Extended abstract: Quantum Policy Gradient in Reproducing Kernel Hilbert Space

David M. Bossens Kishor Bharti Jayne Thompson

Reinforcement learning (RL) is a technique that is successful across a wide range of interactive applications. A key limitation of RL is that it requires a large number of samples before a high-performing policy is learned. With the aim of reducing the sample complexity, several works have proposed applying RL systems within quantum-accessible environments, where interactions with the environment occur within a quantum system allowing to make use of superpositions across state-action trajectories. While exponential sample complexity improvements have only been shown for a special case environment formulated around Simon's problem [1], recent policy gradient algorithms demonstrate benefits in terms of quadratic sample complexity improvements when applying a parametrised quantum circuit within a quantum environment due to the properties of quantum superpositions [2]. Moreover, several quantum RL works demonstrate that by using parametrised quantum circuits (PQCs), the number of parameters can be reduced compared to using classical neural networks [3, 4] – although this line of work focused primarily on classical environments.

Despite the promise of quadratic or better improvements, limited work has been done in quantum-accessible environments, and especially in the construction of suitable PQCs. Previous work has introduced various PQCs for classical RL [5], using hardware efficient PQCs with alternating layered architecture. Two of these circuit classes, namely Raw-PQC and Softmax-PQC, have then been analysed further in the context of quantum-accessible environments when using quantum policy gradient (QPG) algorithms [2], which yield quadratic improvements in query complexity, i.e. the calls to the oracle to estimate the policy gradient.

Due to quantum states residing in a high-dimensional complex Hilbert space, PQCs have a natural interpretation in terms of kernel methods. While so far, this property has been discussed widely for supervised learning [6, 7], this has not yet been adopted in RL.

Our work is inspired by streams of work in classical RL that use kernel-based formulations of the policy [8, 9]. We formulate Gaussian and softmax policies based on quantum kernels and analyse their efficiency across various optimisation schemes with quantum policy gradient algorithms. While maintaining quadratic speedups associated with QPG, the use of quantum kernels for the policy definition leads to advantages such as analytically available policy gradients, tunable expressiveness, and techniques for sparse non-parametric representations of the policy within the context of vector-valued state and action spaces. This also leads to a quantum actor-critic algorithm with an interpretation as performing a natural gradient. Unlike quantum algorithms for natural policy gradient [10, 11], the proposed algorithm is formulated within the kernel method framework and is tailored to the quantum accessible environment where it can exploit a quadratic sample complexity improvement as well as a variance reduction as is often associated with actor-critic RL.

Using quantum kernels for reinforcement learning policies.– Kernel methods have strong theoretical foundations for functional analysis and supervised learning (see e.g. [12] for an overview). We review some of these useful properties here and how they can be applied to formulate and learn efficient policies for quantum RL.

First, the kernel function determines the expressible function space through its reproducing kernel Hilbert space (RKHS). The choice of the kernel function thereby provides an opportunity to balance the expressiveness, training efficiency, and generalisation. For instance, reducing the bandwidth factor to $c < 1$ of the squared cosine kernel

$$\kappa(s, s') = \prod_{j=1}^{d} \cos^2(c(s_j - s'_j)/2) \tag{1}$$

restricts features to parts of the Bloch sphere, allowing improved generalisation [13]. As this affects the coverage of the Bloch sphere, this tuning factor can be related to expressiveness measures based on the Haar distribution, such as the frame potential or the KL difference [14]. Tuning a single parameter is significantly more convenient compared to redesigning the ansatz of a PQC.

Second, kernel functions inherently define a particular feature-map. This interpretation follows from Mercer's theorem, which states that every kernel function can be written as

$$\kappa(s, s') = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_i(s') \,,$$

where for all i, e_i is an eigenfunction such that $\lambda_i e_i(s') = \mathscr{T}_K[e_i](s') = \int_a^b \kappa(s, s') e_i(s) ds$ with eigenvalue λ_i. Mercer's

theorem leads to the kernel trick,

$$\kappa(s, s') = \langle \phi(s), \phi(s') \rangle \,,$$

which allows writing the kernel function as an inner product based on a feature-map ϕ. For quantum kernels, this allows a definition of kernels in terms of the data encoding as a feature-map. For instance, the basis encoding corresponds to the Kronecker delta kernel, the amplitude encoding corresponds to the inner product quantum kernel, etc. [7]. We construct kernel-based policies which construct kernel computations in quantum circuits both in the explicit view and the implicit (i.e. inner product) view.

Third, the kernel based formalism is based on a representative set of state-action pairs. In a supervised learning setting, the representer theorem guarantees that the optimal function approximator can be written as linear combination of kernel evaluations based on input data pairs, which leads to the formulations of support vector machines and kernel regression. In quantum supervised learning, one uses this property to evaluate the kernel in a quantum device and the model is then computed from the kernel in a quantum or a classical device [6, 15]. In a quantum RL setting, we analogously consider the optimal deterministic policy μ as a linear combination of kernel computations with regard to a select subset of the states:

$$\mu(s) = \sum_{i=1}^{N} \beta_i \kappa(s, s_i) \,. \tag{2}$$

Using this quantity as the mean of a Gaussian distribution allows the quantum analogue of the Gaussian policies that are popular in classical RL with vector-valued action spaces. We show this comes with analytical forms for the gradient and is suitable for various non-parametric optimisation schemes. We will formulate a special class of PQCs which form circuits with Eq. 2 as their expectation, allowing a novel way to form expressive and coherent policies in the context of quantum environments.

Fourth, with the aim of performing ℓ_2-regularisation in the context of kernel ridge regression, we make use of the result that for every RKHS $\mathscr{H}_K$ with reproducing kernel K, and any $g \in \mathscr{H}_K$,

$$||g||^2_{\mathscr{H}_K} = \langle g, g \rangle_{\mathscr{H}_K} = \int (\mathscr{R} g(x))^2 dx$$

where the operator $\mathscr{R} : \mathscr{H}_K \to \mathscr{D}$ can be interpreted as extracting information from the function value which gets penalised during optimisation [12]. For instance, it can penalise large higher or lower order derivatives, large function values, or still other properties, leading to smoother optimisation landscapes and therefore improved query complexity and convergence to the global optimum. This property contributes to an improved query complexity when considering actor-critic algorithms with smooth critic functions and can also be exploited when directly optimising the kernel.

Overview of the contributions.– This work contributes the following theoretical results to the field of quantum RL:

- We propose two classes of quantum kernel policies (QKPs) for learning in quantum-accessible environments. First, we propose Representer PQCs, which incorporate representer theorem based formalisms directly within a quantum circuit and which are suitable for both analytical and numerical gradient based optimisation. Second, we propose Gaussian kernel-based policies based on a classically known mean function and covariance, which due to the mean and covariance being parametrised classically has known analytical gradient, thereby removing the need for expensive estimation procedures required for analytical quantum policy gradient with traditional PQCs. We also provide a formula to scale the number of representers based on a Lipschitz constant.
- We use a central differencing approach on phase oracles of the value function for a numerical quantum policy gradient algorithm [2] based on Representer PQCs. We report a query complexity comparable to Jerbi et al. [2] but note the potentially lower number of parameters.
- We use analytical quantum policy gradient algorithms which perform quantum multivariate Monte Carlo on binary oracles of the policy gradient. We first confirm a query complexity that is comparable to Jerbi et al. [2].
- We propose two further improvements in an algorithm we call Compatible Quantum RKHS Actor-Critic (CQRAC). First, the parameter dimensionality of the policy is reduced by using vector-valued kernel matching pursuit. Second, we formulate a quantum oracle, which we call the state-action occupancy oracle, which computes the policy gradient samples based on the critic's prediction on a particular state-action pair rather than on the cumulative reward of the trajectory, thereby reducing the variance of the estimate produced by analytical quantum policy gradient. Theorem 6.2**a** (see main text) demonstrates that the resulting query complexity depends on the maximal deviation from a baseline estimate, rather than on the maximal cumulative reward. Theorem 6.2**b** (see main text) provides an improved result which exploits an upper bound on the variance of the gradient of the log-policy, and thereby demonstrates how smooth policies such as the Gaussian kernel-based policy can give additional query complexity benefits.

TABLE I: Query complexity of policy gradient estimation with PQCs. General notations: $\mathscr{A}$ is the action space; $r_{\max}$ denotes the maximal absolute reward; T is the horizon; d is the parameter dimensionality; γ is the discount factor; and ϵ is the tolerance for error in the gradient estimate. Specific notations: $\mathscr{T}$ is the temperature of the softmax; D is an upper bound on higher-order derivatives of the policy; Δ_Q is the maximal absolute deviation of the critic's prediction to a baseline estimate; N is the number of representers in a kernel policy; σ_Q is an upper bound on the standard deviation of the critic's prediction to the baseline estimate; upper bounds on p-norms are denoted as B_p for the gradient of the log-policy, as σ_{∇_p} for the standard deviation of the partial derivative of the log-policy, as $\kappa_p^{\max}$ for the kernel computations across policy centres, and as C_p for the gradient of the critic w.r.t. actions; and $\xi(p) = \max\{0, 1/2 - 1/p\}$ is used for converting across p-norms.

Algorithm	**Oracle and estimation**	**Query complexity**
1. Policy gradient with Softmax-PQC [17]	Return oracle, single-qubit parameter shift rule [18], and classical Monte Carlo	$\tilde{\mathcal{O}}\left(\frac{\mathscr{T}^2 r_{\max}^2 T^2}{\epsilon^2(1-\gamma)^2}\right)$
2. Numerical QPG and Raw-PQC [2]	Return oracle, quantum gradient estimation via central differencing [19]	$\tilde{\mathcal{O}}\left(\sqrt{d}\frac{DTr_{\max}}{\epsilon(1-\gamma)}\right)$
3. Analytical QPG and Softmax-PQC [2]	Analytical gradient oracle, bounded quantum multivariate Monte Carlo (Theorem 3.3 [20])	$\tilde{O}\left(d^{\xi(p)}\frac{B_p T r_{\max}}{\epsilon(1-\gamma)}\right)$
Proposed: Numerical QPG in RKHS	cf.2	cf.2 but $d = NA$
Proposed: Analytical QPG in RKHS	cf.3	cf.3 but $d = NA$
Proposed: CQRAC	Analytical gradient oracle, near-optimal quantum multivariate Monte Carlo (Theorem 3.4 [20])	$\tilde{\mathcal{O}}\left(d^{\xi(p)}\frac{\Delta_Q B_p}{(1-\gamma)\epsilon}\right)$ for $d = NA$ $\tilde{\mathcal{O}}\left(\frac{d^{\xi(p)}\sigma_Q\sigma_{\nabla_p}}{(1-\gamma)\epsilon}\right)$ for $d = NA$
Proposed: DCQRAC	cf.3	$\tilde{\mathcal{O}}\left(d^{\xi(p)}\frac{\kappa_p^{\max} C_p}{(1-\gamma)\epsilon}\right)$ for $d = NA$

- We further propose Deterministic Compatible Quantum RKHS Actor-Critic (DCQRAC), which is based on the deterministic policy gradient theorem [16]. The approach makes use of similar formalisms as its non-deterministic counterpart, though with the key differences that it is based on state occupancy rather than state-action occupancy, and that the policy gradient takes a different form. Theorem 6.3 demonstrates that the resulting query complexity depends on the norm of kernel features and the gradient norm of the critic, which illustrates the importance of techniques such as kernel matching pursuit and regularisation.

Our query complexity results compare favourably to other policy gradients methods for PQCs (see Table I).

Why QTML.– This paper presents numeric and analytical optimisation techniques for quantum kernel policies for efficient RL in vector-valued action spaces. We prove quadratic improvements of kernel-based policy gradient and actor-critic algorithms over their classical counterparts, across different formulations of stochastic and deterministic kernel-based policies. Two actor-critic algorithms are proposed that improve on quantum policy gradient algorithms under favourable conditions. The proposed quantum kernel policies allow convenient analytical forms for the gradient and techniques for expressiveness control. Considering the significance for the quantum machine learning community, we believe our work is well-suited for QTML.

[1] Vedran Dunjko, Yi-Kai Liu, Xingyao Wu, and Jacob M. Taylor. Exponential improvements for quantum-accessible reinforcement learning. *arXiv preprint arXiv:1710.11160*, pages 1–27, 2017.

[2] Sofiene Jerbi, Arjan Cornelissen, Māris Ozols, and Vedran Dunjko. Quantum policy gradient algorithms. In *Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2023)*, pages 1–24, 2023.

[3] Qingfeng Lan. Variational quantum soft actor-critic. *arXiv preprint arXiv:2112.11921*, pages 1–8, 2021.

[4] Samuel Yen Chi Chen. Asynchronous training of quantum reinforcement learning. *Procedia Computer Science*, 222:321–330, 2023.

[5] Sofiene Jerbi, Casper Gyurik, Simon C Marshall, and Hans J Briegel. Parametrized Quantum Policies for Reinforcement Learning. In *Advances in Neural Information Processing (NeurIPS 2021)*, pages 1–14, 2021.

[6] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physics Review Letters*, 122:040504, Feb 2019.

[7] Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv preprint arxiv:2101.11020*, pages 1–25, 2021.

[8] Guy Lever and Ronnie Stafford. Modelling policies in MDPs in reproducing kernel Hilbert space. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*, volume 38, pages 590–598, 2015.

[9] J Andrew Bagnell and Jeff Schneider. Policy Search in Kernel Hilbert Space. *Tech. Rep. RI–TR-03–45*, 2003.

[10] Nico Meyer, Daniel D. Scherer, Axel Plinge, Christopher Mutschler, and Michael J. Hartmann. Quantum Natural Policy Gradients: Towards Sample-Efficient Reinforcement Learning. In *IEEE International Conference on Quantum Computing and Engineering (QCE 2023)*, volume 2, pages 36–41, 2023.

[11] André Sequeira, Luis Paulo Santos, and Luis Soares Barbosa. On Quantum Natural Policy Gradients. *arXiv preprint arXiv:2401.08307*, pages 1–14, 2024.

[12] Bernhard Schölkopf and Alexander J. Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2003.

[13] Abdulkadir Canatar, Evan Peters, Cengiz Pehlevan, Stefan M. Wild, and Ruslan Shaydulin. Bandwidth enables generalization in quantum kernel models. *Transactions on Machine Learning Research*, pages 1–31, 2022.

[14] Kouhei Nakaji and Naoki Yamamoto. Expressibility of the alternating layered ansatz for quantum computation. *Quantum*, 5:1–20, 2021.

[15] Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *Nature Communications*, 14(1):1–8, 2023.

[16] David Silver, Guy Lever, Nicholas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic Policy Gradient Algorithms. In *International Conference on Machine Learning (ICML 2014)*, pages 1–9, Bejing, China, 2014.

[17] André Sequeira, Luis Paulo Santos, and Luis Soares Barbosa. Policy gradients using variational quantum circuits. *Quantum Machine Intelligence*, 5(18):18, 2023.

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

[19] Arjan Cornelissen. Quantum gradient estimation of Gevrey functions. *arXiv preprint arXiv:1909.13528*, pages 1–48, 2019.

[20] Arjan Cornelissen, Yassine Hamoudi, and Sofiene Jerbi. Near-optimal quantum algorithms for multivariate mean estimation. In *Annual ACM SIGACT Symposium on Theory of Computing (STOC 2022)*, pages 33–43, New York, NY, USA, 2022. Association for Computing Machinery.