Trainability issues in quantum policy gradients

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 035037 https://doi.org/10.1088/2632-2153/ad6830

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
17 April 2024

REVISED
20 June 2024

ACCEPTED FOR PUBLICATION
26 July 2024

PUBLISHED
6 August 2024

PAPER

Trainability issues in quantum policy gradients

André Sequeira[1,2,3,*], **Luis Paulo Santos**[1,2,3] **and Luis Soares Barbosa**[1,2,3]

[1] Department of Informatics, University of Minho, Braga, Portugal
[2] High Assurance Software Laboratory, INESC TEC, Braga, Portugal
[3] Quantum Linear-optical computation group, International Nanotechnology Laboratory, Braga, Portugal
[*] Author to whom any correspondence should be addressed.

E-mail: andre.sequeira@inl.int

Keywords: quantum policy gradients, barren plateaus, quantum reinforcement learning

Abstract

This research explores the trainability of Parameterized Quantum Circuit-based policies in Reinforcement Learning, an area that has recently seen a surge in empirical exploration. While some studies suggest improved sample complexity using quantum gradient estimation, the efficient trainability of these policies remains an open question. Our findings reveal significant challenges, including standard Barren Plateaus with exponentially small gradients and gradient explosion. These phenomena depend on the type of basis-state partitioning and the mapping of these partitions onto actions. For a polynomial number of actions, a trainable window can be ensured with a polynomial number of measurements if a contiguous-like partitioning of basis-states is employed. These results are empirically validated in a multi-armed bandit environment.

1. Introduction

Variational Quantum Algorithms (VQAs), emerging as a cornerstone in the Noisy Intermediate Scale Quantum (NISQ) era, present a novel approach to overcoming the limitations inherent in quantum computing, such as restricted qubit availability and noise related constraints on circuit depth. Initially proposed as universal computation models [4], VQAs operate through a synergy of quantum and classical mechanisms. They utilize a Parameterized Quantum Circuit (PQC) where the parameters are fine-tuned via a classical optimization routine to achieve the global optimum of a specified objective function [5]. Despite the theoretical allure of VQAs, their practical efficiency is often hampered by the so-called barren plateau (BP) phenomenon, a critical challenge in quantum optimization [13]. This phenomenon, characterized by the exponential suppression of the gradients' magnitude with an increasing number of qubits, requires an exponentially large number of measurements to allow the algorithm to effectively navigate through the optimization landscape.

The BP phenomenon pose a significant hurdle, not only in gradient-based but also in gradient-free optimization approaches, where cost concentration emerges as a parallel challenge [2]. Understanding and mitigating the occurrence of BPs in specific VQAs is thus vital for harnessing any potential quantum advantage. Several factors contribute to the emergence of BPs, including deep and random quantum circuits [13], PQCs adhering to a volume law in entanglement entropy [12] etc. The work of Cerezo *et al* [6] particularly highlights the dependence of the BP phenomenon on the locality of the cost function, showing that local losses measured on a logarithmic number of qubits can retain trainability in shallow circuits [16].

Further complicating the picture, conventional machine learning cost functions like the mean squared error, negative log likelihood, and KL-divergence have been shown to lead to BPs [21]. BPs are typically characterized by the scaling of the variance of partial derivatives of the cost function, which diminishes exponentially with the number of qubits [13]. This scaling results in gradients increasingly concentrating around zero, making optimization exceedingly difficult. Another approach to characterize a BP is through the study of cost concentration [2], where cost differences between randomly selected points in the landscape show an exponential concentration with increasing qubits. In addition, the Fisher Information Matrix (FIM) spectrum, as explored in the work of Abbas *et al* [1], offers valuable insights into the flatness of the loss

landscape in the presence of BPs, with the eigenvalues of the FIM becoming exponentially small as the number of qubits increases.

Recent studies have expanded the application of VQAs to Reinforcement Learning (RL), showing promising results [10, 19, 20]. Notably, the work of Jerbi *et al* [11] demonstrated a quadratic optimization improvement in policy-based RL agents using PQC-based policies over classical agents. However, the trainability of these quantum policies, particularly in the face of BPs, remains an open question. In this context, our study aims to provide a deeper understanding of the trainability issues associated with PQC-based policies in RL, focusing on cost-function dependent BPs and their implications. We explore the challenges faced by specific variations of previously proposed *Raw Policies* [10, 14], and investigate their performance under various conditions. Our findings contribute to the ongoing research on optimizing PQC-based agents in quantum RL, addressing critical questions on the interplay between policy types, number of qubits, action-space size, and the presence of BPs and other trainability issues such as exploding gradients. This research not only advances our understanding of quantum RL but also sets the stage for future investigations into other types of PQC-based policies [10, 19], thereby unlocking the full potential of quantum computing in machine learning applications.

1.1. Related work

Recent advancements have been documented concerning the application of VQAs to RL. There has been considerable empirical evidence supporting the efficacy of VQAs in diverse benchmark environments, encompassing both value-based [8, 20] and policy-based [10, 19] RL paradigms. A significant contribution in this field was made by Jerbi *et al* [11], who demonstrated a quadratic improvement in gradient estimation for optimizing policy-based RL agents using PQC-based policies compared to purely classical agents. In another notable work, Cherrat *et al* [9] introduced quantum neural network architectures featuring orthogonal and compound layers for policy and value functions, notably devoid of BPs in the context of financial hedging. At the same time, Meyer *et al* [14] posited that a global parity-based policy could provide more information to the agent and a more conducive optimization landscape. This proposition challenges the previously held belief that global measurements lead to flatter landscapes, implying further issues on trainability. The emerging divergence in these findings entails the need for further research to fully understand the impact of the BP phenomenon within RL, especially in the context of generalized PQC-based policies, as it may significantly influence optimization efficiency provided by gradient estimation.

This investigation centers on analyzing *cost-function dependent barren plateaus* within the framework of policy-based RL, utilizing both local and global projector-based observables in conjunction with PQC-based policies. The primary objective of this study is to delineate variance limits for the gradient of the REINFORCE policy-dependent objective function [23], especially under the assumption of a PQC-based policy. We re-examine two previously introduced policies, redefined here for enhanced clarity: (1) The *Contiguous-like Born policy*, as referenced in [10], derived from categorizing basis states into a contiguous set proportional to the action-space size, and (2) The *Parity-like Born policy*, detailed in [14], formulated through a recursive parity function applied to measured basis states.

1.2. Contributions

Our findings highlight that both contiguous and parity-like Born policies can potentially face extreme challenges in terms of trainability. On one side, the policy might encounter standard BPs characterized by exponentially vanishing gradients, while on the other, it may face issues of gradient explosion. These phenomena are heavily influenced by the locality of the observables employed that depend on the action-space size. For n qubit policies estimated through $\mathcal{O}(\text{poly}(n))$ measurements, the *contiguous-like Born policy* exhibits a trainable region at logarithmic depth $\mathcal{O}(\log(n))$, assuming the action-space is of $\mathcal{O}(n)$ size. For a $\mathcal{O}(\text{poly}(n)$ number of actions, the policy enters a transition region where the locality of the observables increase but it is still possible to train under polynomially large number of measurements. Conversely, under the same conditions, the *Parity-like Born policy* is untrainable, suffering from a BP. Beyond polynomially-sized action spaces, no policy can be trained using a polynomial number of measurements since the probability of measuring basis states becomes exponentially suppressed with the number of qubits. In such a scenario, the gradient behavior shifts towards exploding gradients due to the exponentially small probabilities.

The trainability of PQC-based policies was further analyzed by inspecting the FIM spectrum. It was observed that, under polynomially sized actions spaces, the FIM spectrum indeed reveals a BP for the *Parity-like Born policy*, as FIM entries shrink exponentially with increasing qubits, resulting in a spectrum highly concentrated at zero, therefore characterizing a flat landscape. Outside polynomial action spaces, the FIM spectrum becomes less informative about BPs due to the exponentially small probabilities that induce large FIM entries, causing a shift in the spectrum with more eigenvalues concentrated away from zero.

Empirical validation of these results was achieved by examining the scaling of the variance of the log likelihood gradient, using the simplified two-design ansatz [6] for the PQC-based policy. Furthermore, the effect of the observables' global nature on a PQC-based agent's trainability was explored in the context of learning to select the optimal arm in a simulated multi-armed bandit environment. The observations confirmed that for a PQC-based agent with a polynomial number of actions, the contiguous-like Born policy is capable of learning the optimal arm, unlike the parity-like policy. However, when extending beyond a polynomial number of actions, both policies were unable to learn the optimal arm, in line with our theoretical predictions.

The rest of the document is organized as follows: section 2 introduces the policy gradient framework in RL and the intricacies behind PQC-based policies such as gradient estimation. Section 3 establishes novel results forming the core of this work. It provides clear lower bounds for the policy gradient's variance. Section 4 resorts to numerical experiments as an empirical validation of the theoretical predictions in section 3. Finally, section 5 concludes the work and outlines future research directions.

2. Quantum policy gradients

Policy Gradient algorithms are designed to optimize a parameterized policy $\pi(a|s,\theta) = \mathbb{P}\{a_t = a|s_t = s, \theta_t = \theta\}$, where $\theta \in \mathbb{R}^k$ denotes the parameter vector with dimension k, s, a, and t represent the state, action and the time step, respectively. The essence of this approach is to enable optimal action selection without relying on a value function, with the primary aim of maximizing a performance measure $J(\theta)$. This is achieved by applying gradient ascent to $J(\theta)$ as follows:

$$\theta_{i+1} = \theta_i + \eta \nabla_{\theta_i} J(\theta_i) \tag{1}$$

where η is the learning rate. For discrete and small action spaces, a Softmax-Policy is commonly used to balance exploration and exploitation. The Monte-Carlo policy gradient, known as REINFORCE, estimates the gradient from samples across N trajectories of length T, or the horizon, under the parameterized policy. A known limitation of REINFORCE is the high variance of its gradient estimation due to the stochastic nature of sampling trajectories. This variance can negatively affect performance in complex settings. Introducing a baseline denoted by $b(s_t)$, such as the average return, can reduce the variance without having to increase the number of samples N. The baseline is subtracted from the returned value to stabilize the optimization process, as shown in equation (2)

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \left(G_t(\tau_i) - b(s_{t_i})\right) \nabla_\theta \log \pi\left(a_{t_i}|s_{t_i}, \theta\right) \tag{2}$$

where $G_t(\tau_i)$ is the cumulative discounted return at time step t in trajectory τ_i. Throughout the rest of the paper, the baseline $b(s_t)$ is considered as the average return across all trajectories

$$b(s_t) = \frac{1}{N} \sum_{i=0}^{N-1} G_t(\tau_i). \tag{3}$$

In this work, we consider PQC-generated policies i.e. policies generated from PQCs. Specifically we consider two variants of the *raw policies* proposed in the literature and redefined here for enhanced clarity: (1)The *Contiguous-like Born policy* [10] and (2) *Parity-like Born policy* [14]. For completion, the Softmax-based PQC policy [10, 19] is also defined but addressing its trainability is outside of the scope of this work. Let us start with the most general definition of a Born policy.

2.1. Born policy

Definition 2.1. Let $s \in \mathcal{S}$ be a state embedded in an n-qubit parameterized quantum state, $|\psi(s,\theta)\rangle$, where $\theta \in \mathbb{R}^k$. The probability associated to a given action $a \in A$ is given by:

$$\pi(a|s,\theta) = \langle P_a \rangle_{s,\theta} = \langle \psi(s,\theta) | P_a | \psi(s,\theta) \rangle \tag{4}$$

where $P_a = \sum_{\nu \in V_a} |\nu\rangle\langle\nu|$ is the projector into partition $V_a \subseteq V$ where $V = \{\nu_0, \nu_1, \ldots, \nu_{2^n-1}\}$ is the set of eigenstates of an observable

$$O = \sum_{i=0}^{2^n-1} \lambda_i |\nu_i\rangle\langle\nu_i|. \tag{5}$$

Moreover, $\bigcup_{a \in A} V_a = V$ and $V_a \cap V_{a'} = \emptyset$, for all $a \neq a'$.

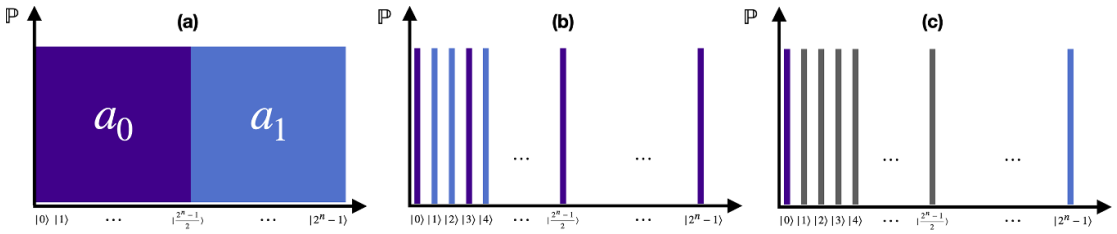

Figure 1. Partitions considered for the base case of $|A| = 2$. (a) and (b) Illustrates a Contiguous-like and parity-like partitioning, respectively, of all 2^n basis states. (c) Action-projector-like partitioning considering just two basis states.

Definition 2.1 introduces the most general definition of a Born policy. However, there could be partitions that do not take into account every eigenstate of a given observable, as described above. In these scenarios, the probability associated to a given action would not be normalized as before since $\sum_{a\in A} P_a \neq I$. Moreover, since the goal of this work is the study of cost-function dependent BPs in quantum policy gradients, different partitions V_a, and the associated globality of the measurement should be further clarified.

Figure 1 illustrates different partitions considered throughout this work for $|A| = 2$, which constitutes the base case in RL.

2.1.1. Contiguous-like Born policy

Consider the action space $A = \{a_0, a_1\}$. The simplest partitioning that fits definition 2.1, would be to separate all basis states in half i.e. $V_{a_0} = \{|0\rangle, |1\rangle, |2\rangle \ldots |\frac{2^n-1}{2}\rangle\}$ and $V_{a_1} = \{|\frac{2^n}{2}\rangle \ldots |2^n - 1\rangle\}$. Such partitioning is illustrated in figure 1(a). In this case, even though every n-bit bitstring is considered to build the policy, a careful analysis of the partitioning indicates that it does not correspond to a global measurement. It is possible to assign a bitstring to its respective set by just measuring the first bit. If the bit is in state $|0\rangle$ (respectively, $|1\rangle$) it corresponds to the set V_{a_0} (respectively, V_{a_1}). Thus, such assignment corresponds to a 1-local measurement, indeed.

In general, for an arbitrary number of actions $|A| \leqslant 2^n$ if we assign each bitstring to $|A|$ contiguous sets, then the measurement will actually be $(\log|A|)$-local, since to assign each bitstring $\log|A|$ bits are required to distinguish between the sets. As an example. let the the number of qubits be $n = 3$. The total number of bitstrings is $2^3 = 8$, corresponding to the set $\{000, 001, 010, 011, 100, 101, 110, 111\}$. Suppose $|A| = 4$ with partition set $V = V_0 \cup V_1 \cup V_2 \cup V_3$. The number of bits needed to distinguish between sets is $\log_2(4) = 2$. Thus, the first 2 bits of each bitstring are considered to assign it to one of the 4 sets. Let a be represented in its binary expansion. Then, the partitioning will be given by $V = \{000, 001\} \cup \{010, 011\} \cup \{100, 101\}$ $\cup\{110, 111\} \cup \{110, 111\}$. and the measurement will be 2-local.

2.1.2. Parity-like Born policy

Notice that for the base case $|A| = 2$, the contiguous-like Born policy loses expressivity since the measurement becomes 1-local. We can actually devise a more expressive assignment by considering a parity function, as illustrated in figure 1(b). The 2^n bitstrings in a n-qubit PQC can be considered assigning each of them by the parity of the bitstring (number of 1's). Thus, the policy is represented as:

$$\pi(a|s,\theta) = \sum_{b\in\{0,1\}^n}^{\oplus b=a} \langle\psi(s,\theta)|b\rangle\langle b|\psi(s,\theta)\rangle. \tag{6}$$

Such an assignment constitutes a global measurement and the authors of [14] showed that it corresponds to the assignment that maximizes the extracted information. Notice that instead of the Pauli-Z measurement on every qubit, one could instead measure either a single-qubit or an ancilla, provided a CNOT cascade prior to the measurement, as highlighted in [14]. For $|A| > 2$, the authors designed a partitioning based on a recursive parity function which they conjecture to be optimal in the sense of extracted information and globality. Let $m = \log|A|$ be the number of recursive calls and $\boldsymbol{b}$ be a n-bit bitstring measured through sampling a PQC. Then, the partition can be defined recursively as,

$$\mathcal{C}^{(m)}_{[a]_2} = \left\{ \boldsymbol{b} \,|\, \bigoplus_{i=m}^{n-1} b_i = a_0 \wedge \boldsymbol{b} \in \mathcal{C}^{(m-1)}_{a_m\cdots a_2(a_1\oplus a_0)} \right\} \tag{7}$$

Table 1. Summary of the locality of the measurement for the different partitions considered in this work.

Born policy	Locality of measurement
Contiguous-like	$\log\|A\|$—local
Parity-like	n-local
Action-projector-like	n-local

where $[a]_2 = a_m \ldots a_0$ is the binary expansion of action a. Since for computing the parity, each of the n bits is necessary, a parity-based policy will be composed of a global measurement (or n-local) for $|A| = 2$ as base case. Thus, it will always be global independently of the number of actions.

2.1.3. Action-projector-like Born policy

There can also be partitions that do not take into account every eigenstate of a given observable. For instance, for the base case $|A| = 2$, we could assign the all-zero state to action a_0, $\langle P_{a_0}\rangle_{s,\theta} = |\langle 0|\psi(s,\theta)\rangle|^2$ and the all-ones state to action a_1, $\langle P_{a_1}\rangle_{s,\theta} = |\langle 1|\psi(s,\theta)\rangle|^2$, and discard all other basis states, as illustrated in figure 1(c). In this case, the probability would need to be further normalized:

$$\pi(a|s,\theta) = \frac{\langle P_a\rangle_{s,\theta}}{\sum_{a'\in A}\langle P_{a'}\rangle_{s,\theta}}. \tag{8}$$

For $|A| = 2^n$ it makes sense to assign each eigenstate to an action. In such case, the measurement would be n-local.

Table 1 summarizes the locality of the measurement for the different partitions considered in this work. The locality is expressed as a function of $|A|$.

2.2. Softmax policy

Definition 2.2. Let $s \in \mathcal{S}$ be a state embedded in an n-qubit parameterized quantum state, $|\psi(s,\theta)\rangle$, where $\theta \in \mathbb{R}^k$. Let O_a be an arbitrary observable composed by the sum of m local/global terms $O_a = \sum_{i=0}^{m-1}\langle O_i\rangle$ representing the numerical preference of action $a \in \mathcal{A}$ and β an hyperparameter. The probability associated to a given action a for a softmax policy is given by:

$$\pi(a|s,\theta) = \frac{e^{\beta\langle O_a\rangle_{s,\theta}}}{\sum_{a'} e^{\beta\langle O_{a'}\rangle_{s,\theta}}} \tag{9}$$

where β is often referred as the inverse temperature hyperparameter that is responsible for the control of the policies greediness. That is, the softmax policy allows for greater control compared to the Born policy , since β can control the degree in which we select what we think to be the best action or explore other actions. The higher the β the more greedy the policy is [10].

2.3. Gradient estimation

The policy gradient (equation (2)) is in its essence classical with the exception of the log policy gradient in which the gradient w.r.t the PQC must be computed. In that regard, the log policy gradient must be expressed as the gradient of the expectation value of an observable and the parameter-shift rule [17] can be applied to compute the gradient using quantum hardware. Let $\langle O\rangle_\theta$ be the parameterized expectation value of the observable O. The parameter-shift rule is a hardware-friendly technique to compute the partial derivative of $\langle O\rangle_\theta$ w.r.t θ. Explicitly, it states the equality

$$\frac{\partial\langle O\rangle_\theta}{\partial\theta_l} = \frac{1}{2\sin\alpha}\left[\langle O\rangle_{\theta+\alpha e_l} - \langle O\rangle_{\theta-\alpha e_l}\right] \tag{10}$$

where e_l indicates that the parameter θ_l is being shifted by α. The partial derivative can be obtained using two expectation value estimates, each requiring a number of quantum circuit evaluations. Thus, for $\theta \in \mathbb{R}^k$, the gradient can be estimated using $2k$ total quantum circuit evaluations. The gradient accuracy is maximized at $\alpha = \frac{\pi}{4}$, since $\frac{1}{\sin\alpha}$ is minimized at this point. For arbitrary functions of expectation values like the log policy gradient, the gradient can be obtained via the standard chain rule.

For the Born policy the chain rule gives the following expression for the log policy gradient partial derivatives

$$\partial_{\theta_l}\log\pi(a|s,\theta) = \partial_{\theta_l}\log\langle P_a\rangle_{s,\theta} = \frac{\partial_{\theta_l}\langle P_a\rangle_{s,\theta}}{\langle P_a\rangle_{s,\theta}} \tag{11}$$

which results clearly in a unbounded gradient expression. The full REINFORCE algorithm with PQC-based policies explored in this work is outlined in algorithm 1.

Algorithm 1: PQC-based REINFORCE with Baseline.

Input: PQC-based policy π_θ with $\theta \in \mathbb{R}^k$. Learning rate η and horizon T
Output: Updated parameters θ^*
/* Loop until the stopping condition is met */
while *True* **do**
  /* Generate trajectories following the policy π_θ */
  for $i = 0 \dots N-1$ **do**
    Generate $\tau_i = \{(s_0, a_0, r_0), \dots, (s_{T-1}, a_{T-1}, r_{T-1})\}$ under PQC-based policy
  Compute gradient with baseline, $\nabla_\theta J(\theta)$ as in equation (2) using the parameter-shift rule (equation (10))
  /*Update parameters via gradient ascent */
  $\theta = \theta + \eta \nabla_\theta J(\theta)$

3. Trainability issues in Born policies

This section presents new findings that form the cornerstone of this study, addressing the trainability of the quantum policy gradient algorithm as outlined in algorithm 1. We focus on Contiguous and Parity-like PQC-based Born policies defined in definition 2.1. A key aspect of this investigation is the analysis of the variance of the log policy gradient for these policies, considering the impact of the number of qubits and actions, which subsequently influences the globality of associated observables, as detailed in table 1. The analysis proceeds as follows:

1. *Analysis of Product States* (section 3.1): We begin with an examination of product states as an instructive case, discussing the behavior and characteristics of the log policy gradient variance in this simplified scenario.
2. *Consideration of Entangled States* (section 3.2): We extend the analysis to include entangled states, comparing and contrasting the findings with those from the product states to highlight the effects of entanglement on trainability.
3. *Unified Variance Analysis* (section 3.3): We conduct a unified analysis of variance as a function of the number of actions, providing a comprehensive overview of how the variance scales with an increasing number of actions and its implications for the trainability of PQC-based policies.

By systematically analyzing these cases, we aim to provide a thorough understanding of the factors influencing the trainability of quantum policy gradient algorithms and offer insights into optimizing PQC-based policies for practical applications. Since the variance of the log policy partial derivative is desired, we start with a simplification of the REINFORCE policy gradient objective, expressed in equation (2), to an expression that depends only on the variance of the policy. This approach allows for an accurate study of trainability as a function of different PQC-based policies. In the following, we consider the trivial upper bound in terms of relevant quantities in RL to rephrase the variance expression as a function of the policy.

Lemma 3.1. *Let $\pi(a|s,\theta)$ be a n-qubit PQC-based policy with $\theta \in \mathbb{R}^k$. Let T be the trajectories horizon, R_{max} be the maximum reward and γ the trajectories discount factor. Then, the policy gradient variance w.r.t variational parameters θ is upper bounded by*

$$\mathbb{V}_\theta \left[\partial_\theta v_\pi(s)\right] \leqslant \frac{R_{max}^2 T^4}{(1-\gamma)^4} \mathbb{V}_\theta \left[\partial_\theta \log \pi(a|s,\theta)\right] \tag{12}$$

Proof.

$$\begin{aligned}
\mathbb{V}_\theta \left[\partial_\theta v_\pi(s)\right] &= \mathbb{V}_\theta \left[\frac{1}{N} \sum_{i=0}^{N-1} \sum_{t=0}^{T-1} G_t(\tau_i)\, \partial_\theta \log \pi\left(a_t^i | s_t^i, \theta\right)\right] \\
&= \frac{1}{N^2} \mathbb{V}_\theta \left[\sum_{i=0}^{N-1} \sum_{t=0}^{T-1} G_t(\tau_i)\, \partial_\theta \log \pi\left(a_t^i | s_t^i, \theta\right)\right] \\
&\leqslant \frac{1}{N^2} \left(\sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \sqrt{G_t^2 \mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a_t^i | s_t^i, \theta\right)\right]}\right)^2 \qquad \text{(A)} \\
&= \frac{G_t^2}{N^2} \left(\sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \sqrt{\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a_t^i | s_t^i, \theta\right)\right]}\right)^2 \qquad \text{(B)}
\end{aligned}$$

$$\leqslant G_t^2 T^2 \mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \tag{C}$$

$$= \frac{R_{max}^2 T^4}{\left(1-\gamma\right)^4} \mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \tag{D}$$

where: (A) Follows from the variance of the sum of random variables $\left(\mathbb{V}\left[\sum_i X_i\right] \leqslant \left(\sum_i \sqrt{\mathbb{V}[X_i]}\right)^2\right)$. (B) Follows from variance of a constant a times a random variable $X(\mathbb{V}[aX] = a^2\mathbb{V}[X])$. (C) Considers the upper bound on N and T. (D) Considers the trivial upper bound on the return (see appendix A), following the independence of θ. □

Lemma 3.1 indicates that the variance of the log policy objective function increases with relevant quantities in RL. Specifically, the variance increases with the reachable maximum reward, the horizon and the discount factor. At this stage trainability of PQC-based agents can be evaluated through the scaling of the variance of the log policy gradient $\mathbb{V}_\theta\left[\partial_\theta \log \pi(a|s,\theta)\right]$ as a function on the number of qubits. To that end, let us start by analyzing the behavior of the gradients in the context of product states and build from there towards general entangled quantum states.

3.1. The instructive case of product states

We begin by examining the straightforward scenario of a product state. In [6], the authors explored a simple n-qubit parameterized model described by the unitary $V(\theta) = \bigotimes_{i=0}^{n-1} e^{-i\theta_i \sigma_x}$. They focused on the global observable $O_G = 1 - |0\rangle\langle 0|$ to prepare the all-zero state. Although this PQC corresponds to a single layer of parameterized Pauli rotations forming a separable state, it was shown to suffer from BPs. The global observable results in a cost function $C_G(\theta) = 1 - \prod_{i=0}^{n-1} \cos^2(\theta_i)$, whose variance decays exponentially with the number of qubits due to its global nature. The authors then suggested the local observable composed by individual qubit contributions $O_L = 1 - \sum_{j=0}^{n-1} |0\rangle\langle 0|_j \otimes \mathbb{I}_{\bar{j}}$ with cost function $C_L(\theta) = 1 - \frac{1}{n}\sum_{i=0}^{n-1} cos(\theta)^2$. This modification ensures that the variance of the cost function decays polynomially with the number of qubits, thus avoiding BPs. Such finding emphasizes the critical role of a well-crafted cost function.

In the broader context of machine learning, and policy gradients specifically, the log-likelihood is often preferred over direct probability as considered before. Such cost-function leads to different behavior. For an arbitrary product state $|\psi\rangle$, the probability of the all-zero state is given by:

$$|\langle 0|\psi\rangle|^2 = \prod_{i=0}^{n-1} |\langle 0_i|\psi\rangle|^2. \tag{13}$$

The decomposition into individual qubit contributions enables a product state to avoid BPs since the log likelihood cost-function separates the product into a sum of individual qubit contributions. To apply this reasoning to the REINFORCE cost function in RL, where the focus is on the log policy gradient, consider a Born policy with $|A| = 2^n$ and a global projector $|a\rangle\langle a|$ for action a. The policy is expressed as $\pi(a|s,\theta) = |\langle a|\psi(s,\theta)|^2$. If the parameterized state is a product state, the probability can be decomposed into individual qubit contributions as follows:

$$\pi\left(a|s,\theta\right) = |\langle a|\psi\left(s,\theta\right)|^2 = \prod_{i=0}^{n-1} |\langle a_i|\psi\left(s,\theta\right)|^2 \tag{14}$$

where a_i represents the individual qubit projector $|a_i\rangle\langle a_i| \otimes \mathbb{I}_{\bar{i}}$ on the ith qubit, applying the identity operation to the other qubits. Considering the variance of the log policy gradient:

$$\begin{aligned}
\mathbb{V}_\theta\left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] &= \mathbb{V}_\theta\left[\partial_\theta \log \prod_{i=0}^{n-1} |\langle a_i|\psi\left(s,\theta\right)|^2\right] \\
&= \mathbb{V}_\theta\left[\sum_{i=0}^{n-1} \partial_\theta \log |\langle a_i|\psi\left(s,\theta\right)|^2\right] \\
&= \sum_{i=0}^{n-1} \mathbb{V}_\theta\left[\partial_\theta \log |\langle a_i|\psi\left(s,\theta\right)|^2\right] \qquad \text{(A)}
\end{aligned}$$

where (A) follows from the linearity and independence of the observables [22]. The variance of the log policy gradient becomes the sum of the variances of the log probabilities of each individual qubit. Notice that since we have a product state, the partial derivative would in fact not depend on the number of qubits, provided that different parameters are part of the circuit. Only in the scenarion where the parameters are shared across

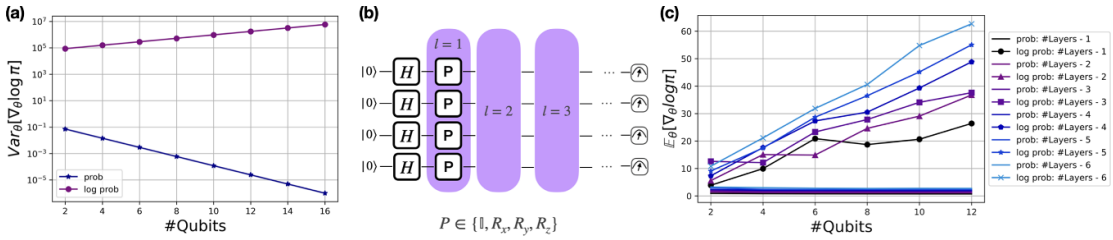

Figure 2. Variance and expectation value of the gradient of log-probability cost-function. (a) Variance for the all-zero state with parameterized state of individual qubit y-rotations. (b) Random product state composed of Pauli rotations sampled uniformly at random. (c) Expectation value for randomly sampled projectors in the random product state in (b).

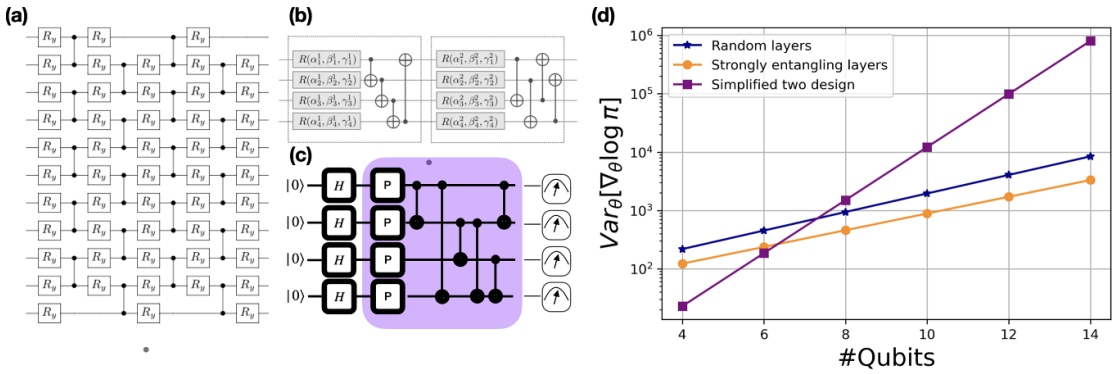

Figure 3. Variance of the log policy gradient for three distinct entangled states. (a) Simplified two design. (b) Strongly entangling layers. (c) Random states composed of Pauli rotations sampled uniformly at random followed by randomly selected CZ gates. (d) Variance as a function of the number of qubits for the circuits (a)–(c).

qubits the partial derivative will sum those terms and increase with the number of qubits, as illustrated in figure 2(a). Such behavior is propelled by the nature of a product state, where the probability of each individual qubit is independent of the other qubits. In figure 2(c) the expectation value of the partial derivative of log probability of the all-zero state is illustrated for the random product state illustrated in figure 2(b) where the parameters are shared per layer. That is $\theta_{i,l} = \theta_l$for all number of layers l. Indeed, the variance increases with both the number of qubits and layers, as expected.

3.2. Generalized behavior for entangled states

In this subsection, we analyze the variance of the log-probability for entangled states. In particular, we focus on the extreme case where $|A| = 2^n$, involving global projectors similar to the product states discussed in section 3.1. It is known that such measurements are susceptible to BPs [6] since the probability of each basis state in this scenario depends on a subset of qubits characterized by the entangled state, derived from an n-qubit global projector, assuming a PQC constituted by local two-design parameterized blocks. In figure 3(d) the variance of the log probability is illustrated as a function of the number of qubits for three distinct entangled quantum states: (1) Simplified 2-design ansatz illustrated in figure 3(a). (2) Strongly entangling layers, depicted in figure 3(b). (3) State generated from Pauli rotations sampled uniformly at random followed by randomly selected CZ gates, as illustrated in figure 3(c). n layers of the blocks shown in their respective figures are employed. Moreover, projectors were sampled uniformly at random from the set of 2^n available ones and the variance illustrated for an average of a thousand experiments.

From figure 3(d), it is evident that in each experiment, the variance of the log-probability increases with the number of qubits when global projectors are considered. This behavior is akin to that observed in product states. However, the variance reaches extremely high levels as a function of n, indicating that although these circuits avoid BP, they are prone to exploding gradients. This phenomenon arises because the probabilities diminish exponentially with an increase in the number of qubits, leading to two major issues: (1) The log-probability gradient becomes exponentially large due to the vanishing probabilities. (2) An exponentially large number of quantum circuit executions is required to accurately estimate both the probability and its gradient. As the number of qubits grows, measuring the eigenstate of interest becomes increasingly challenging due to the exponentially concentrated probabilities [16]. However, recall that in the context of RL, we will need to do a partitioning of possibly all 2^n basis states into the set of available actions

$|A|$. Thus, the previous observation is no longer true once the number of actions is $|A| \in \mathcal{O}(\text{poly}(n))$ since the probabilities will no longer be exponentially small . In such cases, a trainable region could be created depending on the locality of the projector, which in turn is heavily influenced by the type of Born policy implemented. In the following subsection, we examine the variance of the cost function for different Born policies as a function on the number of actions $|A|$.

3.3. Variance as a function of $|A|$

Let us start with an analytical upper bound for the variance of the log likelihood cost function partial derivative, presented in lemma 3.2. Let $f(\pi_\theta) = \log \pi(a|s,\theta)$ for simplicity.

Lemma 3.2. *Consider a n-qubit Born policy $\pi(a|s,\theta)$ as in definition 2.1 with $|A|$ actions. Then, the upper bound for the variance of the log policy gradient is given by*

$$\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \leqslant 2\left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2 \left[\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] + \mathbb{E}_\theta \left[\partial_\theta \pi_\theta\right]^2\right] \tag{15}$$

Proof.

$$\begin{aligned}
\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] &= \mathbb{V}_\theta \left[\partial_\theta f(\pi_\theta)\right] \\
&= \mathbb{V}_\theta \left[\partial_{\pi_\theta} f(\pi_\theta)\, \partial_\theta \pi_\theta\right] && \text{(A)} \\
&\leqslant 2\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2 + 2\mathbb{E}_\theta \left[\partial_\theta \pi_\theta\right]^2 \mathbb{V}_\theta \left[\partial_{\pi_\theta} f \pi_\theta\right] && \text{(B)} \\
&\leqslant 2\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2 + 2\mathbb{E}_\theta \left[\partial_\theta \pi_\theta\right]^2 \left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2 && \text{(C)} \\
&= 2\left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2 \left[\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] + \mathbb{E}_\theta \left[\partial_\theta \pi_\theta\right]^2\right] && \text{(D)}
\end{aligned}$$

where (A) Follows from the chain rule; (B) Follows from variance of product of random variables $\mathbb{V}\left[XY\right] \leqslant 2\mathbb{V}\left[X\right]\left|Y^2\right|_\infty + 2\mathbb{E}\left[X\right]^2 \mathbb{V}\left[Y\right]$ as proposed in [21]; (C) Upper bound on the variance; (D) Algebraic manipulation. □

The upper bound depends entirely on the total number of actions and the observable considered to estimate the policy. Let us assume a global projector on n-qubits and either parameterized blocks before/after parameter θ form a 1-design. That way, the average of the partial derivative $E_\theta \left[\partial_\theta \pi_\theta\right] = 0$ [6]. If $|A| \in \mathcal{O}(\text{poly}(n))$, then w.l.g we can assume that $\pi_{\min} \in [b, 1]$ with $b \in \Omega(\frac{1}{\text{poly}(n)})$ [21]. In the context of RL, the log policy gradient is only computed for sampled actions. Thus, the probability of an action cannot be zero in the gradient estimation phase. Nevertheless, it can be arbitrarily close to zero. To avoid such an issue, clipping is often considered in practice. Thus, assuming $b \in \Omega(\frac{1}{\text{poly}(n)})$ works as some sort of clipping of probabilities. In general, provided discrete action spaces, $|A| \ll 2^n$. Therefore it is only reasonable to assume the possible number of outcomes to be $|A| \sim \text{poly}(n)$. Thus, considering $b \in \Omega(\frac{1}{\text{poly(n)}})$ as proposed in [21] does not pose any issues under these conditions. However, this assumption breaks down in the most general case when $|A| = 2^n$, associating each basis state projector in an n-qubit PQC to an action, as the probabilities become exponentially small with the number of qubits. Consequently, for action spaces larger than $\text{poly}(n)$, the number of quantum circuit executions required to accurately estimate the policy becomes impractical, eventually scaling exponentially with the number of qubits. Given that the total number of features in an RL agent's state, s_f, is typically large and $s_f \gg |A|$ for a discrete action space, several qubits are often required to encode the state of the agent. If the standard angle encoding scheme is employed, as seen in most literature [8, 10, 11, 19], then $n \sim s_f$, which implies that $|A| \ll 2^n$, validating the $\text{poly}(n)$ clipping assumption. Therefore, provided $|A| \in \mathcal{O}(\text{poly}(n))$ the absolute value of the log policy gradient $\left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2$ falls within $\mathcal{O}(\text{poly}(n))$, with $b \in \Omega(\frac{1}{\text{poly}(n)})$. If each parameterized block forms a local 2-design, the variance $\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \in \mathcal{O}(\frac{1}{\alpha^n})$ for $\alpha > 1$. Thus, in this setting the overall variance vanishes exponentially with the number of qubits.

However, outside of $\text{poly}(n)$ actions, the polynomial clipping assumption no longer holds. In such cases, the probability of a given action can be exponentially small, leading to a concentration with the number of qubits [16]. Thus, for $\beta > 1$, $\pi_{\min} \in \Omega(\frac{1}{\beta^n})$ and $\left|\partial_{\pi_\theta} f(\pi_\theta)\right|_\infty^2 \in \mathcal{O}(\beta^n)$. Consequently, the variance $\mathbb{V}_\theta \left[\partial_\theta \log \pi(a|s,\theta)\right]$ scales as $\mathcal{O}\left(\left(\frac{\beta}{\alpha}\right)^n\right)$ and the upper bound becomes too lose since β could actually be greater than α which implies that the variance increases with the number of qubits. Nonetheless, the behavior analyzed above corresponds exactly to the behavior observed with a parity-like Born policies, since this policy is always composed of a n-local measurement. Thus, for $|A| \in \text{poly}(n)$ the policy is prone to BPs. Beyond $\text{poly}(n)$ actions, the exponentially small probabilities cause a phase transition from BPs to exploding gradients. However, the need for an exponentially large number of shots to accurately estimate the policy renders it untrainable for a large number of qubits and actions.

In stark contrast, the base case $|A| = 2$ in a contiguous-like Born policy hinges on single qubit measurements which implies right from the start, a markedly different trainability profile compared to the parity-like policy. It is expected that the contiguous-like policy becomes harder to train with an increase in the number of actions since that translates directly in the globality of the employed observables. However, there exists a window of trainability that is contingent on maintaining a relatively small number of actions. Specifically, when the number of actions equals $|A| = n$, the contiguous-like policy utilizes $\log(|A|)$-local measurements. In practical terms, this equates to measuring at most $\log(n)$ adjacent qubits, a scenario that is typically local and known to circumvent BPs [16]. The trainability window is described in lemma 3.3.

Lemma 3.3. *Consider a n-qubit contiguous-like Born policy $\pi(a|s,\theta)$ with $|A|$ actions as in definition 2.1. Then, if each block in the parameterized quantum circuit forms a local 2-design, the policy gradient variance is given by*

$$\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \in \Omega\left(\frac{1}{poly(n)}\right) \tag{16}$$

for $|A| \in \mathcal{O}(n)$ and depth $\mathcal{O}(\log(n))$. On the other hand, the policy gradient variance scales as

$$\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \in \Omega\left(2^{-poly(\log(n))}\right) \tag{17}$$

for $|A| \in \mathcal{O}(n)$ and depth $\mathcal{O}(poly\log(n))$.

The detailed proof of this lemma can be found in appendix B. It provides a lower bound for the variance of the policy gradient for contiguous-like Born policies, under the condition that each parameterized block in the circuit constitutes a local 2-design. For $|A| \in \mathcal{O}(n)$, the variance declines at most polynomially with the number of qubits, and the number of quantum circuit executions needed for accurate policy and gradient estimation does not grow exponentially, thanks to $\log(n)$-local measurement. Hence, the policy remains trainable up to a depth of $\mathcal{O}(\log(n))$. Thus, in general, for $|A|$ within $\mathcal{O}(\text{poly}(n))$, the variance diminishes more rapidly than polynomially but less so than exponentially with the number of qubits, due to the observables' locality being greater than $\log(n)$ but less than n.

3.4. Analysis of the Fisher Information spectrum

In the realm of computational learning theory, the FIM is instrumental for evaluating the information garnered from a parameterized statistical model. In RL, the FIM must account for states sampled from the distribution generated under the policy, our parameterized model. Defining $\pi(a|s,\theta)$ as the parameterized policy and d_s^π as the distribution of states under the policy, the FIM is formulated as the expected value of the outer product of the gradient of the log-likelihood function:

$$\mathcal{I}\left(\theta\right) = \mathbb{E}_{s\sim d_s^\pi} \mathbb{E}_{a\sim\pi(.|s,\theta)} \left[\nabla_\theta \log \pi\left(a|s,\theta\right) \nabla_\theta \log \pi\left(a|s,\theta\right)^T\right]. \tag{18}$$

The FIM efficiently indicates how changes in parameters impact the model's output. Notably, the FIM spectrum is crucial for characterizing the BP phenomenon in PQC-based statistical models employing log-likelihood loss functions [1] which in turn can also be used to characterize BPs in the RL paradigm. However, in RL, where the loss function is influenced by cumulative rewards, the FIM does not consider this weighting. Still, assuming non-zero rewards, the FIM spectrum remains a valuable tool for BP characterization. In a BP, the eigenvalues on the FIM will be exponentially concentrated around zero [1]. The expected value of a diagonal entry k of the FIM can be written as,

$$\begin{aligned}\mathbb{E}_\theta \left[\mathcal{I}_{kk}\left(\theta\right)\right] &= \mathbb{E}_\theta \left[\left(\partial_{\theta_k} \log \pi\left(a|s,\theta\right)\right)^2\right] \\ &= \mathbb{V}_\theta \left[\partial_{\theta_k} \log \pi\left(a|s,\theta\right)\right] + \left(\mathbb{E}_\theta \left[\partial_{\theta_k} \log \pi\left(a|s,\theta\right)\right]\right)^2 \end{aligned} \tag{A}$$

where (A) is obtained from the definition of the variance. As a result, the FIM's diagonal entry can be lower bounded by the variance of the log likelihood:

$$\mathbb{E}_\theta \left[\mathcal{I}_{kk}\left(\theta\right)\right] \geqslant \mathbb{V}_\theta \left[\partial_{\theta_k} \log \pi\left(a|s,\theta\right)\right]. \tag{19}$$

This implies that the trace of the FIM, being the sum of its eigenvalues, has a lower bound as follows: $\text{Tr}\left[\mathbb{E}_\theta[\mathcal{I}(\theta)]\right] \geqslant \sum_{k=0}^{K-1} \mathbb{V}\theta \left[\partial\theta_k \log\pi(a|s,\theta)\right]$ for $\theta \in \mathbb{R}^K$. Therefore, the lower bound presented in lemma 3.3 can be directly used to identify the conditions under which the FIM spectrum indicates a BP for PQC-based policies. In a BP scenario, each entry of the FIM will vanish exponentially with the number of qubits, necessitating an exponentially increasing number of measurements to estimate each entry of the FIM

accurately. Based on lemma 3.3, for a Contiguous-like Born policy with $|A| \in \mathcal{O}(\text{poly}(n))$, the variance vanishes at most polylogarithmically with the number of qubits, affecting also the eigenvalues of the FIM as indicated by equation (19). Consequently, the FIM spectrum in such cases fails to indicate a BP. Conversely, for a Parity-like Born policy with $|A| \in \mathcal{O}(\text{poly}(n))$, the variance shrinks exponentially with the number of qubits, as do the eigenvalues of the FIM highlighting a BP. In situations where the number of actions exceeds $\text{poly}(n)$, not only the number of required measurements for accurate policy estimation are prohibitively large, but also the probabilities associated with actions remain exponentially small. This implies that, despite avoiding BPs, these scenarios are more likely to encounter exploding gradients rather than BPs, reflected in increasing FIM entries and a less concentrated spectrum around zero. Therefore, it is crucial to note that in cases where the number of actions is large, the FIM spectrum may not indicate BPs, but significant trainability issues can still arise from the necessity of a large number of measurements and the possibility of exploding gradients.

3.5. Summary

In this section, we characterized the trainability landscape of PQC-based policies. We demonstrated that under the realistic condition of a polynomial number of measurements, the variance of the log policy gradient for a contiguous-like Born policy decays at most polylogarithmically with the number of qubits, while for a parity-like Born policy, the variance vanishes exponentially with the number of qubits, provided the action space is also of polynomial size. Outside of the regime of polynomially sized action spaces, both policies becomes become untrainable since exponentially small probabilities associated with actions arise, which induce an exploding gradient on the log policy gradient objective function. Such phenomena was also captured and characterized through the inspection of the Fisher Information spectrum. A spectrum highly concentrated around zero indicates a BP. However, we highlight that in a regime where the number of actions is outside of $\mathcal{O}(\text{poly}(n))$, the FIM spectrum will not faithfully indicate a trainability problem. In such cases since probabilities are exponentially small, the FIM spectrum will be less concentrated around zero, moving away from BPs. Nevertheless, the gradient explodes at the same time as an exponentially large number of measurements is needed to accurately estimate both the policy and gradient.

4. Numerical experiments

In this section, we conduct an in-depth evaluation of the trainability issues in quantum policy gradients, as posited in lemma 3.3, through experimental validation in two distinct scenarios:

- *Trainability issues using a simplified 2-design*—We resort to the simplified two-design ansatz, as described in [6] and depicted in figure 3(a). This task involves examining the variance of the log likelihood partial derivative with respect to the type of policy and the range of available actions. Additionally, we investigate the FIM spectrum for both policies, focusing on how it varies with the number of actions.
- *Multi armed bandits*—To assess the practical performance of both Born policy formulations in an RL context, we designed a synthetic multi-armed bandit environment. The objective here is to evaluate the policies' ability to discern the optimal arm through actions sampling.

For the first task, the chosen two-design ansatz, while not strictly a two-design, it has been previously demonstrated to encounter cost-function BPs [6]. This setup is particularly advantageous for simulation purposes, especially with a large number of qubits and increased depth. We opted for a depth of $\mathcal{O}(n^2)$ in our experiments. Given the scarcity of large-scale RL problems amenable to efficient resolution via PQC-based policies, and more critically, those employing a sufficient number of qubits to investigate trainability in policy gradients, we selected the multi-armed bandit for the second task. This choice allows us to maintain the same objective function while affording the flexibility to adjust the total number of qubits, thus facilitating a thorough examination of trainability issues. It's noteworthy that all experiments were conducted using Pennylane's quantum simulator [3], with gradient estimations carried out via Parameter-shift rules [17] and considering $\mathcal{O}(\text{poly}(n))$ number of measurements.

4.1. Trainability issues using a simplified 2-design

In section 3.3, we meticulously examine the conditions under which trainability issues emerge in the training of learning agents using quantum policy gradients, focusing on the type of policy employed and the range of actions available. This assessment hinges on the globality of the observable, and for this investigation, we adopt the simplified two-design ansatz referenced in [6]. To align closely with a RL setting, we incorporate the initial rotation set from the ansatz, shown in purple in figure 4, to encode the agent's state. In this

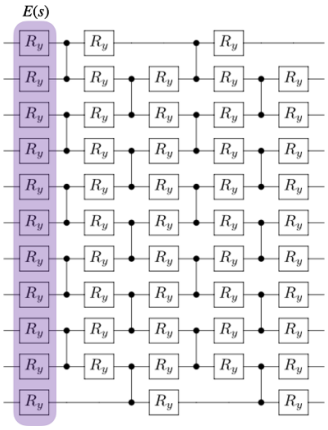

Figure 4. Simplified two-design ansatz with the first set of rotations in purple representing the unitary $E(s)$ responsible for the encoding of the state of the agent, s.

artificial setup, each of the agent's n state features is encoded using n qubits through angle-encoding. Both the agent's state, s, and the trainable parameters, θ, are uniformly sampled from the interval $\{s, \theta\} \sim U(-\pi, \pi)$.

Mimicking an RL environment, we execute the PQC a polynomial number of times to construct the policy from the measurements. An action is then sampled from the resulting policy distribution, and the probability of the selected action is used to estimate the variance of the log likelihood's partial derivatives. This variance is evaluated as a function of the number of actions and qubits, with the number of qubits, n, considered within the finite set $n = \{4, 6, 8, 10, 12, 14\}$. The action set, A, varies as a function of the number of qubits, defined as $A = \{2^i | i \in \mathbb{Z}, 1 \leqslant i \leqslant n\}$. Considering the assumption that the minimum probability is bounded polynomially with n, a scenario typical in RL where actions are sampled from the policy's probability distribution, the probability cannot be zero but may approach it closely. For this reason, $\pi_{\min} \in \Omega(\frac{1}{\text{poly}(n)})$ is assumed, provided the total number of actions is also polynomially large ($|A| \in \mathcal{O}(\text{poly}(n))$). This assumption acts as a practical clipping of probabilities, as often employed in RL contexts [21]. However, as the number of qubits increases, probabilities become exponentially small, requiring an analysis of the log-likelihood cost function's variance under both clipped and unclipped probability scenarios. With probability clipping, $\pi_{\min}$ is set to be within $\Omega(\frac{1}{n^2})$. Each experiment is repeated and averaged over 2000 iterations with a randomly generated parameter set. Regarding the FIM spectrum analysis, the same PQC is used under similar conditions. The only difference is in the number of executions, reduced to 10 to manage the computational cost of computing the FIM.

4.1.1. Contiguous-like Born policy

Firstly, we explore the variance of the log-likelihood cost function for the contiguous-like Born policy. Figures 6 and 5 display this variance as a function of the number of actions, considering both clipped and unclipped probabilities. In the unclipped probability scenario (figure 5(a)), the variance increases with the number of actions due to the $\log(|A|)$-local nature of the contiguous-like Born policy. However, despite the local nature of the measurements, as we increase the number of actions, probabilities diminish, leading to an escalation in the gradient. Figure 5(c) reveals a semi-logarithmic plot for the variance illustrating a polynomial decay in variance with the number of qubits, confirming lemma 3.3's predictions. Additionally, figure 5(b) demonstrates that the variance surges dramatically with an increase in actions, signaling the onset of exploding gradient phenomena. With a large number of qubits and actions, the gradients become unmanageable since the probabilities become exponentially small and thus a polynomial number of measurements will not be sufficient to perform learning. For that matter, consider a polynomial clipping of probabilities, illustrated in figure 6. In this setting, it can be observed from figure 6(a) a different behavior for the scaling of the variance as a function of the number of actions, compared to the originally non-clipped probabilities. Firstly, for a small number of actions, the variance increases in the same way as before. However, at some point the variance starts decreasing heavily. Such behavior can be explained from the polynomial clipping considered. Recall that in this experimental setup $\pi_{\min} \in \Omega(\frac{1}{\text{poly}(n)})$, which means that once the the number of actions increases, the probabilities decrease and eventually at some point every of the probability vector will have the same value which corresponds to the clipping. That is the reason for the decrease in variance with the number of actions since changing the parameters leads eventually to no further change in the output probabilities. With respect to the variance as a function of the number of qubits, it was predicted in lemma 3.3 that for a polynomial number of actions $|A| \in \mathcal{O}(\text{poly(n)})$ the variance should decay

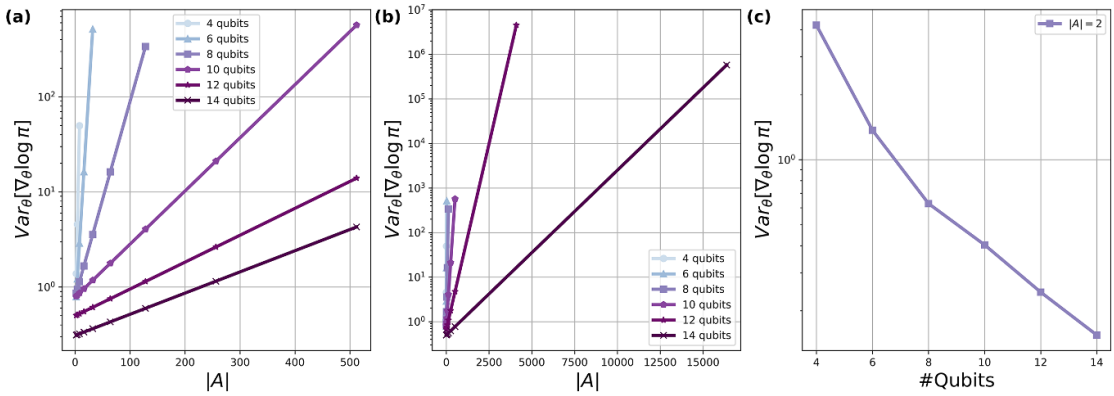

Figure 5. Variance of the log policy gradient for contiguous-like Born policies: (a) and (b) as a function of $|A|$ and (c) semi-logarithmic plot for varying number of qubits.

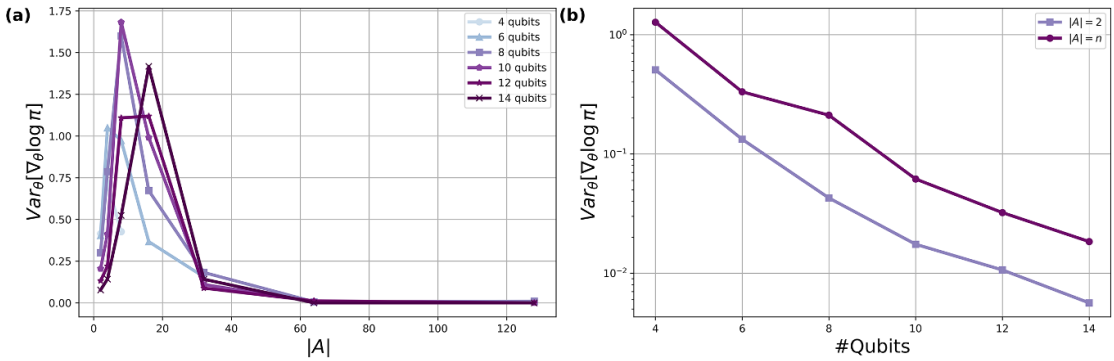

Figure 6. Variance of the log policy gradient for contiguous-like Born policies as (a) a function of $|A|$ and (b) for varying number of qubits with polynomial clipping.

at most polylogarithmically with the number of qubits. Moreover, if the number of actions $|A| = n$ this correponds exactly to measuring $\log(n)$ qubits and thus it was predicted that the variance should also decrease at most polynomially with the number of qubits. Such prediction is indeed confirmed through the scaling of the variance presented in the semi-logarithmic plot of figure 6(b) which illustrates the variance for a polynomially fixed number of actions.

4.1.2. Parity-like Born policy

Let us now do a similar analysis for the case dealing with a parity-like Born policy. Recall that in this setting, the policy is obtained always from a global measurement. Thus, it was predicted in section 3 that the upper bound for the variance already would be exponentially vanishing providing the number of actions is at most polynomial $|A| \in \mathcal{O}(\text{poly(n)})$ proving that the Parity-like policy indeed suffers from the BP phenomenon. Such conclusion came from the fact that probabilities can also be considered as polynomially vanishing with the number of qubits. Outside a polynomial number of actions, the probabilities would be too small for the polynomial assumption to be valid and thus the upper bound would lose its meaning. As predicted, the variance for this policy, illustrated in figure 7(a) for non-clipped probabilities, increases with the number of actions. Furthermore, figure 7(c) shows that for the base case of $|A| = 2$, illustrated in semi-logarithmic plot, that the variance diminishes exponentially with the number of qubits, confirming the susceptibility of the parity-like Born policy to BPs even with a minimal number of actions. However, as depicted in figure 7(b), the variance escalates for a fixed number of actions, deviating from a vanishing trend with an increase in qubit numbers, thus hinting at exploding gradients rather than BPs. Let us now analyze the scenario in which the probabilities are clipped. In figure 8(a), we observe a similar trend in variance as a function of the number of actions for parity-like policies with polynomially clipped probabilities. The variance initially increases with the number of actions but then undergoes a sharp decline, due to the equalization of probability values due to clipping. This behavior aligns with the hypothesis that clipping leads to uniform probabilities at higher action counts, reducing the variance. Figure 8(b) confirms the exponential decrease in

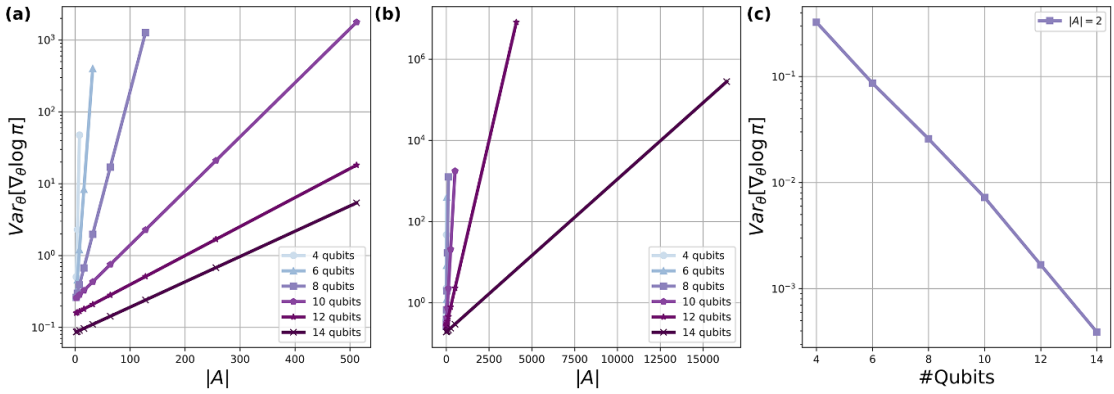

Figure 7. Variance of the log policy gradient for parity-like Born policies: (a) and (b) as a function of $|A|$ and (c) in a semi-logarithmic plot for varying number of qubits.

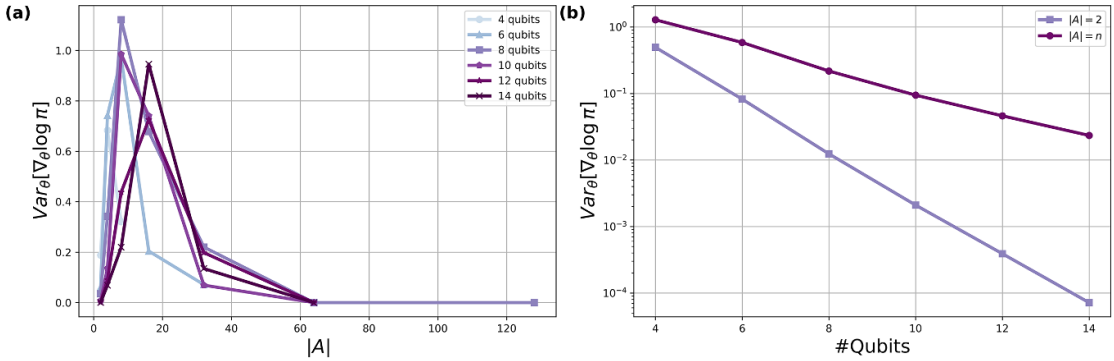

Figure 8. Variance of the log policy gradient for parity-like Born policies as (a) a function of $|A|$ and (b) for varying number of qubits with polynomial clipping.

variance with the number of qubits for polynomially large actions, consistent with the theoretical predictions made in section 3.

4.1.3. Analysis of the FIM spectrum

The FIM spectrum analysis, as discussed in section 3.4, is crucial for identifying trainability issues in different policy types. Focusing on the parity-like policy first, if this policy is prone to BPs with a polynomial number of actions, as inferred from previous sections, then the FIM entries should diminish exponentially with the number of qubits. This would confirm the presence of a BP, characterized by eigenvalues concentrated around zero. Conversely, beyond a polynomial number of actions and without probability clipping, the variance and subsequently the FIM entries would increase, leading to a less concentrated eigenvalue spectrum around zero, indicating the absence of a BP. These predicted behaviors are indeed evident in the FIM spectrum for the parity-like policy, as shown in figure 9.

In figure 9(a) it can be observed that for $|A| = 2$ eigenvalues concentrate in zero as the number of qubits increase, confirming the presence of a BP. Moreover, in figure 9(b) it can be observed the opposite behavior for $|A| = 2^n$. The eigenvalues are moving away from zero as the number of qubits increase, thus confirming that outside a polynomial number of actions, since the variance increases with the number of qubits the FIM spectrum does not indicate the presence of a BP.

Let us now inspect the spectrum of the FIM for the contiguous-like policy. The contiguous-like policy exhibits a distinct variance trend compared to the parity-like policy, especially for a smaller number of polynomial actions. Unlike the parity-like policy, this variance decreases polynomially with an increase in the number of qubits, suggesting the absence of a BP. Consequently, the FIM spectrum is expected to show eigenvalues concentrating around zero as the number of qubits increases, albeit more gradually than in the parity-like policy. However, beyond polynomial action numbers, similar to the parity-like policy, the eigenvalues will deviate from zero as the number of qubits increases due to the rising variance. Figure 10 corroborates these predictions. Particularly, figure 10(a) illustrates the eigenvalue concentration around zero

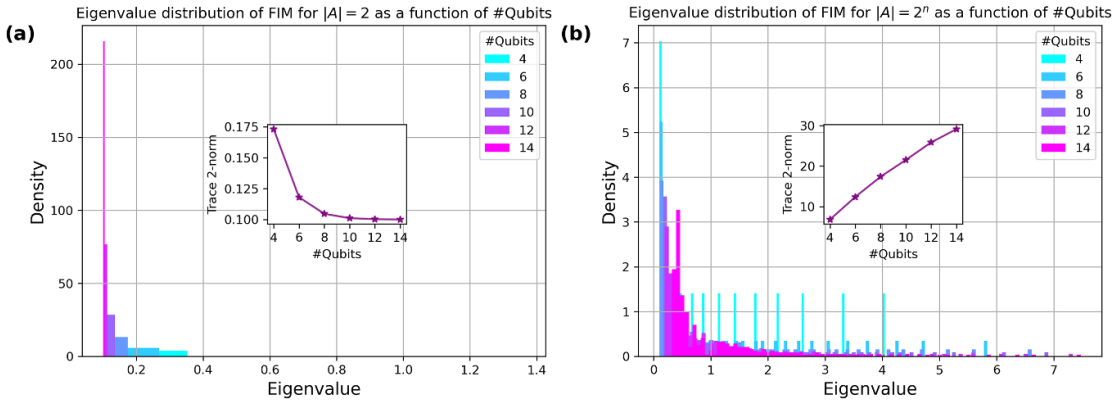

Figure 9. Eigenvalue distribution for the FIM as a function of the number of qubits for the parity-like Born policy in the extreme cases of $|A| = 2$ and $|A| = 2^n$.

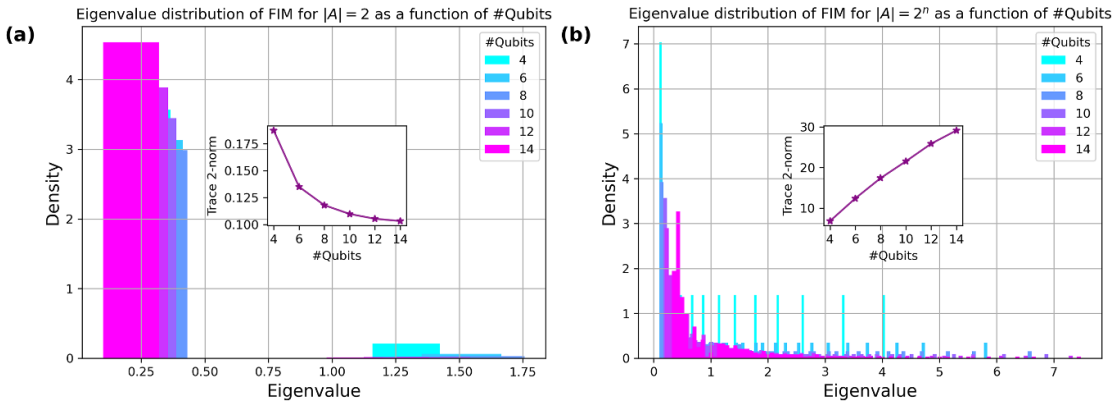

Figure 10. Eigenvalue distribution for the FIM as a function of the number of qubits for the contiguous-like Born policy in the extreme cases of $|A| = 2$ and $|A| = 2^n$.

for increasing qubits, but with a noticeably lower density compared to the parity-like case shown in figure 9(a).

4.2. Multi armed bandits

This section discusses trainability in the context of PQC-based policies, particularly what quantum systems with a substantial number of qubits, and evaluates their effectiveness in learning optimal actions in a reward-based context akin to RL. We consider a multi-armed bandit environment featuring a simple linear reward function: for each arm a, the deterministic reward $R(a)$ is given by $R(a) = 2a$. This setup allows us to scrutinize the learning capabilities of PQC-based policies with respect to the number of available actions. The PQC-based policy architecture we examine comprises a single layer of σ_z and σ_y single-qubit rotations, followed by an all-to-all CZ entanglement pattern. For these experiments, we use 16 qubits ($n = 16$) to gauge the impact of PQC depth on trainability in the scenario we have a contiguous or a parity-like policy. We analyze two scenarios: (1) a bandit environment with $|A| = n$ arms, which results in a contiguous-like policy that is $\log(n)$-local, and (2) a bandit environment with $|A| = 2^{n-4}$ arms, leading to a contiguous-like policy involving measurements over a polynomial number of qubits. The performance of contiguous-like policies considered in those scenarios is compared with the global parity-like policy with the same number of qubits. Each scenario incorporates a polynomial number of measurements. Gradient estimation is conducted using parameter-shift rules.

We assess the performance of the PQC-based agent by tracking the probability of choosing the best arm over a fixed number of episodes. An episode in this bandit environment involves performing a single action, collecting the associated reward, and using this information to update the PQC-based policy parameters via gradient-based methods. We conduct 100 episodes, comprising 100 action-steps, and average the probability of selecting the best arm over 50 different trials with randomly selected parameters. Figure 11 presents the outcomes when $|A| = n$ arms. In subfigures 11(a) and (b), the probability of choosing the best arm is depicted for contiguous-like and parity-like policies, respectively. The contiguous-like policy generally

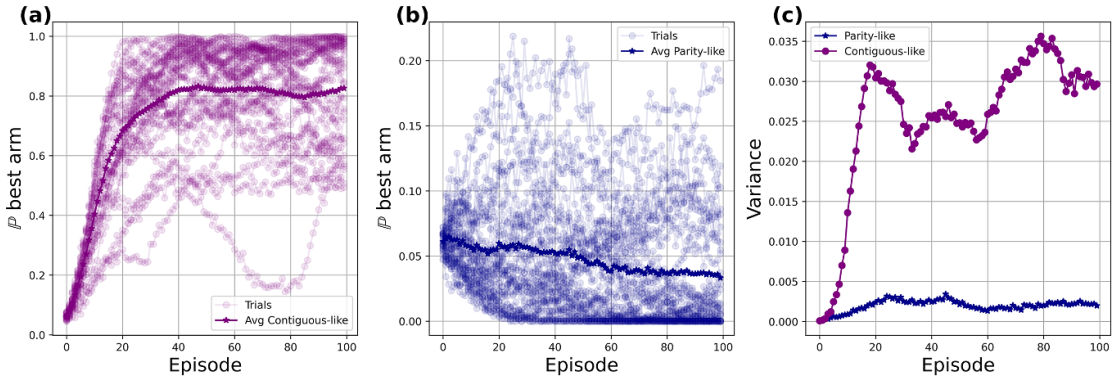

Figure 11. Experimental results for the Multi armed bandit environment with $|A| = n$ actions. In (a) and (b) the probability of selecting the best arm is illustrated for the contiguous-like and parity-like policies, respectively. (c) Illustrates the variance of the log policy gradient for both policies.

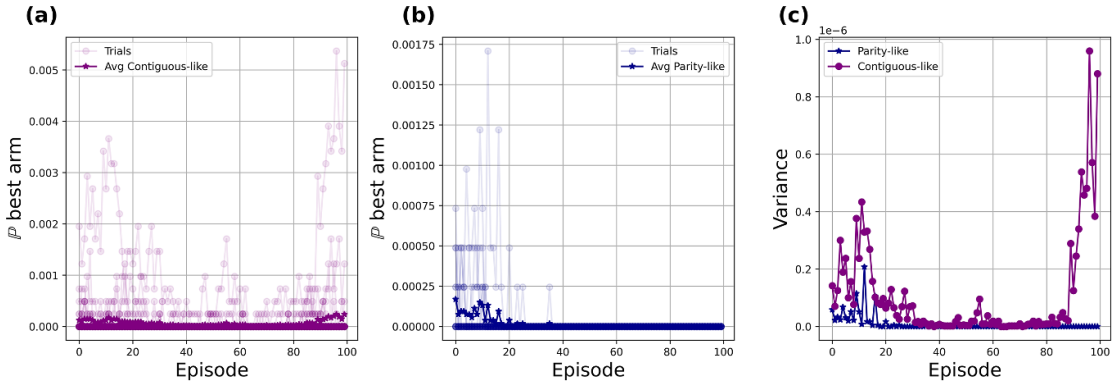

Figure 12. Experimental results for the Multi armed bandit environment with $|A| = 2^{n-4}$ actions. In (a) and (b) the probability of selecting the best arm is illustrated for the contiguous-like and parity-like policies, respectively. (c) illustrates the variance of the log policy gradient for both policies.

achieves a selection probability above 0.8 for the best arm, with some instances reaching a deterministic policy (probability of 1). The parity-like policy, however, struggles to exceed a 0.5 selection probability throughout training. This disparity can be attributed to the differing localities of each policy. Since $|A| = n$, the contiguous-like policy involves $\log(n)$-local measurements, contrasting with the parity-like policy's n-local approach. The effect of these differing observables on trainability is further illuminated by examining the variance of the log policy gradient, as shown in figure 11(c). While the variance remains relatively low for both policies, it is notably smaller and close to zero for the parity-like policy. Figure 12 depicts results for the bandit environment with $|A| = 2^{n-4}$ arms. Subfigures 12(a) and (b) depict the probability of selecting the best arm over a series of episodes for both contiguous-like and parity-like policies. In this scenario, neither policy demonstrates the ability to learn the optimal arm effectively, with probabilities of selecting the best arm consistently below one percent. This outcome is explained by examining the employed observables. The parity-like policy is always globally measured, but now, with more arms, the probabilities associated with each arm are significantly reduced. Furthermore, the contiguous-like policy, which previously measured $\log(n)$ qubits, engages in a measurement over a polylog number of qubits. Despite not experiencing exponential decay in variance, the large number of qubits and actions places the contiguous-like policy in a BP, akin to the parity-like policy. This is further evidenced by the variance of the gradient, as shown in figure 12(c). The variance for both policies is minimal, indicating an inability to learn the optimal policy. Thus, a polynomial number of measurements, as utilized in these experiments, proves insufficient for learning the optimal policy in such complex settings.

5. Conclusion

In conclusion, our research provides pivotal insights into the trainability of PQC-based policies in the realm of policy-based RL. A significant aspect of our findings concerns the trainability challenges faced by two specific types of policies: the Contiguous-like and Parity-like Born policies. These challenges manifest in two

distinct forms: the occurrence of standard BPs characterized by exponentially diminishing gradients, and the potential for gradient explosion.

The nature and extent of these challenges are influenced by two key factors: the specific type of Born policy used, and the interplay between the number of qubits and the action-space size. Notably, our study reveals that with n qubits, when a polynomial number of measurements ($\mathcal{O}(\text{poly}(n))$) is considered, the Contiguous-like Born policy demonstrates trainable regions at a logarithmic depth ($\mathcal{O}(\log(n))$), provided that the action-space remains within polynomial bounds ($\mathcal{O}(\text{poly}(n))$). This observation is crucial as it delineates a specific scenario where this policy remains effective and trainable. In contrast, under similar conditions, the Parity-like Born policy consistently exhibits a BP, indicating its inherent limitations in certain settings. Furthermore, we noticed a striking shift in gradient behavior for actions beyond polynomial size. In these instances, the gradients transition from diminishing to exploding, attributed to exceedingly low probabilities. This shift renders the polynomial number of measurements inadequate for differentiating actions, thus presenting a significant challenge to the practical application of these policies.

We would like to emphasize that part of the gradient behavior observed throughout this work is due to classical post-processing. Recall that in section 3.1 we analyzed the variance for the trivial scenario of product states. We concluded that product states would not lead to trainability issues in policy gradient optimization. This is given by the logarithm post-processing of the probability of basis states, which separates the product of individual qubit contributions into a sum. However, for entangled states, the product factorization is no longer true, and indeed, as the number of qubits increases, the probabilities potentially become exponentially small. The logarithm post-processing, in turn, makes the gradient explode. Thus, the classical-post processing indeed impacts what we observe. We highlight that such behavior is also expected to manifest in the classical policy gradient algorithm. Indeed, for very small probabilities, the gradient would also explode. This is one reason advanced policy gradient algorithms such as PPO [18] are much more stable to train. The crucial difference compared with PQC-based policies stems from the fact that the probabilities derived from quantum systems will eventually concentrate given more expressivity and depth of the circuit [16], leading to other sorts of optimization problems that we do not know to be as severe in the classical setting. In addition, we would also like to stress that the trainability analysis presented in this work neglected the effect of the reward. Indeed, it was assumed the presence of a maximum reward to simplify the policy gradient REINFORCE objective expressed in equation (2) to an expression dependent only on the policy. Nevertheless, in practice, there are problems such as the sparsity of the reward. For instance, environments where the reward is assigned only at a goal state and no reward in the middle. In such scenarios, small or even null rewards would lead to vanishing gradients and, indeed, hide the behavior of the quantum system. For that matter, we did not consider these cases.

In [7], the authors showed that the BP phenomenon results from a curse of dimensionality and that the cases where one can impose trainability guarantees also lead to classically simulable models. We suspect that this is most likely the case for PQC-based policies since we are still considering hardware-efficient ansatze, and under the measurement conditions outlined through this work, it would fall under the same category explored in [7]. However, it raises the question of whether there are other types of ansatze that strike a perfect balance between trainability and non-efficient classical simulation, combined with specific PQC-based optimization that could be used to circumvent this issue.

There are several other promising directions for future work. For instance, one should address the trainability issues associated with softmax policies [10, 19], where the freedom to measure the expectation values of $|A|$ different observables presents an intriguing avenue for exploration since more sophisticated results in the trainability of PQCs [15] can be considered.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/andre-sequeira10/Trainability-issues-in-QPGs.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020 (DOI 10.54499/UIDB/50014/2020). This work is financed by National Funds through FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project IBEX, with reference PTDC/CCI-COM/4280/2021 (DOI 10.54499/PTDC/CCI-COM/4280/2021).

Appendix A. Upper bound on the return

Let $R_{\max}$ be the maximum possible reward at any time step. Thus, the return is upper bounded by:

$$G(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1} \leqslant R_{\max} \sum_{t=0}^{T-1} \gamma^t = R_{\max} \frac{\gamma^T - 1}{\gamma - 1}. \tag{20}$$

This enables the following upper bound on the return per time step

$$\sum_{t=0}^{T-1} G_t(\tau) \leqslant R_{\max} \sum_{t=0}^{T-1} \frac{\gamma^{T-t} - 1}{(\gamma - 1)} \leqslant R_{\max} \frac{T}{(\gamma - 1)^2}. \tag{21}$$

Appendix B. Proof of lemma 3.3

Lemma 3.3. *Consider a n-qubit contiguous-like Born policy $\pi(a|s,\theta)$ with $|A|$ actions as in definition 2.1. Then, if each block in the parameterized quantum circuit forms a local 2-design, the policy gradient variance is given by*

$$\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \in \Omega\left(\frac{1}{poly(n)}\right) \tag{16}$$

for $|A| \in \mathcal{O}(n)$ and depth $\mathcal{O}(\log(n))$. On the other hand, the policy gradient variance scales as

$$\mathbb{V}_\theta \left[\partial_\theta \log \pi\left(a|s,\theta\right)\right] \in \Omega\left(2^{-poly(\log(n))}\right) \tag{17}$$

for $|A| \in \mathcal{O}(n)$ and depth $\mathcal{O}(poly\log(n))$.

Proof. Let us start with the expansion of the standard expression of the variance. For the sake of simplicity let $\pi_\theta = \pi(a|s,\theta)$ and the partial derivative $\partial_\theta \log \pi_\theta = \frac{\partial_\theta \pi_\theta}{\pi_\theta}$

$$\begin{aligned}
\mathbb{V}_\theta \left[\partial_\theta \log \pi_\theta\right] &= \mathbb{E}_\theta \left[\left(\frac{\partial_\theta \pi_\theta}{\pi_\theta}\right)^2\right] - \mathbb{E}_\theta \left[\frac{\partial_\theta \pi_\theta}{\pi_\theta}\right]^2 \\
&\geqslant \mathbb{E}_\theta \left[\left(\partial_\theta \pi_\theta\right)^2\right] \mathbb{E}_\theta \left[\frac{1}{\pi_\theta^2}\right] - \mathbb{V}_\theta \left[\left(\partial_\theta \pi_\theta\right)^2\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta^2}\right] - \mathbb{E}_\theta \left[\frac{\partial_\theta \pi_\theta}{\pi_\theta}\right]^2 && \text{(A)} \\
&\geqslant \mathbb{E}_\theta \left[\left(\partial_\theta \pi_\theta\right)^2\right] \mathbb{E}_\theta \left[\frac{1}{\pi_\theta^2}\right] - \mathbb{V}_\theta \left[\left(\partial_\theta \pi_\theta\right)^2\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta^2}\right] - \mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta}\right] && \text{(B)} \\
&= \mathbb{E}_\theta \left[\left(\partial_\theta \pi_\theta\right)^2\right] \mathbb{E}_\theta \left[\frac{1}{\pi_\theta^2}\right] - \underbrace{\left(\mathbb{V}_\theta \left[\left(\partial_\theta \pi_\theta\right)^2\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta^2}\right] + \mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta}\right]\right)}_{(a)}
\end{aligned}$$

where (A) is obtained from the lower bound of the expectation value of the product of two non-negative random variables $\mathbb{E}_\theta[XY] \geqslant \mathbb{E}_\theta[X]\mathbb{E}_\theta[Y] - \mathbb{V}_\theta[X]\mathbb{V}_\theta[Y]$ and (B) from the upper bound of the variance of the product of two random variables via Cauchy-Schwarz $\mathbb{V}_\theta[XY] \leqslant \sqrt{\mathbb{V}_\theta[X]\mathbb{V}_\theta[Y]}$ [21]. The variance is lower bounded taking the upper bound of (a) that can be simplified to:

$$\begin{aligned}
(a) &\leqslant \left(2\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \left|\partial_\theta \pi_\theta\right|_{\max}^2 + 2\mathbb{E}_\theta \left[\partial_\theta \pi_\theta\right] \mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right]\right) \left|\frac{1}{\pi_\theta^2}\right|_{\max} + \mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta}\right] && \text{(A)} \\
&\leqslant \frac{1}{2}\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \left|\frac{1}{\pi_\theta^2}\right|_{\max} + \mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \mathbb{V}_\theta \left[\frac{1}{\pi_\theta}\right] && \text{(B)} \\
&\leqslant \frac{3}{2}\mathbb{V}_\theta \left[\partial_\theta \pi_\theta\right] \left|\frac{1}{\pi_\theta^2}\right|_{\max} && \text{(C)}
\end{aligned}$$

where (A) is obtained from the upper bound of the variance pf the product of two random variables, (B) from the assumption that either parameterized block before/after θ forms a 1-design and thus $\mathbb{E}_\theta[\partial_\theta \pi_\theta] = 0$ and (C) from the upper bound on the variance $\mathbb{V}_\theta[\frac{1}{\pi_\theta}] \leqslant |\frac{1}{\pi_\theta^2}|_{\max}$.

The lower bound on the variance of the policy gradient can thus be further simplified to:

$$\begin{aligned}
\mathbb{V}_\theta\left[\partial_\theta \log \pi_\theta\right] &\geqslant \mathbb{E}_\theta\left[\left(\partial_\theta \pi_\theta\right)^2\right]\mathbb{E}_\theta\left[\frac{1}{\pi_\theta^2}\right] - \frac{3}{2}\mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]\left|\frac{1}{\pi_\theta^2}\right|_{\max} \\
&\geqslant \left(\mathbb{E}_\theta\left[\partial_\theta \pi_\theta\right]^2 - \mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]^2\right)\left(\mathbb{E}_\theta\left[\frac{1}{\pi_\theta}\right]^2 - \mathbb{V}_\theta\left[\frac{1}{\pi_\theta}\right]^2\right) - \frac{3}{2}\mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]\left|\frac{1}{\pi_\theta}\right|_{\max}^2 \qquad \text{(A)} \\
&= \mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]^2 \mathbb{V}_\theta\left[\frac{1}{\pi_\theta}\right]^2 - \left(\mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]^2 \mathbb{E}_\theta\left[\frac{1}{\pi_\theta}\right]^2 + \frac{3}{2}\mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]\left|\frac{1}{\pi_\theta}\right|_{\max}^2\right) \qquad \text{(B)} \\
&\geqslant \mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]^2 \mathbb{V}_\theta\left[\frac{1}{\pi_\theta}\right]^2 - 3\mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]^2 \mathbb{E}_\theta\left[\frac{1}{\pi_\theta}\right]^2 \qquad \text{(C)} \\
&= \mathbb{V}_\theta\left[\partial_\theta \pi_\theta\right]^2 \underbrace{\left(\mathbb{V}_\theta\left[\frac{1}{\pi_\theta}\right]^2 - 3\mathbb{E}_\theta\left[\frac{1}{\pi_\theta}\right]^2\right)}_{(a)} \qquad \text{(D)}
\end{aligned}$$

where (A) is obtained from the lower bound of the expectation value of the product of two non-negative random variables, (B) from the assumption that either parameterized block before/after θ forms a 1-design and thus $\mathbb{E}_\theta[\partial_\theta \pi_\theta] = 0$ and reorganizing terms and (C) from the upper bound on the expectation value and joining terms.

Since the variance is non-negative it implies that $(a) \geqslant 0$. Therefore the variance will be lower bounded depending on the number of actions and corresponding globality of the observable. For $|A| \in \mathcal{O}(n)$, $\mathbb{V}_\theta[\partial_\theta \pi_\theta]^2 \in \Omega(\frac{1}{\text{poly}(n)}^2)$ for $\mathcal{O}(\log(n))$ depth. It decays polynomially with the number of qubits since we are measuring $\log(n)$ (adjacent) qubits [6]. Moreover, $(a) \leqslant \text{poly}(n)^2$. Thus, the overall variance deacay at most polynomially with the number of qubits. When the number of actions $|A| \in \mathcal{O}(\text{poly}(n))$, $\mathbb{V}_\theta[\partial_\theta \pi_\theta]^2 \in \Omega(2^{-\text{poly}(\log(n))^2})$. It decays faster than polynomially but slower than exponentially since we are measuring $\log(\text{poly}(n))$ qubits [6]. In this case $(a) \leqslant 2^{\text{poly}(\log(n))^2}$ since we have $|A| \in \text{poly}(n)$. Therefore the overall variance decay at most polylogarithmically with the number of qubits. Thus, completing the proof. □

ORCID iD

André Sequeira https://orcid.org/0000-0002-6659-9277

References

[1] Abbas A, Sutter D, Zoufal C, Lucchi A, Figalli A and Woerner S 2021 The power of quantum neural networks *Nat. Comput. Sci.* **1** 403–9

[2] Arrasmith A, Cerezo M, Czarnik P, Cincio L and Coles P J 2021 Effect of barren plateaus on gradient-free optimization *Quantum* **5** 558

[3] Bergholm V *et al* 2022 PennyLane: automatic differentiation of hybrid quantum-classical computations (arXiv:1811.04968v4 [quant-ph])

[4] Biamonte J 2021 Universal variational quantum computation *Phys. Rev.* A **103** L030401

[5] Cerezo M *et al* 2021 Variational quantum algorithms *Nat. Rev. Phys.* **3** 625–44

[6] Cerezo M, Sone A, Volkoff T, Cincio L and Coles P J 2021 Cost function dependent barren plateaus in shallow parametrized quantum circuits *Nat. Commun.* **12** 1791

[7] Cerezo M *et al* 2024 Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing (arXiv:2312.09121 [quant-ph])

[8] Chen S Y C, Huck Yang C-H, Qi J, Chen P-Y, Ma X and Goan H-S 2020 Variational quantum circuits for deep reinforcement learning *IEEE Access* **8** 141007–24

[9] Cherrat E A *et al* 2023 Quantum deep hedging (arXiv:2303.16585 [quant-ph])

[10] Jerbi S, Gyurik C, Marshall S, Briegel H J and Dunjko V 2021 Variational quantum policies for reinforcement learning (arXiv:2103.05577)

[11] Jerbi S, Cornelissen A, Ozols Māris and Dunjko V 2022 Quantum policy gradient algorithms (arXiv:2212.09328 [quant-ph])

[12] Leone L, Oliviero S F E, Cincio L and Cerezo M 2022 On the practical usefulness of the hardware efficient ansatz *Phys. Rev. Lett.* **128** 050402

[13] McClean J R, Boixo S, Smelyanskiy V N, Babbush R and Neven H 2018 Barren plateaus in quantum neural network training landscapes *Nat. Commun.* **9** 4812

[14] Meyer N, Scherer D D, Plinge A, Mutschler C and Hartmann M J 2023 Quantum policy gradient algorithm with optimized action decoding (arXiv:2212.06663 [quant-ph])

[15] Ragone M, Bakalov B N, Sauvage F'eric, Kemper A F, Ortiz Marrero C, Larocca M and Cerezo M 2023 A unified theory of barren plateaus for deep parametrized quantum circuits (arXiv:2309.09342 [quant-ph])

[16] Rudolph M S, Lerch S, Thanasilp S, Kiss O, Vallecorsa S, Grossi M and Holmes Z 2023 Trainability barriers and opportunities in quantum generative modeling

[17] Schuld M, Bergholm V, Gogolin C, Izaac J and Killoran N 2019 Evaluating analytic gradients on quantum hardware *Phys. Rev.* A **99** 032331

[18] Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O 2017 Proximal policy optimization algorithms (arXiv:1707.06347 [cs.LG])
[19] Sequeira A, Paulo Santos L and Soares Barbosa L 2023 Policy gradients using variational quantum circuits *Quantum Mach. Intell.* **5** 18
[20] Skolik A, Jerbi S and Dunjko V 2022 Quantum agents in the gym: a variational quantum algorithm for deep q-learning *Quantum* **6** 720
[21] Thanasilp S, Wang S, Nghiem N A, Coles P J and Cerezo M 2021 Subtleties in the trainability of quantum machine learning models (arXiv:2110.14753 [quant-ph])
[22] Uvarov A V and Biamonte J D 2021 On barren plateaus and cost function locality in variational quantum algorithms *J. Phys. A: Math. Theor.* **54** 245301
[23] Williams R J 1992 Simple statistical gradient-following algorithms for connectionist reinforcement learning *Mach. Learn.* **8** 229–56