

Towards Surrogate Based Dequantization of Quantum Reinforcement Learning

Pablo Rodriguez-Grasa, Mikel Sanz & Ryan Sweke

1 Introduction

Given rapid recent progress in both quantum computing and machine learning, the field of quantum machine learning (QML) has emerged as an extremely active area of research. Among the many proposals for quantum-enhanced machine learning, variational quantum algorithms, which optimize parameterized quantum circuits (PQCs) to solve learning tasks, have attracted particular attention, primarily as a result of their near-term implementability on noisy intermediate-scale quantum (NISQ) devices [1–3]. However, despite significant research, the question of whether and to which extent such variational quantum algorithms can gain meaningful advantages over classical approaches remains a central open question.

A particularly active line of research aimed at addressing this question is that of *dequantization*, i.e. the construction of classical algorithms which can match the performance of their quantum counterparts with only polynomial overhead in computational resources [4–14]. For variational QML in particular, the *surrogate-based* approach to dequantization exploits the observation that PQC models are linear models with respect to a feature map defined by the data-encoding strategy of the circuit [15], and therefore lie within a reproducing kernel Hilbert space (RKHS) which can, under certain conditions, be optimized over efficiently using classical kernel methods [7–10]. To date, however, this approach has been studied only in the context of supervised learning [7–10] and generative modelling [11].

In parallel, there has been growing interest in the application of quantum algorithms to reinforcement learning (RL) [16]. In particular, quantum Q-learning replaces the classical neural network in deep Q-learning [17] with a PQC as a function approximator for the action-value function [18–25]. While empirical studies have reported promising results, and provable advantages can be obtained in artificial environments [20, 26], the theoretical foundations of quantum Q-learning remain largely unexplored. In particular, to the best of our knowledge, no prior work has addressed whether quantum Q-learning can be dequantized. This motivates the central question of this work:

What are sufficient conditions for surrogate-based dequantization of quantum Q-learning?

2 Setup: Quantum Q-learning with PQC Models

We consider a discounted Markov Decision Process $(S, \mathcal{A}, P, R, \gamma)$ with state space $S \subseteq [0, 1]^{d_S}$, finite action set $\mathcal{A}$, and discount factor $\gamma \in (0, 1)$. The optimal action-value function Q^* satisfies the Bellman optimality equation $Q^* = TQ^*$, where $(TQ)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)}[\max_{a'} Q(s', a')]$. We measure the quality of an approximate action-value function Q via its distance to the optimal action-value function, i.e. via the loss $\mathcal{L}(Q) = \mathbb{E}_{(s, a) \sim U(S \times \mathcal{A})}[|Q^*(s, a) - Q(s, a)|]$.

In quantum Q-learning, the Q -function is parameterized by a PQC model of the form

$$Q_\theta(s, a) = \langle 0 | U^\dagger(s, \theta) O_a U(s, \theta) | 0 \rangle, \quad (2.1)$$

for some set of observables $\{O_a\}$. Under mild assumptions, the above ansatz can be written in the form

$$Q_\theta(s, a) = \langle 0 | U^\dagger(s, \theta) V^\dagger(a) O V(a) U(s, \theta) | 0 \rangle, \quad (2.2)$$

where $U(s, \theta)$ and $V(a)$ use a *Hamiltonian data-encoding strategy* $\mathcal{D}$. Under this assumption, such models can be written as linear models $Q_\theta(z) = \langle c(\theta), \phi_{\mathcal{D}}(z) \rangle$ with respect to the feature map

$$\phi_{\mathcal{D}}(z) = (1, \cos(\langle \omega_1, z \rangle), \sin(\langle \omega_1, z \rangle), \dots, \cos(\langle \omega_M, z \rangle), \sin(\langle \omega_M, z \rangle)), \quad (2.3)$$

where $z = (s, a)$ and the frequency vectors $\{\omega_i\}_{i=1}^M \subset \mathbb{R}^{d_S + d_A}$ are determined entirely by the data-encoding strategy $\mathcal{D}$ [15]. Given any vector $w = (w_0, \dots, w_M)$ with all entries non-zero, one can define the normalized, shift-invariant PQC-inspired kernel

$$K_{(\mathcal{D}, w)}(z, z') = \frac{1}{\|w\|_2^2} \sum_{i=0}^M w_i^2 \cos(\langle \omega_i, z - z' \rangle), \quad (2.4)$$

with associated reproducing kernel Hilbert space $\mathcal{H}_{(\mathcal{D}, w)}$. By construction, the PQC model class $\mathcal{F}_{\Theta, \mathcal{D}, O}$ satisfies $\mathcal{F}_{\Theta, \mathcal{D}, O} \subset \mathcal{H}_{(\mathcal{D}, w)}$ for any such w , meaning the RKHS contains all functions expressible by the PQC.

3 Dequantization of quantum Q-learning

We say that a classical algorithm provides a *relative-to-best dequantization* of quantum Q-learning if, using polynomial time and samples, it outputs with high probability a policy π satisfying $\mathcal{L}(Q^\pi) \leq \mathcal{L}(Q_{\theta^*}) + \epsilon$, where Q_{θ^*} is the best Q-function achievable by the PQC model class. Motivated by the observation that the RKHS $\mathcal{H}_{(\mathcal{D}, w)}$ contains all functions expressible by the PQC model, the natural idea is to perform classical optimization over $\mathcal{H}_{(\mathcal{D}, w)}$ for some chosen w .

A simplifying assumption: uniform environment samples. In order to provide rigorous guarantees, we make the simplifying assumption that uniformly random state-action samples from $U(S \times \mathcal{A})$ are available. This models the regime of sampling from a large experience replay buffer after sufficient exploration, and is an assumption often made in the theoretical analysis of reinforcement learning, where it is referred to as the assumption of *access to a generative model* for the environment. In particular, this assumption has been used to provide the first rigorous analysis of standard deep Q-learning [27].

Under this assumption, our proposed classical dequantization algorithm is Fitted Q-Iteration (FQI) with Kernel Ridge Regression (KRR) via the PQC-inspired kernel $K_{(\mathcal{D}, w)}$, where w is a reweighting vector chosen by the classical learner. The algorithm mirrors quantum Q-learning's structure but replaces optimization over the PQC model class with KRR in the associated RKHS $\mathcal{H}_{(\mathcal{D}, w)}$.

Algorithm 1 Fitted Q-Iteration with PQC-inspired kernel $K_{(\mathcal{D}, w)}$

- 1: **Input:** Data encoding strategy $\mathcal{D}$, discount factor γ , iterations K , samples per iteration m , regularization $\lambda > 0$
 - 2: **Choose** Reweighting vector w defining the kernel $K_{(\mathcal{D}, w)}$
 - 3: Initialize $\hat{Q}_0 \equiv 0$
 - 4: **for** $k = 0, \dots, K - 1$ **do**
 - 5: Sample $\{(s_i, a_i)\}_{i=1}^m \sim U(S \times \mathcal{A})$; observe rewards r_i and next states s'_i
 - 6: Compute Bellman targets: $y_i \leftarrow r_i + \gamma \max_{a \in \mathcal{A}} \hat{Q}_k(s'_i, a)$
 - 7: Solve KRR: $F_{k+1} \leftarrow \arg \min_{f \in \mathcal{H}_{(\mathcal{D}, w)}} \left\{ \sum_{i=1}^m (y_i - f(z_i))^2 + \lambda^2 \|f\|_{\mathcal{H}_{(\mathcal{D}, w)}}^2 \right\}$
 - 8: Clip: $\hat{Q}_{k+1} \leftarrow \Pi_{[-V_{\max}, V_{\max}]} F_{k+1}$
 - 9: **Output:** Greedy policy $\pi_K(s) \in \arg \max_{a \in \mathcal{A}} \hat{Q}_K(s, a)$
-

4 Main Result

The main result of this work provides a set of sufficient conditions under which Algorithm 1 is both time and sample efficient, under the assumption of access to a uniform generative model for the environment. As such, given the relation between the RKHS $\mathcal{H}_{(\mathcal{D}, w)}$ and the class of PQC models $\mathcal{F}_{\Theta, \mathcal{D}, O}$, these conditions are also sufficient for relative-to-best dequantization of quantum Q-learning in the uniform sampling setting.

Theorem 4.1: (Sufficient conditions for efficient dequantization – informal) *Assume access to a uniform generative model for the environment. Under mild technical assumptions on the data-encoding strategy $\mathcal{D}$ and MDP structure, suppose the following conditions hold:*

1. *The reweighting vector w has polynomially decaying components.*
2. *The Bellman targets encountered during learning are well-aligned with w :*

$$\max_{1 \leq k \leq K} \|T\hat{Q}_{k-1}\|_{\mathcal{H}(\mathcal{D}, w)} = \mathcal{O}(\text{poly}(d_S, |\mathcal{A}|)). \quad (4.1)$$

3. *The weight vector w is induced from a symmetric matrix product state with polynomial bond-dimension.*

Then, with

$$K = \mathcal{O}\left(\log \frac{1}{(1-\gamma)^2 \epsilon}\right), \quad m = \tilde{\mathcal{O}}(\text{poly}(d_S, |\mathcal{A}|, 1/\epsilon, 1/\delta)), \quad (4.2)$$

Algorithm 1 yields, with probability at least $1 - \delta$, a policy π_K satisfying

$$\mathcal{L}(Q^{\pi_K}) \leq \mathcal{L}(Q_{\theta^*}) + \epsilon. \quad (4.3)$$

Moreover, the algorithm runs in time polynomial in m , K , and $|\mathcal{A}|$.

5 Interpretation and Implications

Theorem 4.1 identifies the regime in which quantum Q-learning can be efficiently dequantized, via sufficient conditions on the reweighting vector w that the classical learner (Alice) chooses when given the data-encoding strategy $\mathcal{D}$ of the quantum learner’s (Bob’s) PQC. We now discuss the role of each condition.

Polynomial eigenvalue decay. The first condition controls the intrinsic complexity of the kernel $K_{(\mathcal{D}, w)}$ and guarantees that the maximum information gain grows sublinearly in the number of samples, ensuring polynomial sample complexity. Alice can always choose w to satisfy this condition.

Efficient kernel evaluation via tensor networks. The PQC-inspired kernel $K_{(\mathcal{D}, w)}$ is defined via an inner product in an exponentially large feature space, making naive evaluation intractable. However, when w is induced from a symmetric matrix product state with polynomial bond-dimension [9], the kernel admits exact and efficient evaluation via tensor network contractions. Alice can always choose w with this structure.

Alignment. The most problem-dependent condition requires that the Bellman targets encountered during learning are well-aligned with the kernel reweighting. This captures the intuition that dequantization is possible when the problem structure is well matched to the inductive bias of the classical surrogate. Unlike the previous two conditions, this cannot be verified a priori without solving the problem.

Summary. The first two conditions are structural requirements that Alice can satisfy by design. The third is problem-dependent. When all three are met, no exponential quantum advantage via quantum Q-learning is possible in the uniform sampling setting. Conversely, our result identifies, by contrapositive, that a genuine quantum advantage would require either (i) a problem whose relevant frequency content is systematically misaligned with any efficiently computable kernel reweighting, or (ii) a PQC kernel that cannot be evaluated efficiently by any classical means.

We note that, despite the significantly different analysis required in the reinforcement learning setting, these conditions are very similar in spirit to the sufficient conditions identified for kernel-based dequantization of supervised learning [8–10]. We stress that while our results are proven only in the simplified setting of uniform state-action samples, they lay the theoretical groundwork for extending surrogate-based dequantization to reinforcement learning more broadly, and motivate future exploration of kernelized FQI as a practical heuristic in standard settings.

References

- [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”. In: *Nature Reviews Physics* 3.9 (2021), pp. 625–644. ISSN: 2522-5820. DOI: [10.1038/s42254-021-00348-9](https://doi.org/10.1038/s42254-021-00348-9). URL: <http://dx.doi.org/10.1038/s42254-021-00348-9> (page 1).
- [2] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. “Parameterized quantum circuits as machine learning models”. In: *Quantum Science and Technology* 4.4 (2019), p. 043001. DOI: [10.1088/2058-9565/ab4eb5](https://doi.org/10.1088/2058-9565/ab4eb5). URL: <https://doi.org/10.1088/2058-9565/ab4eb5> (page 1).
- [3] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermann Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. “Noisy intermediate-scale quantum algorithms”. In: *Rev. Mod. Phys.* 94 (1 2022), p. 015004. DOI: [10.1103/RevModPhys.94.015004](https://link.aps.org/doi/10.1103/RevModPhys.94.015004). URL: <https://link.aps.org/doi/10.1103/RevModPhys.94.015004> (page 1).
- [4] Ewin Tang. “A quantum-inspired classical algorithm for recommendation systems”. In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. STOC ’19. ACM, 2019, pp. 217–228. DOI: [10.1145/3313276.3316310](https://doi.org/10.1145/3313276.3316310). URL: <http://dx.doi.org/10.1145/3313276.3316310> (page 1).
- [5] Nai-Hui Chia, András Gilyén, Tongyang Li, Han-Hsuan Lin, Ewin Tang, and Chunhao Wang. “Sampling-based sublinear low-rank matrix arithmetic framework for dequantizing Quantum machine learning”. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2020. Chicago, IL, USA: Association for Computing Machinery, 2020, pp. 387–400. ISBN: 9781450369794. DOI: [10.1145/3357713.3384314](https://doi.org/10.1145/3357713.3384314). URL: <https://doi.org/10.1145/3357713.3384314> (page 1).
- [6] Ewin Tang. “Quantum Principal Component Analysis Only Achieves an Exponential Speedup Because of Its State Preparation Assumptions”. In: *Physical Review Letters* 127.6 (2021). ISSN: 1079-7114. DOI: [10.1103/PhysRevLett.127.060503](https://doi.org/10.1103/PhysRevLett.127.060503). URL: <http://dx.doi.org/10.1103/PhysRevLett.127.060503> (page 1).
- [7] Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. *Classically Approximating Variational Quantum Machine Learning with Random Fourier Features*. 2022. arXiv: [2210.13200](https://arxiv.org/abs/2210.13200) [quant-ph]. URL: <https://arxiv.org/abs/2210.13200> (page 1).
- [8] Ryan Sweke, Erik Recio-Armengol, Sofiene Jerbi, Elies Gil-Fuster, Bryce Fuller, Jens Eisert, and Johannes Jakob Meyer. “Potential and limitations of random Fourier features for dequantizing quantum machine learning”. In: *Quantum* 9 (2025), p. 1640. ISSN: 2521-327X. DOI: [10.22331/q-2025-02-20-1640](https://doi.org/10.22331/q-2025-02-20-1640). URL: <https://doi.org/10.22331/q-2025-02-20-1640> (pages 1, 3).
- [9] Ryan Sweke, Seongwook Shin, and Elies Gil-Fuster. *Kernel-based dequantization of variational QML without Random Fourier Features*. 2025. arXiv: [2503.23931](https://arxiv.org/abs/2503.23931) [quant-ph]. URL: <https://arxiv.org/abs/2503.23931> (pages 1, 3).
- [10] Mehrad Sahebi, Alice Barthe, Yudai Suzuki, Zoë Holmes, and Michele Grossi. *On Dequantization of Supervised Quantum Machine Learning via Random Fourier Features*. 2025. arXiv: [2505.15902](https://arxiv.org/abs/2505.15902) [quant-ph]. URL: <https://arxiv.org/abs/2505.15902> (pages 1, 3).
- [11] Mario Herrero-Gonzalez, Brian Coyle, Kieran McDowall, Ross Grassie, Sjoerd Beentjes, Ava Khamseh, and Elham Kashefi. *The Born Ultimatum: Conditions for Classical Surrogation of Quantum Generative Models with Correlators*. 2025. arXiv: [2511.01845](https://arxiv.org/abs/2511.01845) [quant-ph]. URL: <https://arxiv.org/abs/2511.01845> (page 1).
- [12] Sergi Masot-Llima, Elies Gil-Fuster, Carlos Bravo-Prieto, Jens Eisert, and Tommaso Guaita. *Prospects for quantum advantage in machine learning from the representability of functions*. 2025. arXiv: [2512.15661](https://arxiv.org/abs/2512.15661) [quant-ph]. URL: <https://arxiv.org/abs/2512.15661> (page 1).
- [13] Seongwook Shin, Yong Siah Teo, and Hyunseok Jeong. “Dequantizing quantum machine learning models using tensor networks”. In: *Phys. Rev. Res.* 6 (2 2024), p. 023218. DOI: [10.1103/PhysRevResearch.6.023218](https://link.aps.org/doi/10.1103/PhysRevResearch.6.023218). URL: <https://link.aps.org/doi/10.1103/PhysRevResearch.6.023218> (page 1).
- [14] Elies Gil-Fuster, Casper Gyurik, Adrian Perez-Salinas, and Vedran Dunjko. “On the Relation between Trainability and Dequantization of Variational Quantum Learning Models”. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: <https://openreview.net/forum?id=TdqaZbQvdi> (page 1).

- [15] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. “Effect of data encoding on the expressive power of variational quantum-machine-learning models”. In: *Phys. Rev. A* 103 (3 2021), p. 032430. DOI: [10.1103/PhysRevA.103.032430](https://doi.org/10.1103/PhysRevA.103.032430). URL: <https://link.aps.org/doi/10.1103/PhysRevA.103.032430> (pages 1, 2).
- [16] Nico Meyer, Christian Ufrecht, Maniraman Periyasamy, Daniel D. Scherer, Axel Plinge, and Christopher Mutschler. *A Survey on Quantum Reinforcement Learning*. 2024. arXiv: [2211.03464](https://arxiv.org/abs/2211.03464) [quant-ph]. URL: <https://arxiv.org/abs/2211.03464> (page 1).
- [17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmashan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (2015), pp. 529–533. ISSN: 1476-4687. DOI: [10.1038/nature14236](https://doi.org/10.1038/nature14236). URL: <https://doi.org/10.1038/nature14236> (page 1).
- [18] Samuel Yen-Chi Chen, Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, and Hsi-Sheng Goan. “Variational Quantum Circuits for Deep Reinforcement Learning”. In: *IEEE Access* 8 (2020), pp. 141007–141024. DOI: [10.1109/ACCESS.2020.3010470](https://doi.org/10.1109/ACCESS.2020.3010470) (page 1).
- [19] Owen Lockwood and Mei Si. “Reinforcement Learning with Quantum Variational Circuit”. In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 16.1 (2020), pp. 245–251. DOI: [10.1609/aiide.v16i1.7437](https://doi.org/10.1609/aiide.v16i1.7437). URL: <https://ojs.aaai.org/index.php/AIIDE/article/view/7437> (page 1).
- [20] Andrea Skolik, Sofiene Jerbi, and Vedran Dunjko. “Quantum agents in the Gym: a variational quantum algorithm for deep Q-learning”. In: *Quantum* 6 (2022), p. 720. ISSN: 2521-327X. DOI: [10.22331/q-2022-05-24-720](https://doi.org/10.22331/q-2022-05-24-720). URL: <https://doi.org/10.22331/q-2022-05-24-720> (page 1).
- [21] Owen Lockwood and Mei Si. “Playing Atari with Hybrid Quantum-Classical Reinforcement Learning”. In: *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*. Ed. by Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol. Vol. 148. Proceedings of Machine Learning Research. PMLR, 2021, pp. 285–301. URL: <https://proceedings.mlr.press/v148/lockwood21a.html> (page 1).
- [22] Andrea Skolik, Stefano Mangini, Thomas Bäck, Chiara Macchiavello, and Vedran Dunjko. “Robustness of quantum reinforcement learning under hardware errors”. In: *EPJ Quantum Technology* 10.1 (2023), p. 8. ISSN: 2196-0763. DOI: [10.1140/epjqt/s40507-023-00166-1](https://doi.org/10.1140/epjqt/s40507-023-00166-1). URL: <https://doi.org/10.1140/epjqt/s40507-023-00166-1> (page 1).
- [23] S Lokesh, C Sakthi Jay Mahenthara, S Parvatha Kumaran, Palaniyappan Sathyaprakash, and Vaithiyashankar Jayakumar. “Implementation of Quantum Deep Reinforcement Learning Using Variational Quantum Circuits”. In: *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*. 2022, pp. 1–4. DOI: [10.1109/TQCEBT54229.2022.10041479](https://doi.org/10.1109/TQCEBT54229.2022.10041479) (page 1).
- [24] Samuel Yen-Chi Chen. *Quantum deep Q learning with distributed prioritized experience replay*. 2023. arXiv: [2304.09648](https://arxiv.org/abs/2304.09648) [quant-ph]. URL: <https://arxiv.org/abs/2304.09648> (page 1).
- [25] Hao-Yuan Chen, Yen-Jui Chang, Shih-Wei Liao, and Ching-Ray Chang. “Deep Q-learning with hybrid quantum neural network on solving maze problems”. In: *Quantum Machine Intelligence* 6.1 (2024), p. 2. ISSN: 2524-4914. DOI: [10.1007/s42484-023-00137-w](https://doi.org/10.1007/s42484-023-00137-w). URL: <https://doi.org/10.1007/s42484-023-00137-w> (page 1).
- [26] Sofene Jerbi, Casper Gyurik, Simon C. Marshall, Hans J. Briegel, and Vedran Dunjko. “Parametrized quantum policies for reinforcement learning”. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS ’21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393 (page 1).
- [27] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. “A Theoretical Analysis of Deep Q-Learning”. In: *Proceedings of the 2nd Conference on Learning for Dynamics and Control*. Ed. by Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger. Vol. 120. Proceedings of Machine Learning Research. PMLR, 2020, pp. 486–489. URL: <https://proceedings.mlr.press/v120/yang20a.html> (page 2).