

Tight gradient bounds in Parameterized Quantum Reinforcement Learning policies

André Sequeira¹

¹*High Assurance Software Laboratory, INESC TEC, Braga, Portugal*

Parameterized quantum circuit (PQC) based policies have emerged as a promising route toward quantum advantage in reinforcement learning (RL), with policy classes based either on Born-rule action decoding or on softmax post-processed quantum expectation values [1, 3]. However, recent work has shown that such policies may suffer from Barren Plateaus (BPs), in particular under local 2-design assumptions, leading to exponentially vanishing policy gradients with the number of qubits, hindering optimization and quantum advantage at scale [4, 5]. In this work, we analyze the trainability of PQC-based policies beyond the 2-design setting by deriving tight gradient and policy-variance bounds for standard PQCs used in practice by the community, building on light-cone techniques of Ref. [2].

PQC-based Born policies are obtained from partitioned Hilbert spaces into disjoint action subspaces. Given a n -qubit parameterized state $|\psi(s, \theta)\rangle$, the probability of action a is defined as $\pi(a|s, \theta) = \langle \psi(s, \theta) | P_a | \psi(s, \theta) \rangle$, where $P_a = \sum_{v \in V_a} |v\rangle\langle v|$ projects onto the subspace V_a associated with action a . For PQCs with L gates such that $U(\theta) = \prod_{l=0}^{L-1} W_l R_l(\theta_l)$ where R_l is a parameterized rotation gate followed by potentially entangling and Clifford operation W_l . Under uniformly initialized variational parameters $\theta_l \sim U[-\pi, \pi]$, and for data encoded through *angle encoding with input scaling* or *basis encoding*, we prove that the variance of both the policy and its gradients decomposes into independent contributions from the basis elements of the action subspace. In particular, $\text{Var}_\theta[\pi(a|s, \theta)] = \sum_{v \in V_a} \text{Var}_\theta[\langle \psi(s, \theta) | v \rangle \langle v | \psi(s, \theta) \rangle] \in \Omega(\frac{1}{\text{poly}(n)})$ provided that *at least one basis element in the corresponding action subspace has a polynomially vanishing contribution*. For k -body projectors where $k \in \mathcal{O}(\log(n))$ expressed in the Pauli basis $|v\rangle\langle v| = \frac{1}{2^k} \sum_{i=0}^{2^k-1} \alpha_i P_i$, where $P_i \in \{I, Z\}^{\otimes k}$ and $\alpha_i \in \{-1, 1\}$, we show that polynomially vanishing contribution occurs whenever the backwards light-cone of a single Pauli string P_i is $\mathcal{O}(\log(n))$.

These results provide a constructive recipe for designing BP-free PQC-based policies at initialization. In particular, they show that partition functions V_a that realize policies composed of local and increasingly global terms can retain trainable local contributions at initialization while still incorporating global structure in later stages of training. We analyze a previously proposed recursive parity-based action partitioning scheme [3] and prove that, while a standard parity function associated with the two-action base case suffers from exponential concentration and thus cannot be trained at scale, recursive parities introduced for larger action spaces embed local terms whose variance contributions can place the policy in a trainable regime at initialization. Numerical experiments across several ansätze far from 2-designs confirm the predicted polynomial scaling of policy variance and support the tightness of the subspace-variance bounds.

We further investigate the optimization dynamics induced by this *mixed* local-global policy structure. In probability maximization tasks, we observe that local terms drive the policy away from the initial concentration regime, after which the gradients of more global terms start contributing later in training once the local terms update the loss function landscape. We show that global term contribution is indeed crucial to allow the target action probability to approach one. Furthermore, increasing the number of local and intermediate-locality terms improves convergence speed and stability, which is obtained from global terms kicking in earlier during optimization thanks to adding more local contributions, thus molding the loss landscape to favorable regions earlier in time. This behavior is also observed in RL training settings such as CartPole and in a synthetic contextual bandit environment with 20 actions, where richer locality structure systematically improved the performance of PQC-based agents. Our results suggest that mixed-locality quantum models offer a principled path toward trainable and potentially classically hard PQC-based RL policies. Future work will address rigorous classical hardness guarantees and extend the analysis to softmax-activated quantum policies with controllable greediness.

References

- [1] S. Jerbi, C. Gyurik, S. Marshall, H. Briegel, and V. Dunjko, “Parametrized Quantum Policies for Reinforcement Learning,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 28362–28375. Available: <https://proceedings.neurips.cc/paper/2021/hash/eec96a7f788e88184c0e713456026f3f-Abstract.html>.
- [2] A. Letcher, S. Woerner, and C. Zoufal, “Tight and Efficient Gradient Bounds for Parameterized Quantum Circuits,” *Quantum*, vol. 8, p. 1484, Sep. 2024. doi: 10.22331/q-2024-09-25-1484.
- [3] N. Meyer, D. D. Scherer, A. Plinge, C. Mutschler, and M. J. Hartmann, “Quantum Policy Gradient Algorithm with Optimized Action Decoding,” May 22, 2023, arXiv:2212.06663. doi: 10.48550/arXiv.2212.06663.
- [4] A. Sequeira, L. Paulo Santos, and L. Soares Barbosa, “Trainability Issues in Quantum Policy Gradients with Softmax Activations,” in *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, Montreal, QC, Canada: IEEE, Sep. 2024, pp. 335–340. doi: 10.1109/QCE60285.2024.10302.
- [5] A. Sequeira, L. P. Santos, and L. S. Barbosa, “Trainability issues in quantum policy gradients,” Jun. 13, 2024, arXiv:2406.09614. Available: <http://arxiv.org/abs/2406.09614>.