{\theta_0}$, we alternatively let the system evolve with $\hat{\mathcal{H}}_{\mathcal{G}}$ (for a duration τ_i) and $\hat{\mathcal{H}}_{\theta_i}$. This time evolution can be summed up in the set of parameters $\Lambda = \{\theta_0, t_1, \theta_1, \ldots, t_p, \theta_p\}$ After the time-evolution the system ends up in the state

$$|\psi_f\rangle = \prod_{i=1}^{p} \left(\mathrm{e}^{-\mathrm{i}\hat{\mathcal{H}}_{\theta_i}} \mathrm{e}^{-\mathrm{i}\hat{\mathcal{H}}_{\mathcal{G}} t_i} \right) \mathrm{e}^{-\mathrm{i}\hat{\mathcal{H}}_{\theta_0}} |\psi_0\rangle \tag{1}$$

Once the system has been prepared in the final state $|\psi_f\rangle$, an observable $\hat{\mathcal{O}}$ is measured, to be used in the construction of the probability distribution. We present here several possible choices for this distribution.

For two graphs $\mathcal{G}$ and $\mathcal{G}'$, and their respective probability distributions $\mathcal{P}$ and $\mathcal{P}'$, we define the graph kernel as $\mathcal{K}_{\mu}(\mathcal{G}, \mathcal{G}') = \exp\left[-\mu\, JS(\mathcal{P}, \mathcal{P}')\right]$ where we have chosen the distance between the two distributions as their Jensen-Shannon divergence $JS(\mathcal{P}, \mathcal{P}')$.

II. EXPERIMENTS AND RESULTS

We tested the protocol on a graph classification task in association with a Support Vector Machine (SVM) on several datasets. We compared the accuracy score with other graph kernels on the same task, the Random Walk (RW) [25] kernel and the Graphlet Subsampling (GS)[26, 27] kernel. The parameter Λ is trained by Bayesian Optimization on the final score function. The datasets are taken from the repository of the Technical University of Dortmund [28]. For all datasets, at least one quantum kernel is better than the best classical kernel.

Furthermore, we show that this method is readily amenable to an already existing platform, based on Rydberg atoms, through direct emulation of its dynamics. In this implementation, the topology of the graph is encoded in the interactions between individual atoms, determined by their spatial arrangement. Our results show that, even in presence of noise, reasonable accuracies can be achieved, as compared to already existing kernels.

FULL WORK AND CODE AVAILABILITY

The full manuscript can be accessed at https://arxiv.org/abs/2107.03247. It has been accepted for publication in Physical Review A. A tutorial following the structure of the paper is available at https://pulser.readthedocs.io/en/stable/tutorials/qek.html. The extensive code is available at https://github.com/pasqal-io/qgraph.

[1] Alexandre Varnek and Igor Baskin. Machine learning methods for property prediction in chemoinformatics: Quo vadis? *Journal of Chemical Information and Modeling*, 52(6):1413–1437, 2012. doi: 10.1021/ci200409x. URL https://doi.org/10.1021/ci200409x. PMID: 22582859.

[2] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/gilmer17a.html.

[3] Giulia Muzio, Leslie O'Bray, and Karsten Borgwardt. Biological network analysis with deep learning. *Briefings in Bioinformatics*, 22(2):1515–1530, 11 2020. ISSN 1477-4054. doi:10.1093/bib/bbaa257. URL https://doi.org/10.1093/bib/bbaa257.

[4] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. Vishwanathan, A. J. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21, 2005. doi:10.1093/bioinformatics/bti1007.

[5] John Scott. Social network analysis: developments, advances, and prospects. *Social Network Analysis and Mining*, 1(1), 01 2011. doi:10.1007/s13278-010-0012-6. URL https://doi.org/10.1007/s13278-010-0012-6.

[6] Zaïd Harchaoui and Francis Bach. Image classification with segmentation graph kernels. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[7] Giannis Nikolentzos, Polykarpos Meladianos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Shortest-path graph kernels for document similarity. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1890–1900, 2017.

[8] Goran Glavaš and Jan Šnajder. Recognizing identical events with graph kernels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 797–803, 2013.

[9] Pierre Latouche and Fabrice Rossi. Graphs in machine learning: an introduction. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 207–218, April 2015.

[10] Niels M. Kriege, Fredrik D. Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(6), 2020. doi:10.1007/s41109-019-0195-3. URL https://doi.org/10.1007/s41109-019-0195-3.

[11] Karsten M. Borgwardt, M. Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Rieck. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6), 2020. doi:10.1561/2200000076. URL https://doi.org/10.1561/2200000076.

[12] B. Schlköpf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001.

[13] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Phys. Rev. Lett.*, 122:040504, Feb 2019. doi:10.1103/PhysRevLett.122.040504. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.040504.

[14] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi:10.1038/s41586-019-0980-2. URL https://doi.org/10.1038/s41586-019-0980-2.

[15] Kaito Kishi, Takahiko Satoh, Rudy Raymond, Naoki Yamamoto, and Yasubumi Sakakibara. Graph kernels encoding features of all subgraphs by quantum superposition, 2021.

[16] Guillaume Verdon, Trevor McCourt, Enxhell Luzhnica, Vikash Singh, Stefan Leichenauer, and Jack Hidary. Quantum Graph Neural Networks. *arXiv e-prints*, art. arXiv:1909.12264, September 2019.

[17] Luca Rossi, Andrea Torsello, and Edwin Hancock. Measuring graph similarity through continuous-time quantum walks and the quantum jensen-shannon divergence. *Physical Review E*, 91:022815, 02 2015. doi:10.1103/PhysRevE.91.022815.

[18] Lu Bai, Luca Rossi, Peng Ren, Zhihong Zhang, and Edwin R. Hancock. A Quantum Jensen-Shannon Graph Kernel Using Discrete-Time Quantum Walks. In Cheng-Lin Liu, Bin Luo, Walter G. Kropatsch, and Jian Cheng, editors, *Graph-Based Representations in Pattern Recognition*, pages 252–261, Cham, 2015. Springer International Publishing. ISBN 978-3-319-18224-7. URL https://doi.org/10.1007/978-3-319-18224-7_25.

[19] Maria Schuld, Kamil Brádler, Robert Israel, Daiqin Su, and Brajesh Gupt. Measuring the similarity of graphs with a gaussian boson sampler. *Phys. Rev. A*, 101:032314, Mar 2020. doi:10.1103/PhysRevA.101.032314. URL https://link.aps.org/doi/10.1103/PhysRevA.101.032314.

[20] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A Quantum Approximate Optimization Algorithm. *arXiv e-prints*, art. arXiv:1411.4028, Nov 2014.

[21] J Werschnik and E K U Gross. Quantum optimal control theory. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 40(18):R175–R211, sep 2007. doi:10.1088/0953-4075/40/18/r01. URL https://doi.org/10.1088/0953-4075/40/18/r01.

[22] Hendrik Weimer, Igor Müller, MarkusLesanovsky, Peter Zoller, and Hans Peter Büchler. A rydberg quantum simulator. *Nature Physics*, 6:382–388, Mar 2010. doi:10.1038/nphys1614. URL https://doi.org/10.1038/nphys1614.

[23] Hannes Bernien, Sylvain Schwartz, Alexander Keesling, Harry Levine, Ahmed Omran, Hannes Pichler, Soonwon Choi, Alexander S. Zibrov, Manuel Endres, Markus Greiner, Vladan Vuletić, and Mikhail D. Lukin. Probing many-body dynamics on a 51-atom quantum simulator. *Nature*, 551(7682):579–584, November 2017. doi:10.1038/nature24622.

[24] Henning Labuhn, Daniel Barredo, Sylvain Ravets, Sylvain de Léséleuc, Tommaso Macrì, Thierry Lahaye, and Antoine Browaeys. Tunable two-dimensional arrays of single Rydberg atoms for realizing quantum Ising models. *Nature*, 534(7609):667–670, Jun 2016. doi:10.1038/nature18274.

[25] S.V.N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(40):1201–1242, 2010. URL http://jmlr.org/papers/v11/vishwanathan10a.html.

[26] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23 (2):e177–e183, 01 2007. ISSN 1367-4803. doi:10.1093/bioinformatics/btl301. URL https://doi.org/10.1093/bioinformatics/btl301.

[27] Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 488–495, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/shervashidze09a.html.

[28] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL www.graphlearning.io.

Testing identity of collections of quantum states: sampling complexity analysis

Marco Fanizza * Raffaele Salvia† Vittorio Giovannetti ‡

Full version at `arXiv:2103.14511`

Introduction

We study the problem of testing identity of a collection of unknown quantum states given sample access to the collection. We show that for a collection of d-dimensional quantum states of cardinality N, the *sample complexity* is $O(\sqrt{N}d/\epsilon^2)$, which is optimal up to a constant. We assume a *sampling model access*, where each state appears with some known probability, adapting [LRR13, DK16] to the quantum case. The test is obtained by estimating the mean squared Hilbert-Schmidt distance between the states, thanks to a suitable generalization of the estimator of the Hilbert-Schmidt distance between two unknown states of [BOW19].

This problem is an example of *property testing*, a concept developed in computer science [Gol17], and applied to hypothesis testing of distributions [Can20] and quantum states and channels [MdW16]. At variance with optimal asymptotic error rates studied in statistical classical and quantum hypothesis testing [LR06, Hay16], the sample complexity captures finite size effect in inference problems, as it expresses the number of samples required to successfully execute an inference task in terms of the extensive parameters of the problem, in our case the dimension d and the cardinality N of the collection. The interest in these kind of questions in the classical case has been motivated by the importance of the study of big data sources; a similar motivation holds for the quantum case, since outputs of fully functional quantum computers will also live in high-dimensional spaces.

Related work

The problem of learning properties of classical distributions in the property testing approach is a vast research area [Gol17, Can20]. Since learning a classical distribution on a set of cardinality d, in total variational distance, can be done in $O(d/\epsilon^2)$ samples [Gol17], the interest in testing properties is to get a sample complexity $o(d)$. Identity testing for two unknown distribution has a sample complexity $\Theta(\max(d^{1/2}/\epsilon^2, d^{2/3}/\epsilon^{4/3}))$ [CDVV14]. The problem of testing identity of collection of N distributions was introduced in the classical case in [LRR13] and solved in [DK16], obtaining $\Theta(\max(\sqrt{dN}/\epsilon^2, d^{2/3}N^{1/3}/\epsilon^{4/3}))$ for the sampling model, where at each sample the tester receives one of N distributions with probability p_i, and $\Theta(\max(\sqrt{d}/\epsilon^2, d^{2/3}/\epsilon^{4/3}))$ for the query model, where the tester can choose the distribution to call at each sample. In the quantum case, *quantum tomography*, requires $\Theta(d^2/\epsilon^2)$ copies of the state [HHJ+17, OW16, OW17]. These algorithms require spectrum learning as a subroutine [ARS88, KW01, HM02, CM06, Key06], which

*marco.fanizza@sns.it, NEST, Scuola Normale Superiore and Istituto Nanoscienze-CNR, I-56127 Pisa, Italy.
†raffaele.salvia@sns.it, Scuola Normale Superiore, I-56127 Pisa, Italy.
‡NEST, Scuola Normale Superiore and Istituto Nanoscienze-CNR, I-56127 Pisa, Italy.

has the sample complexity $O(d^2/\epsilon^2)$ [OW15]. The measurement used for spectrum learning is known as weak Schur sampling, and it can be efficiently implemented, with gate complexity $O(n, \log d, \log 1/\delta)$ [BCH06, Har05, Kro19], where n is the number of copies of the state, and δ is the precision of the implementation. Weak Schur sampling is a key ingredient in measurement with optimal sample complexity for testing identity to the completely mixed state $O(d/\epsilon^2)$ [OW15] and identity testing between unknown states [BOW19]. In [BOW19], identity testing between unknown states is done by first estimating their Hilbert-Schmidt distance with a minimum variance unbiased estimator, developing a general framework for efficient estimators of traces of polynomials of states. This improves on a primitive way to estimate the overlap $\mathrm{Tr}[\rho\sigma]$ between two unknown states, the swap test [BCWdW01], while optimal estimation of the overlap between pure states with average error figures of merit has been addressed by a series of works [BRS04, BIMT06, LSB06, GI06, FRS$^+$20]. The present work develops the method of [BOW19], building a test based on a nested weak Schur sampling measurement. The sample complexity of testing identity of collections of quantum states in the query model was established to be $\Theta(d/\epsilon^2)$ in [Yu19], with an algorithm which makes direct use of the Hilbert-Schmidt estimator of [BOW19] on chosen couples ρ_i, ρ_j. This measurement is different from the collective nested weak Schur sampling measurement we employ in our test.

Results

Given a collection of d-dimensional quantum states $\{\rho_i\}_{i=1,\dots,N}$, and a probability distribution p_i, we consider a *sampling model* [LRR13, DKN15] where we have access to copies of the state

$$\rho = \sum_{i=1}^{N} p_i \, |i\rangle\langle i| \otimes \rho_i. \tag{1}$$

where $\{|i\rangle\}_{i=1,\dots,N}$ is an orthonormal basis of $\mathbb{C}^N$. We are promised that one of the two following properties holds:

- **Case** A: $\rho_1 = \rho_2 = ... = \rho_N; \sum_i p_i D_{\mathrm{Tr}}(\rho_i, \sigma) = 0$, with D_{Tr} the trace distance;
- **Case** B: For any d-dimensional state σ it holds $\sum_i p_i D_{\mathrm{Tr}}(\rho_i, \sigma) > \epsilon$.

Where D_{Tr} is the trace distance. The goal is to find M such that there is a two-outcome test using M copies of ρ, which can return either `"accept"` or `"reject"`, and which accepts (that is, it returns the outcome `"accept"`) with probability larger than $2/3$ in case A, and it accepts with probability smaller than $1/3$ in case B. We prove the following results:

Theorem 1.1. Given access to $O(\frac{\sqrt{N}d}{\epsilon^2})$ samples of the density matrix ρ of Eq. (1), there is an algorithm which can distinguish with high probability whether $\sum_i p_i D_{\mathrm{Tr}}(\rho_i, \sigma) > \epsilon$ for every state σ, or there exists a state σ such that $\sum_i p_i D_{\mathrm{Tr}}(\rho_i, \sigma) = 0$ (that is, all the states ρ_i are equal).

Moreover, the observable used in the test of the theorem can be implemented efficiently by nested weak Schur sampling.

Theorem 1.2. Any algorithm which can distinguish with high probability whether $\sum_i p_i D_{\mathrm{Tr}}(\rho_i, \sigma) > \epsilon$ for every state σ, or there exists a state σ such that $\sum_i p_i D_{\mathrm{Tr}}(\rho_i, \sigma) = 0$ (that is, all the states ρ_i are equal), given access to M copies of the density matrix ρ of Eq. (1), requires at least $M = \Omega(\frac{\sqrt{N}d}{\epsilon^2})$ copies.

References

[ARS88] Robert Alicki, Slawomir Rudnicki, and Slawomir Sadowski. Symmetry properties of product states for the system of N n -level atoms. *Journal of Mathematical Physics*, 29(5):1158–1162, may 1988.

[BCH06] Dave Bacon, Isaac L. Chuang, and Aram W. Harrow. Efficient Quantum Circuits for Schur and Clebsch-Gordan Transforms. *Physical Review Letters*, 97(17):170502, oct 2006.

[BCWdW01] Harry Buhrman, Richard Cleve, John Watrous, and Ronald de Wolf. Quantum Fingerprinting. *Physical Review Letters*, 87(16):167902, sep 2001.

[BIMT06] E. Bagan, S. Iblisdir, and R. Muñoz-Tapia. Relative states, quantum axes, and quantum references. *Physical Review A*, 73(2):022341, feb 2006.

[BOW19] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, New York, NY, USA, jun 2019. ACM.

[BRS04] Stephen D. Bartlett, Terry Rudolph, and Robert W. Spekkens. Optimal measurements for relative quantum information. *Physical Review A*, 70(3):032321, sep 2004.

[Can20] Clement L. Canonne. A Survey on Distribution Testing: Your Data is Big. But is it Blue? *Theory of Computing*, 1(1):1–100, 2020.

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, Philadelphia, PA, jan 2014. Society for Industrial and Applied Mathematics.

[CM06] Matthias Christandl and Graeme Mitchison. The Spectra of Quantum States and the Kronecker Coefficients of the Symmetric Group. *Communications in Mathematical Physics*, 261(3):789–797, feb 2006.

[DK16] Ilias Diakonikolas and Daniel M. Kane. A New Approach for Testing Properties of Discrete Distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, oct 2016.

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1841–1854, Philadelphia, PA, oct 2015. Society for Industrial and Applied Mathematics.

[FRS+20] Marco Fanizza, Matteo. Rosati, Michalis Skotiniotis, John Calsamiglia, and Vittorio Giovannetti. Beyond the Swap Test: Optimal Estimation of Quantum State Overlap. *Physical Review Letters*, 124(6):060503, feb 2020.

[GI06] N. Gisin and S. Iblisdir. Quantum relative states. *The European Physical Journal D*, 39(2):321–327, aug 2006.

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, nov 2017.

[Har05] Aram W. Harrow. *Ph.D. thesis. arXiv preprint: quant-ph/0512255.* PhD thesis, MIT, Cambridge, 2005.

[Hay16] Masahito Hayashi. *Quantum Information Theory: Mathematical Foundation*. Springer, 2016.

[HHJ+17] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-Optimal Tomography of Quantum States. *IEEE Transactions on Information Theory*, 63(9):5628–5641, sep 2017.

[HM02] Masahito Hayashi and Keiji Matsumoto. Quantum universal variable-length source coding. *Physical Review A*, 66(2):022311, aug 2002.

[Key06] M. Keyl. Quantum state estimation and large deviations. *Reviews in Mathematical Physics*, 2006.

[Kro19] Hari Krovi. An efficient high dimensional quantum Schur transform. *Quantum*, 3:122, feb 2019.

[KW01] M. Keyl and R. F. Werner. Estimating the spectrum of a density operator. *Physical Review A*, 64(5):052311, oct 2001.

[LR06] Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[LRR13] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(1):295–347, 2013.

[LSB06] Netanel H. Lindner, Petra F. Scudo, and Dagmar Bruss. Quantum estimation of relative information. *International Journal of Quantum Information*, 04(01):131–149, feb 2006.

[MdW16] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *Theory of Computing*, 2016.

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2015.

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, volume 19-21-June, pages 899–912. Association for Computing Machinery, jun 2016.

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography II. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2017.

[Yu19] Nengkun Yu. Quantum closeness testing: A streaming algorithm and applications, 2019.

Complexity of Quantum Support Vector Machines and Quantum Neural Networks

Arne Thomsen, David Sutter, Amira Abbas, and Stefan Woerner

Full version: Master's Thesis "Comparing Quantum Neural Networks and Quantum Support Vector Machines" by Arne Thomsen. Please contact the authors if interested.

Abstract

We prove a polynomial speedup compared to [1] for training of noisy quantum support vector machines via the dual optimization problem. We introduce the PEGASOS algorithm [2] as an alternative and derive bounds on its runtime, which scales favorably. In addition, we analyze quantum neural networks numerically from the same perspective.

At the heart of *quantum support vector machines* (QSVMs) [3, 1] and *quantum neural networks* (QNNs) [3, 4] are quantum expectation values. By Born's rule, even on an ideal fault-tolerant quantum computer, these can fundamentally only be determined approximately for any finite number of measurement shots, which invariably introduces statistical uncertainty into the algorithms. In this work, the computational complexity for the models of both training and prediction is analyzed taking this into account.

The machine learning task under consideration is *binary supervised classification*. There, the learner is given a training set $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_M, y_M)\}$ with elements in $\mathbb{R}^s \times \{+1, -1\}$ (data and labels) and a test set $S = \{\mathbf{x}'_1, \mathbf{x}'_2 ..., \mathbf{x}'_m\}$ with elements in $\mathbb{R}^s$, where the labels are unknown during the training procedure. The goal is to accurately predict the labels of the elements in the test set.

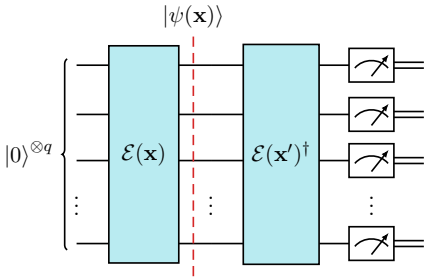

Figure 1: Quantum kernel circuit.

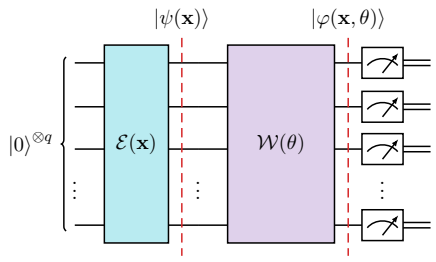

Figure 2: Quantum neural network.

Quantum support vector machines

A standard approach to do binary supervised classification is via a support vector machine (SVM). The sole difference between a QSVM and an SVM is that the so-called *kernel function* is evaluated on a quantum computer in the former case. More precisely, consider a kernel

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{tr}\left[|\psi(\mathbf{x}')\rangle\langle\psi(\mathbf{x}')|\ |\psi(\mathbf{x})\rangle\langle\psi(\mathbf{x})|\right] = |\langle\psi(\mathbf{x}')|\psi(\mathbf{x})\rangle|^2 = |\langle 0|\mathcal{E}(\mathbf{x}')^\dagger\mathcal{E}(\mathbf{x})|0\rangle|^2,$$

that can be computed via the circuit given in Figure 1. Note that whenever the mapping $\mathcal{E}$ is such that it can be implemented efficiently on a quantum computer but not on any classical device (such as IQP [5] or Forrelation [6] type circuits), we can evaluate $k(\mathbf{x}, \mathbf{x}')$ with a quantum speedup. Because of the stochastic nature of quantum mechanics, and since we can only run a finite number of measurement shots, the kernel can only be evaluated approximately. This is a fundamental difference to classical kernels.

Dual optimization problem The so-called *dual optimization problem* in QSVMs denotes the following convex optimization problem

$$\begin{aligned} \underset{\alpha_i \in \mathbb{R}}{\text{maximize}} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \, k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2}\sum_i \frac{\alpha_i^2}{C} \\ \text{subject to} \quad & 0 \leq \alpha_i \; \forall i. \end{aligned} \tag{1}$$

whose solution defines the fully trained classifier. In Table 1, we state the overall computational cost for solving (1) such that the resulting classification function is ε close with high probability to the ideal one stemming from exact kernel evaluations. We note that our bound improves on the best previously known results for the same setting scaling as $\mathcal{O}\left(M^6/\varepsilon^2\right)$ [1].

Pegasos Algorithm Alternatively to (1), the primal optimization problem

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^s,\, b, \xi_i \in \mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \; \forall i \, . \end{aligned} \tag{2}$$

can be approximately solved with the kernelized Pegasos algorithm [2], which is a form of stochastic sub-gradient descent. This work marks the first usage of the Pegasos algorithm for QSVMs and we display its overall runtime in Table 1. Unlike in an entirely classical setting (where Pegasos was introduced), we adapt the algorithm to noisy kernels.

Quantum neural networks

An alternative to QSVMs are QNNs that are defined like the *variational quantum classifiers* in [3] as the succession of a feature map circuit $\mathcal{E}(\mathbf{x})$ and a variational circuit $\mathcal{W}(\theta)$. In that case, the circuit parameters that are fixed by the input data $\mathbf{x}$ and trainable ones denoted by θ are strictly separated. See Figure 2 for a generic circuit. In Table 1, we highlight the computational cost for training and prediction via QNNs.

Results

Table 1 summarizes the computational complexity for training and prediction using QSVMs or QNNs. We observe that QNNs and for QSVMs Pegasos are particularly competitive.

	QSVM (dual)	QSVM (Pegasos)	QNN
training	$\mathcal{O}\left(M^{4.67}/\varepsilon^2\right)$	$\mathcal{O}\left(\min\left[M^2/\delta^3,\, 1/\delta^5\right]\right)$	$\mathcal{O}\left(d/\varepsilon^n\right)$
prediction	$\mathcal{O}\left(S^2/\varepsilon^2\right)$	$\mathcal{O}\left(S^2/\varepsilon^2\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

Table 1: Asymptotic complexity of the number of ideal quantum circuit evaluations necessary for training and prediction. M is the size of the training set, S the number of support vectors, d the number of trainable parameters, $n = 3$ conjectured, ε a bound on the difference between the ideal and noisy classification function that holds with probability $> \frac{1}{2}$ and δ the accuracy with which (2) is approximated.

References

[1] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. "A rigorous and robust quantum speed-up in supervised machine learning." In: *Nature Physics* (2021). ISSN: 1745-2481. DOI: 10.1038/s41567-021-01287-z. URL: https://doi.org/10.1038/s41567-021-01287-z.

[2] Shai Shalev-Shwartz et al. "Pegasos: Primal estimated sub-gradient solver for SVM." In: *Mathematical Programming* 127.1 (2011), pp. 3–30. ISSN: 00255610. DOI: 10.1007/s10107-010-0420-4.

[3] Vojtěch Havlíček et al. "Supervised learning with quantum-enhanced feature spaces." In: *Nature* 567.7747 (2019), pp. 209–212. ISSN: 14764687. DOI: 10.1038/s41586-019-0980-2. arXiv: 1804.11326. URL: http://dx.doi.org/10.1038/s41586-019-0980-2.

[4] Amira Abbas et al. "The power of quantum neural networks." In: *Nature Computational Science* 1.June (2020). ISSN: 2662-8457. DOI: 10.1038/s43588-021-00084-1. arXiv: 2011.00027. URL: http://arxiv.org/abs/2011.00027{\%}0Ahttp://dx.doi.org/10.1038/s43588-021-00084-1.

[5] Dan Shepherd and Michael J. Bremner. "Temporally unstructured quantum computation." In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 465.2105 (2009), pp. 1413–1439. DOI: 10.1098/rspa.2008.0443. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2008.0443. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2008.0443.

[6] Scott Aaronson and Andris Ambainis. "Forrelation: A Problem That Optimally Separates Quantum from Classical Computing." In: *SIAM Journal on Computing* 47.3 (2018), pp. 982–1038. DOI: 10.1137/15M1050902. eprint: https://doi.org/10.1137/15M1050902. URL: https://doi.org/10.1137/15M1050902.

Linear Regression by Quantum Amplitude Estimation and its Extension to Convex Optimization

Kazuya Kaneko,[1] Koichi Miyamoto,[2,1,*] Naoyuki Takeda,[1] and Kazuyoshi Yoshino[1]

[1]*Mizuho-DL Financial Technology Co., Ltd., Tokyo, Japan*
[2]*Center for Quantum Information and Quantum Biology, Osaka University, Osaka, Japan*

(Dated: October 19, 2021)

We propose a quantum algorithm for linear regression based on quantum amplitude estimation. This outputs regression coefficients as classical data with complexity depending on the number of data logarithmically and on the tolerance ϵ as $O(\epsilon^{-1})$, in contrast to $O(\epsilon^{-2})$ in existing methods. Additionally, we generalize it for convex optimization.

(The full version of this extended abstract is available at [1].)

Linear regression is a basic tool for natural and social sciences. This can be described as follows: given N_D data points $\{(\vec{x}_i, y_i)\}_{i=1,...,N_D}$, each of which consists of a vector of d explanatory variables $\vec{x}_i = (x_i^{(1)}, ..., x_i^{(d)})^T \in \mathbb{R}^d$ and an objective variable $y_i \in \mathbb{R}$, find an approximation of y_i as a linear function of $\vec{x}_i$, that is, $y_i \approx \vec{a} \cdot \vec{x}_i$ with some *regression coefficients* $\vec{a} \in \mathbb{R}^d$. More strictly, we find $\vec{a}$ which minimize $\left\|\vec{y} - X\vec{a}\right\|^2$, where $\|\cdot\|$ is the Euclidean norm, $X := (\vec{x}_1, ..., \vec{x}_{N_D})^T$ is a $N_D \times d$ matrix called the *design matrix*, and $\vec{y} := (y_1, ..., y_{N_D})^T$. Such $\vec{a}$ is given by $\vec{a} = W^{-1}\vec{z}$, where $W := \frac{1}{N_D}X^T X$ and $\vec{z} := \frac{1}{N_D}X^T\vec{y}$. Therefore, we obtain $\vec{a}$ by classically calculating elements of W and $\vec{z}$ as

$$w_{ij} = \frac{1}{N_D}\sum_{k=1}^{N_D} x_k^{(i)} x_k^{(j)}, z_i = \frac{1}{N_D}\sum_{k=1}^{N_D} x_k^{(i)} y_k, \tag{1}$$

respectively, and then $W^{-1}\vec{z}$. We hereafter call this method the *naive classical method*. In the ordinary situation where $d \ll N_D$, the bottleneck part of this is calculating (1), which obviously takes $O(d^2 N_D)$ computational time.

On the other hand, there are some quantum algorithms for this with complexity depending on N_D as $O(\mathrm{polylog}(N_D))$ [2–7][1], which means the exponential speedup compared with the naive classical method. However, we should note that many of existing quantum methods create quantum states in which the values of the regression coefficients are encoded in the amplitudes of basis states. Therefore, if we obtain the coefficients as *classical data* from such a state, the estimated coefficients are accompanied by errors inevitably, and high-accuracy estimation leads to large complexity. The existing method for calculating the coefficients as classical data with the best complexity with respect to the error tolerance ϵ is [4], and its complexity is

$$O\left(\frac{d^{5/2}\kappa^3}{\epsilon^2}\mathrm{polylog}\left(\frac{d\kappa}{\epsilon}\right)\right), \tag{2}$$

where κ is the condition number of X. If N_D is intermediate, say $10^3 - 10^5$, and $\epsilon \lesssim 10^{-3}$, a naive classical method can have a smaller complexity than (2).

In [1], we present a new quantum algorithm for linear regression, focusing on reducing the order of ϵ^{-1} in the expression of the complexity. In our method, unlike existing methods, we do not perform all calculation on a quantum computer. Instead, we use a quantum computer *only to perform a bottleneck part in the naive classical method*. That is, we estimate the sums (1) by *quantum amplitude estimation (QAE)* [11–17], which is known for providing quantum speedups for some kinds of computation such as Monte Carlo integration [12, 18]. Then, we classically solve the d-dimensional system of linear equations $W\vec{a} = \vec{z}$, and obtain $\vec{a}$ as classical data.

For more strict discussion, let us make some assumptions.

Assumption 1. *We can use the oracles P_x and P_y such that, for any $i \in \{1, ..., d\}$ and $k \in \{1, ..., N_D\}$, $P_x : |i\rangle|k\rangle|0\rangle \mapsto |i\rangle|k\rangle|x_k^{(i)}\rangle$ and $P_y : |k\rangle|0\rangle \mapsto |k\rangle|y_k\rangle$.*

Assumption 2. *The design matrix X is full-rank.*

Assumption 3. $0 \le x_k^{(i)} \le 1, 0 \le y_k \le 1$ *for any $i \in \{1, ..., d\}$ and $k \in \{1, ..., N_D\}$.*

Assumption 4. *There exists a positive real number c, which is independent of N_D, ϵ, κ and d, such that $w_{ii} = \frac{1}{N_D}\sum_{k=1}^{N_D}\left(x_k^{(i)}\right)^2 > c$ for any $i \in \{1, ..., d\}$.*

---

* koichi.miyamoto@qiqb.osaka-u.ac.jp

[1] There are also quantum-inspired classical methods [8–10].

Assumption 3 and 4 are seemingly strong but actually natural, since, for successful regression, we should know "typical scales" of explanatory and objective variables for preprocessings such as outlier handling, and then we can use them to rescale the variables. Then, denoting the max norm as $\|\cdot\|_\infty$, we have the following theorem.

Theorem 1. *Let ϵ be a given positive number. Under Assumption 2 to 4, there is a quantum algorithm that makes*

$$O\left(\max\left\{\frac{d^{3/2}\kappa^4}{\epsilon}, d\kappa^2\right\} \times d^2 \log(d)\right) \tag{3}$$

uses of P_x and

$$O\left(\max\left\{\frac{d^{3/2}\kappa^4}{\epsilon}, d\kappa^2\right\} \times d \log(d)\right) \tag{4}$$

uses of P_y, and outputs $\vec{a} \in \mathbb{R}^d$ such that $\|\vec{a} - W^{-1}\vec{z}\|_\infty = O(\epsilon)$ with a probability larger than 99%.

Besides, inspired by the above linear regression algorithm, we propose a quantum algorithm for some class of convex optimization. Suppose that we want to find the minimum point $\vec{a}^\star \in \mathbb{R}^d$ of the following objective function

$$F(\vec{a}) = \frac{1}{N_D}\sum_{k=1}^{N_D} f(\vec{a}, \vec{c}_k). \tag{5}$$

Here, f is a real-valued twice-differentiable function, which are shared by all the terms. Its inputs are the optimization variables $\vec{a} \in \mathbb{R}^d$ and some parameters $\vec{c}_k$, which are different in each term. We can see that linear regression falls into this type of problem. For such an objective function, the gradient $\vec{g}_F(\vec{a}) = (g_{F,1}(\vec{a}), ..., g_{F,d}(\vec{a}))^T$ and the Hessian $H_F(\vec{a}) = (h_{F,ij}(\vec{a}))_{\substack{1\le i\le d\\ 1\le j\le d}}$ are given as

$$g_{F,i}(\vec{a}) = \frac{\partial F}{\partial a_i}(\vec{a}) = \frac{1}{N_D}\sum_{k=1}^{N_D}\frac{\partial f}{\partial a_i}(\vec{a}, \vec{c}_k),\ h_{F,ij}(\vec{a}) = \frac{\partial^2 F}{\partial a_i \partial a_i}(\vec{a}) = \frac{1}{N_D}\sum_{k=1}^{N_D}\frac{\partial^2 f}{\partial a_i \partial a_j}(\vec{a}, \vec{c}_k). \tag{6}$$

Since these are the sums of many terms, each of which is same function with different inputs, we can estimate (6) by QAE with a complexity depending on N_D logarithmically and on the error tolerance ϵ as $O(\epsilon^{-1})$, if the following oracles are available

$$P_c : |k\rangle|0\rangle \mapsto |k\rangle|\vec{c}_k\rangle,\ P_i : |\vec{a}\rangle|\vec{c}_k\rangle|0\rangle \mapsto |\vec{a}\rangle|\vec{c}_k\rangle\left|\frac{\partial f}{\partial a_i}(\vec{a}, \vec{c}_k)\right\rangle,\ P_{ij} : |\vec{a}\rangle|\vec{c}_k\rangle|0\rangle \mapsto |\vec{a}\rangle|\vec{c}_k\rangle\left|\frac{\partial^2 f}{\partial a_i \partial a_j}(\vec{a}, \vec{c}_k)\right\rangle. \tag{7}$$

Then, like Newton's method, we update $\vec{a}$ by

$$\vec{a} \leftarrow \vec{a} - \hat{H}_F^{-1}(\vec{a})\vec{\hat{g}}_F(\vec{a}), \tag{8}$$

where $\vec{\hat{g}}_F$ and $\hat{H}_F$ are estimates of $\vec{g}_F$ and H_F by QAE, respectively. For this *QAE-based Newton's method*, we have the following.

Theorem 2. *Assume that $F(\vec{a})$ is twice-differentiable and μ-strongly convex, and that $H_F(\vec{a})$ is M-Lipschitz. Then, for any $\epsilon \in (0, \mu/2M)$ and $\vec{a}_0 \in \mathbb{R}^d$ satisfying $\delta_0 := \|\vec{a}_0 - \vec{a}^\star\| < \mu/M$, the QAE-based Newton's method with initial value $\vec{a} = \vec{a}_0$ outputs $\vec{\hat{a}} \in \mathbb{R}^d$ such that $\left\|\vec{\hat{a}} - \vec{a}^\star\right\|_\infty \le 2\epsilon$ by n_{it}-times updates, where*

$$n_{\mathrm{it}} := \max\left\{\left\lceil \log_2\left(\frac{\log\left(\frac{2M\epsilon}{\mu}\right)}{2\log\left(\frac{M\delta_0}{\mu}\right)}\right)\right\rceil + 1, 1\right\}, \tag{9}$$

with a success probability higher than 99%. In the process, the total numbers of calls to $\{P_i\}_{i=1,...,d}$, $\{P_{ij}\}_{i,j=1,...,d}$ and P_c are

$$N_{1\mathrm{stDer}} = O\left(\frac{d^{3/2}}{\mu\epsilon} n_{\mathrm{it}} \log(n_{\mathrm{it}} d^2)\right),\ N_{2\mathrm{ndDer}} = O\left(\frac{\tilde{\delta}_0 d^3}{\mu\epsilon} n_{\mathrm{it}} \log(n_{\mathrm{it}} d^2)\right),\ N_c = N_{1\mathrm{stDer}} + N_{2\mathrm{ndDer}}, \tag{10}$$

respectively.

Acknowledgment — This work was supported by MEXT Quantum Leap Flagship Program (MEXT Q-LEAP) Grant Number JPMXS0120319794.

[1] K. Kaneko et al., "Linear regression by quantum amplitude estimation and its extension to convex optimization", Phys. Rev. A 104, 022430 (2021)

[2] N. Wiebe et al., “Quantum Data Fitting”, Phys. Rev. Lett. 109, 050505 (2012)
[3] M. Schuld et al., “Prediction by linear regression on a quantum computer”, Phys. Rev. A 94, 022342 (2016)
[4] G. Wang, “Quantum Algorithm for Linear Regression”, Phys. Rev. A 96, 012335 (2017)
[5] C.-H. Yu et al., “Quantum algorithms for ridge regression”, IEEE Transactions on Knowledge and Data Engineering 29, 37491 (2019)
[6] S. Chakraborty, “The power of block-encoded matrix powers: improved regression techniques via faster Hamiltonian simulation”, Proceedings of the 46th International Colloquium on Automata, Languages, and Programming (ICALP), pp. 33:1-33:14 (2019)
[7] I. Kerenidis and A. Prakash, “Quantum gradient descent for linear systems and least squares”, Phys. Rev. A 101, 022316 (2020)
[8] N.-H. Chia, et al., “Sampling-based sublinear low-rank matrix arithmetic framework for dequantizing quantum machine learning”, Proceedings of the 52nd ACM Symposium on the Theory of Computing (STOC), 387 (2020)
[9] A. Gilyen et al., “An improved quantum-inspired algorithm for linear regression”, arXiv:2009.07268
[10] C. Shao and A. Montanaro, “Faster quantum-inspired algorithms for solving linear systems”, arXiv:2103.10309
[11] G. Brassard et. al., “Quantum amplitude amplification and estimation”, Contemporary Mathematics, 305, 53 (2002)
[12] Y. Suzuki et. al., “Amplitude Estimation without Phase Estimation”, Quantum Information Processing, 19, 75 (2020)
[13] S. Aaronson and P. Rall, “Quantum approximate counting, simplified”, Symposium on Simplicity in Algorithms”, 24-32, SIAM (2020)
[14] D. Grinko et al., “Iterative quantum amplitude estimation”, arXiv:1912.05559
[15] K. Nakaji, “Faster Amplitude Estimation”, arXiv:2003.02417
[16] E. G. Brown et al., “Quantum Amplitude Estimation in the Presence of Noise”, arXiv:2006.14145
[17] T. Tanaka, et al., “Amplitude estimation via maximum likelihood on noisy quantum computer”, arXiv:2006.16223
[18] A. Montanaro, “Quantum speedup of Monte Carlo methods”, Proc. Roy. Soc. Ser. A, 471, 2181 (2015)