Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/TQE.2020.DOI

On Quantum Natural 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 (HASLab), INESC TEC, Braga, Portugal
[3]Quantum Linear and Optical Computation group, International Nanotechnology Laboratory (INL), Braga, Portugal

Corresponding author: André Sequeira (email: andresequeira401@gmail.com).

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within grants LA/P/0063/2020, UI/BD/152698/2022 and project IBEX, with reference PTDC/CC1-COM/4280/2021

ABSTRACT This research delves into the role of the quantum Fisher Information Matrix (FIM) in enhancing the performance of Parameterized Quantum Circuit (PQC)-based reinforcement learning agents. While previous studies have highlighted the effectiveness of PQC-based policies preconditioned with the quantum FIM in contextual bandits, its impact in broader reinforcement learning contexts, such as Markov Decision Processes, is less clear. Through a detailed analysis of Löwner inequalities between quantum and classical FIMs, this study uncovers the nuanced distinctions and implications of using each type of FIM. Our results indicate that a PQC-based agent using the quantum FIM without additional insights typically incurs a larger approximation error and does not guarantee improved performance compared to the classical FIM. Empirical evaluations in classic control benchmarks suggest even though quantum FIM preconditioning outperforms standard gradient ascent, in general it is not superior to classical FIM preconditioning.

INDEX TERMS Quantum Policy Gradients, Fisher Information, Quantum Reinforcement Learning, Natural Policy Gradients

I. INTRODUCTION

Reinforcement Learning (RL) emerged as a pivotal technology in modern artificial intelligence, driving progress in diverse fields Subramanian et al. [2022], Sutton and Barto [1998]. Deep RL, in particular, exceeded human performance in complex tasks, demonstrating its efficacy in games like Atari, Go, and No-limit poker, among others. The integration of RL and Deep Neural Networks (DNNs) placed positioning RL at the forefront of AI for complex sequential tasks in uncertain environments Russell and Norvig [2020]. RL's strength lies in its ability to allow software agents to adapt and optimize decision-making in unknown environments. This versatility has led to significant applications in healthcare, autonomous driving, and robotics Arulkumaran et al. [2017].

In the landscape of RL, the inception of the Natural Policy Gradient (NPG) algorithm Kakade [2001] marks a seminal advancement. This algorithm addresses stability and sample complexity issues – identified as intrinsic shortcomings of traditional policy gradient methods Sutton et al. [1999], Williams [1992]. Specifically, NPG enhances the stability of policy gradient methods by preconditioning the gradient with the inverse of the Fisher Information Matrix (FIM), facilitating updates directly in the policy space and thereby emerging as a highly sample-efficient RL algorithm Agarwal et al. [2021]. However, the efficacy of NPG is still tethered to the curse of dimensionality, exacerbated by the estimation and inversion of the FIM. This limitation catalyzed the evolution of various NPG derivatives, including Trust Region Policy Optimization (TRPO) Schulman et al. [2017a] and Proximal Policy Optimization (PPO) Schulman et al. [2017b], which have been crucial in advancing Deep RL.

Quantum RL models, employing Parameterized Quantum Circuits (PQCs) demonstrated empirically superior sample complexity in addressing fully visible environments compared to a subset of conventional DNNs, as evidenced in standard classical control benchmarking scenarios Chen et al. [2020], Jerbi et al. [2021], Meyer et al. [2023b], Sequeira et al. [2023], Skolik et al. [2022]. In Cherrat et al. [2023], the authors elucidated that certain PQC-based policies, composed of compound layers, are devoid of barren plateaus, rendering them conducive for financial tasks such as hedging, where deep RL proves instrumental in real market frameworks. Moreover, a quadratic separation in gradient estimation between classical and quantum RL models, provided oracle access to environmental dynamics, was established in Jerbi et al. [2022]. In Meyer et al. [2023a], the authors demonstrate empirically that a PQC-based agent doing gradient updates preconditioned by quantum FIM, has better performance compared to standard euclidean updates. Despite

these strides, a number of questions remain: Can the sample complexity of PQC-based policies be surely improved by employing quantum natural gradients Stokes et al. [2020]? What is the actual role of the quantum FIM? This paper aims at contributing to address these questions through exploiting well-known Löwner inequalities Meyer [2021] between the classical and quantum FIM and its impact in the regret of a PQC-based agent. These questions pivots on the potential of quantum NPG as a possible alternative to the classical NPG algorithm, with the prospect of significantly impacting practical applications. This is particular relevant in quantum control Niu et al. [2019], in which the transition from classical to quantum natural gradients opens a perspective of exploration, potentially harboring enhanced algorithmic stability and sample complexity, thus elevating the robustness and efficiency of RL frameworks.

RELATED WORK

In Meyer et al. [2023a] it was empirically demonstrated within the contextual bandits framework that PQC-based policies, performing gradient updates preconditioned by the quantum FIM, exhibit enhanced sample complexity and training stability in comparison to standard Euclidean updates. However, the efficacy of quantum natural policy gradients in broader RL domains beyond contextual bandits, particularly in conventional Markov Decision Processes, remains unexplored. Furthermore, a comprehensive understanding of the quantum FIM's role, as juxtaposed with the classical FIM employed in the original NPG algorithm Kakade [2001], is yet to be attained. Given the distinct nature of these two information matrices, a pivotal question emerges, which becomes crucial to our investigation:

Does a PQC-based agent accrue tangible benefits from employing updates in state-space with the quantum FIM as opposed to updates in policy-space with the classical FIM?

CONTRIBUTIONS

This paper seeks to elucidate the aforementioned query by harnessing well-established Löwner inequalities between the two information matrices Meyer [2021]. Subsequently, we delineate inequalities concerning the regret of PQC-based agents employing natural gradients preconditioned by both the classical and quantum FIMs. In summary, our main contributions are:

- In the absence of additional insights regarding the nature of the information matrices, a PQC-based agent using the quantum FIM will have a large approximation error compared to the classical FIM and in general not assuring an enhanced regret and thus poorer sample complexity.
- If the square root of the information matrices is considered rather than the conventional inverse, the larger approximation error mentioned above could be compensated. However, this does not inherently imply the attainment of the optimal policy.
- The performance of PQC-based policies resorting to natural gradients was empirically examined in standard classic control benchmarking environments Sutton and Barto [1998], with gradient preconditioning using 1) the inverse and 2) the square root inverse of the information matrices. It was not observed a substantial improvement when considering the quantum FIM inverse. However, if the square root inverse is employed, the quantum FIM provides an improved sample complexity compared to the square root of classical FIM preconditioning. This indicates that in this setting the matrix compensates for the approximation error.
- Sample complexity analysis for the estimation of both quantum and classical FIM, indicates that the quantum FIM is independent of the total number of actions of a given environment, as opposed to the classical FIM. This may be interesting in large action spaces, where samples are expensive to obtain.

Section II provides a comprehensive introduction to policy gradient methods and elaborates on the PQC-based policies under consideration. Section III forms the crux of this paper, introducing the QNPG algorithm and discussing key lemmas pertaining to the significance of the quantum FIM in NPG optimization. Section IV details the experimental framework and shares the findings from these experiments. The paper concludes with Section VI, where we summarize our findings and explore potential avenues for future research.

II. QUANTUM POLICY GRADIENTS

Policy Gradients aim to learn a parameterized probability distribution over actions given states, a *policy* denoted as $\pi(a|s,\theta)$, where $\theta \in \mathbb{R}^k$ represents the parameter vector of size k, $s \in S$ denotes the state and $a \in A$ the action. The main goal is to perform gradient ascent on a performance metric $J(\theta)$:

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

The REINFORCE algorithm Williams [1992] is the simplest policy gradient algorithm, that estimates the gradient of samples obtained from N trajectories of length T—also known as the horizon - under the parameterized policy, as in Equation (2).

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

where $b(s_{t_i})$ is an action-independent control variate also known as baseline that is subtracted from the return, resulting in a variance reduction. In this work, the average return was considered as the baseline, computed by Equation (3).

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

In the sequel, the policy $\pi(a|s,\theta)$ shall be regarded as a PQC-based policy, i.e. the policy is being generated from the output of measurements of PQC's. Specifically, two formulations of such a policy be: the *Born policy* (Definition II.1) and the *softmax-policy* (Definition II.2).

Definition II.1. Let $s \in \mathcal{S}$ be a state embedded in an n-qubit parameterized quantum state, $\psi(s,\theta)$, where $\theta \in \mathbb{R}^k$. The probability associated to a given action $a \in \mathcal{A}$ in the Born framework 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 the Hilbert space is partitioned into $|A|$ disjoint subspaces spanned by the computational basis states and P_a is a projection into the subspace associated with action a.

Definition II.1 presents the most general definition Born policy. However, there could be partitions that do not take into account every eigenstate. In these scenarios, the probability associated to a given action would not be normalized as before since $\sum_{a\in A} P_a \neq I$. In this setting the policy would need to further normalized as follows:

$$\pi(a|s,\theta) = \frac{\langle P_a \rangle_{s,\theta}}{\sum_{a'\in\mathcal{A}} \langle P_{a'} \rangle_{s,\theta}} \tag{5}$$

such that $\sum_{a\in A} \pi(a|s,\theta) = 1$.

Definition II.2. Let $s \in \mathcal{S}$ be a state embedded in an n-qubit parameterized quantum state, $\psi(s,\theta)$, where $\theta \in \mathbb{R}^k$. Let O_a be an arbitrary observable representing the numerical preference of action $a \in \mathcal{A}$ and β the inverse temperature hyperparameter. The probability associated to a given action a in a softmax policy is given by:

$$\pi(a|s,\theta) = \frac{e^{\beta\langle O_a\rangle_{s,\theta}}}{\sum_{a'} e^{\beta\langle O'_a\rangle_{s,\theta}}} \tag{6}$$

$\mathcal{O}(|\mathcal{A}|)$ different observables may be used to attribute the action's numerical preference.

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 Schuld et al. [2019] can then 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, for gates with two eigenvalues, it corresponds to:

$$\frac{\partial\langle O\rangle_\theta}{\partial\theta_l} = \frac{1}{2}\left[\langle O\rangle_{\theta+\frac{\pi}{2}e_l} - \langle O\rangle_{\theta-\frac{\pi}{2}e_l}\right] \tag{7}$$

where e_l indicates that the parameter θ_l is being shifted. The equality indicates that the partial derivative can be obtained using two quantum circuit evaluations. Thus, for $\theta \in \mathbb{R}^k$, the gradient can be estimated ideally using $2k$ total quantum circuit evaluations. However, it is known that the expectation value itself can be estimated up to additive error $\mathcal{O}(\epsilon^{-2})$ Schuld and Petruccione [2021]. Thus, $\mathcal{O}(2k\epsilon^{-2})$ quantum circuit calls are needed. For arbitrary functions of expectation values like the log policy gradient, the gradient can be obtained via standard chain rule. For the softmax policy, the log policy gradient takes a peculiar form expressed as a centered version of the gradient of the expectation values encoding the numerical preference of each action Jerbi et al. [2021]:

$$\nabla_\theta \log\pi(a|s,\theta) = \beta\left[\nabla_\theta\langle O_a\rangle_{s,\theta} - \sum_{a'\in A}\pi(a'|s,\theta)\nabla_\theta\langle O_{a'}\rangle_{s,\theta}\right] \tag{8}$$

III. NATURAL GRADIENTS IN POLICY OPTIMIZATION

This section introduces the QNPG algorithm and delineates its theoretical advantages over the conventional classical NPG. Initially, we discuss the classical NPG algorithm and analyze the regret associated with smooth policies Agarwal et al. [2021]. Subsequently, we propose a reformulation that incorporates the quantum FIM. The derivation of the regret bound for the QNPG algorithm is then grounded in established Löwner inequalities, which compare the classical and quantum FIMs, as detailed in Meyer [2021].

NATURAL POLICY GRADIENTS

The Natural Policy Gradient algorithm (NPG) Kakade [2001] is a rescaled version of the policy gradient that performs gradient updates in the geometry induced by the information matrix associated to the policy, the *Fisher Information matrix* (FIM) as follows:

$$\theta^{t+1} \leftarrow \theta^t + \eta F^{-1}\nabla_\theta V^{\pi_\theta}(\rho) \tag{9}$$

where F is the average FIM on the sampled states and actions under policy π_θ as follows:

$$F = \mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi(\cdot|s,\theta)}\left[\nabla_\theta\log\pi(a|s,\theta)\nabla_\theta\log\pi(a|s,\theta)^T\right] \tag{10}$$

where d^{π_θ} is the distribution of states generated under policy π_θ. Notice that F is positive-definite i.e., $F > 0$, however in practice due to instabilities in approximating the information matrix, the inverse F^{-1} is replaced by the Moore-Penrose pseudoinverse $F^\dagger$ and regularization is often considered. The notion of *regret* is often considered in RL algorithms as a measure of the difference between the policy being followed and an hypothetical optimal policy. Specifically, regret is computed as the difference between the expected reward of an optimal policy and the reward garnered by the agent's policy over a specified number of episodes or time steps as follows,

$$\sum_{t=1}^{T}\left(V^*(s_t) - V^\pi(s_t)\right) \tag{11}$$

where $V^*(s_t)$ denotes the value function under the optimal policy for state s_t at time t, and $V^\pi(s_t)$ denotes the value

function under the policy π employed by the agent. In Agarwal et al. [2021] the authors established a regret bound for the NPG algorithm considering a general class of smooth parameterized policies. The regret lemma is restated below for completeness.

where $\tilde{d}$ is the distribution of states generated under the comparison policy $\tilde{\pi}$. $\left\|w^{(t)}\right\|_2$ is the norm of the vector resulting of the multiplication of the inverse of the classical FIM, F, by the gradient vector, $w^{(t)} = F^{-1}\nabla_\theta \log \pi^{(t)}(a|s,\theta)$. ϵ_t is the approximation error at time step t derived from *compatible function approximation* Sutton et al. [1999]. Lemma III.1 can thus be utilized in the context of PQC-based policies should these policies respect smoothness conditions. Recall that a function $f : \mathbb{R}^k \mapsto \mathbb{R}$ is β-smooth if for all $(x, x') \in \mathbb{R}^k$ Agarwal et al. [2021]:

$$\|\nabla f(x) - \nabla f(x')\|_2 \leq \beta \|x - x'\|_2 \tag{14}$$

The smoothness of both Born and Softmax policies is established in Jerbi et al. [2022] through the Gevrey condition. Since $\pi(a|s,\theta) \in [0,1]$ and in the context of RL, where the action is being sampled from the policies probability distribution, it implies that $\pi \in (0,1]$.

QUANTUM NATURAL POLICY GRADIENTS

The Quantum Natural Policy Gradient algorithm (QNPG) is obtained by replacing the classical FIM with the QFIM, here represented as $\mathcal{F}$. Restricting ourselves to pure quantum states, the QFIM takes the common form Meyer [2021]:

$$\mathcal{F}_{ij} = 4\mathrm{Re}\left[\langle\partial_{\theta_i}\psi|\partial_{\theta_j}\psi\rangle - \langle\partial_{\theta_i}\psi|\psi\rangle\langle\psi|\partial_{\theta_j}\psi\rangle\right] \tag{15}$$

In the context of machine learning, a data-dependent QFIM is needed. Therefore, considering again d^{π_θ} as the distribution of states generated under parameterized policy, the data-dependent QFIM becomes:

$$\begin{aligned}\mathcal{F}_{ij} = \mathbb{E}_{s\sim d^{\pi_\theta}} 4\mathrm{Re}\big[&\langle\partial_{\theta_i}\psi(s,\theta)|\partial_{\theta_j}\psi(s,\theta)\rangle - \\ &\langle\partial_{\theta_i}\psi(s,\theta)|\psi(s,\theta)\rangle\langle\psi(s,\theta)|\partial_{\theta_j}\psi(s,\theta)\rangle\big]\end{aligned} \tag{16}$$

where d^{π_θ} is the distribution of states generated under policy π_θ. Notice that in practice, the empirical QFIM is thus obtained from a finite set of states in a trajectory T, obtained under policy π_θ. It is crucial to understand the differences between the FIM and QFIM. Since they are information matrices, they capture what happens in the neighbourhood of a parameter θ of a given parameterized model by a distance measure. Their difference resorts to what distances are considered within the two different spaces. FIM considers the distance between probability distributions i.e., policies in the context of RL. Thus, the FIM gives information about how the policy changes when infinitesimal changes are performed on a parameter. QFIM, on the other hand, considers distances in the space of quantum states. Thus, it gives information on how the parameterized quantum state changes, given a slight variation of a parameter.

QFIM AS A METRIC FOR POLICY OPTIMIZATION

At first glance, one should say that the FIM is more relevant for policy optimization since it captures changes directly in the policy space. However, even though the QFIM is not actually capturing information in the policy space it could be of independent interest since the policy in our case is derived from the quantum state itself. The use of QFIM in policy gradients can be understood as having different impact depending on the type of PQC-based policy employed. For that matter, consider the Softmax policy as presented in Definition II.2. In its most general form it is comprised of $\mathcal{O}(|A|)$ different expectation values encoding numerical preferences. This makes building the connection between QFIM and expectation value of observables a non-trivial and non-intuitive task. On the other hand, the Born policy (Definition II.1) is derived from projective measurements. Recall that QFIM is derived from the fidelity distance between quantum states Meyer [2021]. Thus, there is an intricate connection between QFIM and the Born policy. For that reason let us start with the Born policy.

Consider a normalized Born policy $\pi(a|s,\theta) = \langle P_a\rangle_{s,\theta}$ as defined in Definition II.1. Let w.l.g V_a be the set of eigenstates associated with action a. The policy can be expressed as:

$$\begin{aligned}\pi(a|s,\theta) &= \sum_{v\in V_a} \langle\psi(s,\theta)|v\rangle\langle v|\psi(s,\theta)\rangle \\ &= \sum_{v\in V_a} |\langle v|\psi(s,\theta)\rangle|^2\end{aligned} \tag{17}$$

Recall that QFIM is a metric that describes changes in state space under variation of θ Haug and Kim [2021] which means that:

$$|\langle\psi(s,\theta)|\psi(s,\theta+\delta)\rangle|^2 = 1 - \frac{1}{4}\mathcal{F}_{ij}\delta_i\delta_j \tag{18}$$

This has a clear impact on policy optimization since the policy is captured in the same way as projectors onto a partition of basis states. More importantly recall that classical FIM corresponds to the information matrix associated to the probability distribution generated from the measurement of the quantum state where we could say that the measurement $\mathcal{M} = \{P_a\}$ where P_a is the a^{th} outcome of the experiment, corresponds to the partition of action a. In this setting, the following matrix inequality Meyer [2021] applies:

$$F \leq \mathcal{F} \tag{19}$$

Inequality (19) expresses the Löwner inequality of positive semi-definite matrices Bhatia [1997] i.e., $\mathcal{F} - F \geq 0$ has only non-negative eigenvalues. The inequality indicates that QFIM is always an upper bound for any information matrix obtained from the outcome of measurements in a parameterized quantum state. The equality happens once the parameterized quantum state prepares a classical probability

Lemma III.1 (NPG Regret Lemma Agarwal et al. [2021])**.** *Fix a comparison policy $\tilde{\pi}$ and a state distribution ρ. Assume for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ that $\log \pi(a|s,\theta)$ is a β-smooth function of θ. Consider $\pi^{(0)}$ the uniform distribution for every state and the sequence of weights $w^{(0)}, \ldots, w^{(T)}$ satisfying $\left\|w^{(t)}\right\|_2 \leq W$. Let ϵ_t be the approximation error at time t:*

$$\epsilon_t = \mathbb{E}_{s\sim\tilde{d}}\mathbb{E}_{a\sim\tilde{\pi}(\cdot|s,\theta)}\left[A^{(t)}(s,a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a|s,\theta)\right] \tag{12}$$

Then the regret at time step t is upper bounded by:

$$\min_{t<T}\left\{V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho)\right\} \leq \frac{1}{1-\gamma}\left(\frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta\beta W^2}{2} + \frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t\right) \tag{13}$$

distribution. The matrix inequality forms the basis for the separation in terms of agent's regret presented in this work.

The NPG objective is optimizing the policy under the log policy gradient, which is slightly different compared to standard quantum natural gradient objective. Nevertheless, recall that natural gradients indeed perform gradient updates using adaptive step sizes, and for that matter if we expand the log policy gradient as $\nabla\log(\pi) = \frac{\nabla\pi}{\pi}$, we could embed the resulting denominator into an adaptive learning rate $\eta' = \frac{\eta}{\pi}$, resulting in approximately the same objective as in standard quantum natural gradient. Thus, for the Born policy QFIM has a direct impact from the state space to the policy space.

For the Softmax policy though the situation is not so intuitive. Recall that to build the policy requires estimating the expectation values of $\mathcal{O}(|A|)$ observables. Therefore, Inequality (19) lost meaning in this scenario since the classical FIM would not be generated from the output of a fixed quantum measurement but from a distribution obtained from $\mathcal{O}(|A|)$ possibly different expectation values. However, recall that the softmax policy was originally considered to overcome the lack of greediness control in the Born policy Jerbi et al. [2021]. That is, at time step T for some environment we could already know everything about the reward function but nonetheless, the type of parametrization could for instance not allow for deterministic policies. For that reason, the Softmax policy is usually considered where an hyperparameter β control its greediness. In this setting, instead of considering the most general softmax formulation and have $\mathcal{O}(|A|)$ expectation values of operators in multiple bases, the observable could then simply be the projectors considered in the Born policy, i.e. $\langle O_a\rangle = \langle P_a\rangle$. In this scenario, expanding the log policy gradient leads to the NPG gradient update for the Softmax policy in Equation (20), where P_a is the projector into a partition of basis states V_a.

QFIM would then have the same impact in policy optimization, however taking into consideration every action as opposed to the Born policy, and modifying the update to take a centered version of the natural gradient into account. Even though Inequality (19) would not in principle apply in this scenario, we would expect such gradient update to be beneficial nonetheless in policy optimization. It remains to be seen in practice the actual role of the QFIM in policy optimization under the Softmax policy.

QFIM FOR IMPROVED REGRET

The NPG regret Lemma III.1 establishes that the regret of agent that uses an arbitrary and smooth parameterized policy is dependent on the vector norm $||w||_2$ and the compatible function approximation error ϵ_t. Thus, to establish bounds on the regret dependently on the information matrix employed, it would suffice to establish bounds on the norms and the approximation errors presented in the regret lemma, induced by those information matrices. Let us start with the norms. Let $||w_{\mathcal{F}}||_2$ and $||w_F||_2$ be the 2-norm induced by QFIM and classical FIM, respectively. The goal of this section is to clarify in which conditions we have the norm inequality

$$||w_{\mathcal{F}}||_2 \leq ||w_F||_2 \tag{21}$$

Thus indicating that the regret associated to PQC-based agent employing NPG optimization benefits from considering the QFIM as the metric instead of the classical FIM.

Let F and $\mathcal{F}$ be two positive semi-definite matrices such that $F \leq \mathcal{F}$ i.e., $\mathcal{F} - F \geq 0$ has only non-negative eigenvalues. Let $v = \nabla_\theta \log \pi_\theta(a|s,\theta)$. Let $||w_F||_2 = ||F^{-1}v||_2$ and $||w_{\mathcal{F}}||_2 = ||\mathcal{F}^{-1}v||_2$ be 2-norm induced by the FIM and QFIM, respectively. Thus:

$$F \leq \mathcal{F} \quad \not\Rightarrow \quad ||w_{\mathcal{F}}||_2 \leq ||w_F||_2 \tag{22}$$

for all $v \in \mathbb{R}^k$. That is, the matrix inequality does not readily imply the vector norm inequality for every gradient vector. Moreover, notice that we are considering positive semi-definite matrices, but the inequality actually considers the inverses and not the pseudoinverses. However, in practice, both QFIM and FIM are ill-conditioned and thus they need to be regularized before inversion i.e., $\mathcal{F} = \mathcal{F} + \epsilon I$ where I is the identity and $\epsilon > 0$ is the regularization term. For that reason, let us consider the inverses from now on. The Löwner partial order inequality guarantees the reverse inequality for the inverses of positive (semi-)definite matrices.

$$F \leq \mathcal{F} \quad \text{iff} \quad F^{-1} \geq \mathcal{F}^{-1} \tag{23}$$

Thus, the inequalities (23) can be used to establish the conditions for which the desired vector norm inequality in Equation (21) is reached.

From the definition of positive semi-definite matrices we have that for any vector $v \in \mathbb{R}^k$ the following applies:

$$v^T F v \geq 0 \quad \text{and} \quad v^T \mathcal{F} v \geq 0 \tag{24}$$

$$\eta\mathcal{F}^{-1}\nabla_\theta \log \pi(a|s,\theta) = \eta\beta\left[\mathcal{F}^{-1}\nabla_\theta\langle P_a\rangle_{s,\theta} - \sum_{a'\in A}\pi(a'|s,\theta)\mathcal{F}^{-1}\nabla_\theta\langle P_{a'}\rangle_{s,\theta}\right] \tag{20}$$

which implies that

$$F \leq \mathcal{F} \quad \Longrightarrow \quad \begin{cases} v^T F v \leq v^T \mathcal{F} v \\ v^T F^{-1} v \geq v^T \mathcal{F}^{-1} v \end{cases} \tag{25}$$

Recall that 2-norm $||Fv||_2^2 = \Big(Fv\Big)^T\Big(Fv\Big)$ and since in this case both information matrices are Hermitian ($F = F^T$) then,

$$||Fv||_2^2 = \Big(Fv\Big)^T\Big(Fv\Big) = v^T F^T F v = v^T F^2 v \tag{26}$$

since $F = F^T$. Thus, the vector norm inequality that we have been seeking implies the matrix norm inequality

$$||Fv||_2^2 \quad \Longrightarrow \quad F^2 \leq \mathcal{F}^2 \tag{27}$$

which is not guaranteed in general if the Löwner inequality $F \leq \mathcal{F}$ is all we have. That is, in general

$$F \leq \mathcal{F} \quad \not\Rightarrow \quad F^2 \leq \mathcal{F}^2 \tag{28}$$

The implication would be guaranteed if either F or $\mathcal{F}$ is idempotent i.e., $\{0, 1\}$ would be their only eigenvalues and the only non-singular matrix (full-rank) would be the identity. This restricts the set of information matrices and thus the set of PQCs needed for the vector norm inequality to be guaranteed. For instance, the PQC

$$|\psi(\theta)\rangle = \bigotimes_{i=1}^{n} \cos(\theta_i)|0\rangle + \sin(\theta_i)|1\rangle \tag{29}$$

has $\mathcal{F} = I$ which would apply. However, it is also true that in this case $F = \mathcal{F}$ thus entailing equality. Therefore, the desired norm inequality would not be in general guaranteed just from the matrix inequality of information matrices. However, notice that the expansion in Equation (26) can also be taken into account considering $F^{\frac{1}{2}}$ instead of F. Thus,

$$||F^{\frac{1}{2}}v||_2^2 = v\Big(F^{\frac{1}{2}}v\Big)^T\Big(F^{\frac{1}{2}}v\Big) = v^T {F^{\frac{1}{2}}}^T F^{\frac{1}{2}} v = v^T F v \tag{30}$$

and since $v^T F v \leq v^T \mathcal{F} v$ and $v^T F^{-1} v \geq v^T \mathcal{F}^{-1} v$ the vector norm inequality is guaranteed:

$$||F^{-\frac{1}{2}}v||_2^2 \geq ||\mathcal{F}^{-\frac{1}{2}}v||_2^2 \quad \Longleftrightarrow \quad F \leq \mathcal{F} \tag{31}$$

Therefore, a norm inequality depends on the type of information matrix inverse considered. In summary:

* $(F^{-1}, \mathcal{F}^{-1})$ - If the standard inverses are considered then the norm inequality is not in general guaranteed, since $F \leq \mathcal{F} \quad \not\Rightarrow \quad ||w_{\mathcal{F}}||_2 \leq ||w_F||_2$, and further information about these matrices is needed.
* $(F^{-\frac{1}{2}}, \mathcal{F}^{-\frac{1}{2}})$ - Norm inequality is guaranteed since $||F^{-\frac{1}{2}}v||_2^2 \geq ||\mathcal{F}^{-\frac{1}{2}}v||_2^2 \quad \Longleftrightarrow \quad F \leq \mathcal{F}$. However its actual utility in solving a RL problem is unknown.

This result motivates the use of a Generalized Quantum Natural Policy Gradient (GQNPG) algorithm, which for $\varphi \in [0, 1]$, the GNQPG algorithm performs the following update:

$$\theta^{t+1} \leftarrow \theta^t + \eta\mathcal{F}^{-\varphi}\nabla_\theta V^{\pi_\theta}(\rho) \tag{32}$$

In Haug and Kim [2021], the authors suggest a similar update for the standard gradient ascent considering QFIM as metric. The authors suggest that $\varphi = \frac{1}{2}$ constitutes a intricate optimization strategy. As previously described, the standard QFIM is usually ill-conditioned and requires to be regularized $\mathcal{F} = \mathcal{F} + \epsilon I$ where $\epsilon > 0$ could have a dramatic impact on sensitivity to parameter updates and lead to an increase in gradient steps to achieve the convergence of the algorithm. The authors show that, for $\varphi = \frac{1}{2}$, QFIM is intrinsically regularized and thus it is full-rank and does not need ϵ, once the fidelity cost-function is considered. The authors observed that for several PQCs the infidelity had a sharp increase for $\varphi \geq 0.6$ due to ill-conditioned QFIM. They suggest however that for small infidelities standard QFIM with $\epsilon = 0.1$ may perform better. However, in the context of policy gradients, it may be very well the case that appears in the beginning of training large infidelities i.e., the policy being far from the optimal policy are expected. Thus, the role of φ and the tradeoff between regularization and performance in the context of RL agents should also be addressed besides the standard preconditioning considered in the NPG algorithm.

The approximation error in the regret lemma of Section III depends on the type of information matrix employed. Same as before, the inequality between the classical and quantum information matrices imply an inequality between the approximation errors induced by these matrices. Recall that the approximation error at time step t, ϵ_t is defined as:

$$\epsilon_t = \mathbb{E}_{s\sim\tilde{d}}\mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)}\left[A^{(t)}(s,a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a|s,\theta)\right] \tag{33}$$

For simplicity, let $v = \nabla_\theta \log \pi^{(t)}(a \mid s)$ and $w^{(t)}$ be expanded as a function of the type of information matrix as before. Let ϵ_F and $\epsilon_{\mathcal{F}}$ be the approximation errors induced by the classical and quantum FIMs, respectively. Consider the difference between the approximation errors induced by the classical and quantum FIM,

$$\begin{aligned} \epsilon_{\mathcal{F}} - \epsilon_F &= -w_{\mathcal{F}} \cdot v + w_F \cdot v \\ &= -\mathcal{F}^{-1}v \cdot v + F^{-1}v \cdot v \\ &= -v^T\mathcal{F}^{-1}v + v^T F^{-1} v \\ &= v^T(F^{-1} - \mathcal{F}^{-1})v \geq 0 \end{aligned} \tag{34}$$

which implies that the approximation error under quantum FIM will always be greater than or equal to the approximation error under classical FIM:

$$\epsilon_{\mathcal{F}} \geq \epsilon_F \tag{35}$$

Therefore, for an agent employing the classical FIM for precondition the gradient will have a regret less than or equal to the regret of an agent employing the quantum FIM since the norm inequality is not guaranteed and the quantum FIM actually provides a greater approximation error. However, recall that the Löwner-Heinz inequality Zhan [2002] implies that:

$$I \leq \mathcal{F} \quad \Longrightarrow \quad I^{-\frac{1}{2}} \geq \mathcal{F}^{-\frac{1}{2}} \tag{36}$$

since for any $0 \leq r \leq 1$, $I^r \leq \mathcal{F}^r$. The approximation error of the square root of classical FIM is then also less than or equal to the square root of quantum FIM,

$$\epsilon_{\mathcal{F}^{\frac{1}{2}}} \geq \epsilon_{F^{\frac{1}{2}}} \tag{37}$$

Therefore, even though the approximation error persists, the regret can be compensated by the norm inequality using the square root of the information matrices. It remains to see in practice now, if the approximation error increases due to quantum FIM can actually be compensated by the norm inequality, since this depends heavily on the problem at hand. The results are summarized in Table 1.

$F/\mathcal{F}$	$\|\|w_{\mathcal{F}}\|\|_2 \leq \|\|w_F\|\|_2$	$\epsilon_{\mathcal{F}} \leq \epsilon_F$	**Improved regret**
$F^{-1}/\mathcal{F}^{-1}$	No	No	No
$F^{-\frac{1}{2}}/\mathcal{F}^{-\frac{1}{2}}$	Yes	No	?

TABLE 1. Summary of results. The first column indicates the type of information matrix considered. The second and third columns indicate wether the norm and approximation error inequalities are guaranteed, respectively. The fourth column indicates if the regret is improved.

IV. PERFORMANCE EVALUATION IN BENCHMARKING ENVIRONMENTS

In this section, we assess the efficacy of the GQNPG algorithm, as introduced in Section III, using two classical control benchmarking environments Sutton and Barto [1998]. We selected the Cartpole and Acrobot environments due to their compact state-action spaces, which have previously been efficiently addressed using PQC-based policies Jerbi et al. [2021].

The Cartpole environment features a four-dimensional state space with two potential actions, while the Acrobot environment has a six-dimensional state with three available actions. Notably, in the Acrobot environment, four of the features represent the sine and cosine values of the two joint angles. To optimize training time and reduce the PQC size, we limited the state representation to the angles, thus reducing it to four features. Consequently, both environments utilize the PQC depicted in Figure 1, as proposed by Jerbi et al. [2021], albeit with different layer configurations and measurement strategies. A comprehensive characterization of the environment and the PQC configurations can be found in Table 4 and Table 5, respectively. We investigated both Born

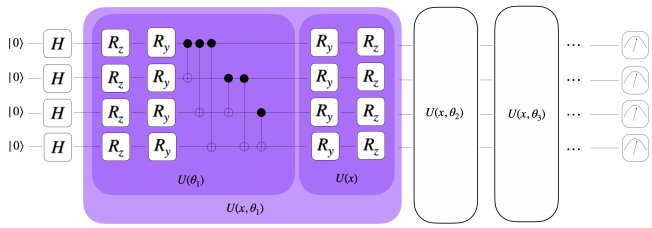

FIGURE 1. The parameterized quantum circuit used in the numerical experiments. Data reuploading is consistent with Jerbi et al. [2021], but input scaling was excluded to improve the estimation of the Quantum FIM matrices.

and Softmax PQC-based policies as discussed in Section II. Simple computational basis measurements were employed to link quantum measurements to their respective policies. For the Cartpole, a single-qubit projector was used. We perform a global measurement decomposition using an ancilla Meyer et al. [2023b] and associate each basis state to an action. For $a \in \{0, 1\}$, the Born policy is defined as:

$$\pi(a|s,\theta)_{\text{cartpole}} = \langle\psi(s,\theta)|a\rangle\langle a|\psi(s,\theta)\rangle \tag{38}$$

In contrast, for the Acrobot, a mod $-$ 3 Born policy was adopted. In this case, each qubit is measured in the computational basis. Let $b \in \{0,1\}^n$ be a n-qubit computational basis state and $\text{int}(b)$ its decimal representation. The basis state is associated with action a if $\text{int}(b) \mod 3 = a$. For $a \in \{0, 1, 2\}$, the Born policy is defined as:

$$\pi(a|s,\theta)_{\text{acrobot}} = \sum_{b\in\{0,1\}^n}^{\text{int}(b) \mod 3=a} \langle\psi(s,\theta)|b\rangle\langle b|\psi(s,\theta)\rangle \tag{39}$$

For the Softmax policy, while the same projectors as in the Born policy were employed, the probability serves as a numerical preference for a specific action. This preference is subsequently processed by the softmax function to yield a probability distribution over actions. It is important to note that the Softmax policy introduces an inverse temperature hyperparameter, β, which influences the policy's greediness, a feature absent in the Born policy. The optimal β value is environment-specific and typically identified through hyperparameter tuning. In our study, we adopted a linear annealing schedule for β, starting at 1 and culminating in the final β value as suggested in Jerbi et al. [2021].

Performance outcomes for five different optimizers in the Cartpole and Acrobot environments are depicted in Figures 2 and 3 respectively. The following optimizers were considered:

- **Adam**: Utilizes the standard Adam optimizer with a learning rate of 10^{-2}.
- **NPG**: Employs the standard NPG algorithm with classical FIM.
- **NPG** $\varphi = 0.5$: Uses the NPG algorithm with the square root of the classical FIM.

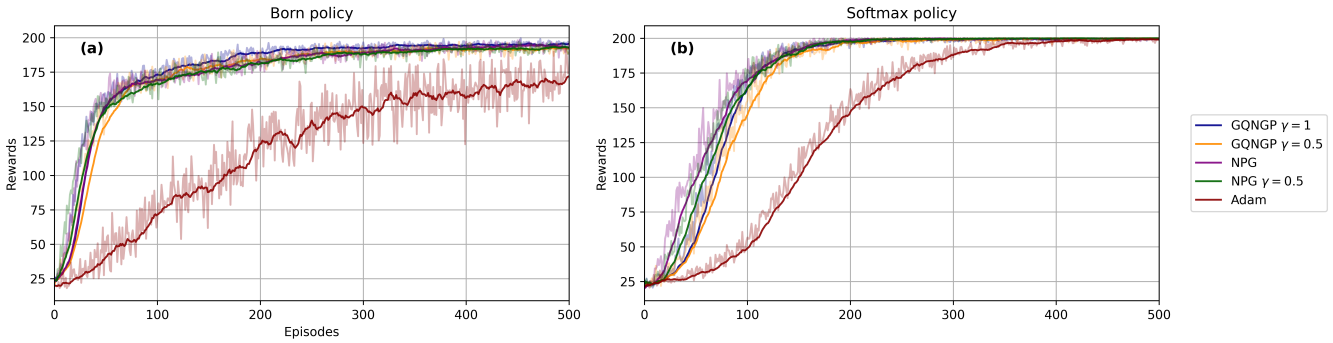

FIGURE 2. Performance of the NPG algorithm (and its generalized quantum counterpart) in the Cartpole environment. Subfigures **(a)** and **(b)** represent the performance of Born and Softmax policies using the cumulative reward as the evaluation metric.

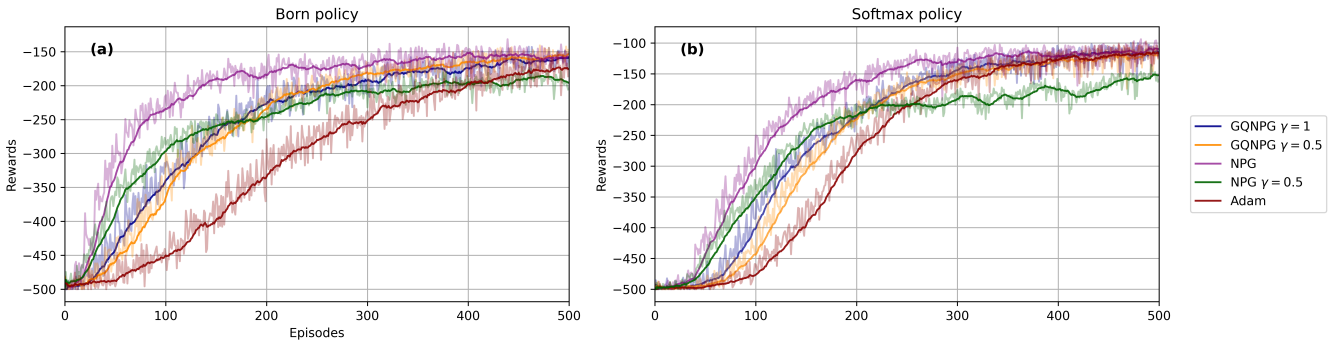

FIGURE 3. Performance of the NPG algorithm (and its generalized quantum counterpart) in the Acrobot environment. Subfigures **(a)** and **(b)** showcase the performance of Born and Softmax policies using cumulative reward as a performance measure.

- **GQNPG**: Integrates the NPG algorithm with quantum FIM.
- **GQNPG** $\varphi = 0.5$: Adopts the NPG algorithm with the square root of the quantum FIM.

The optimizers' performances were benchmarked using the cumulative reward metric, plotted on the y-axis, against the total episode count on the x-axis. Each optimizer's performance was averaged across 50 trials, with each figure displaying a 10-episode running mean and a shaded region representing the standard deviation of the experiments.

Given the deterministic nature of the environments, actions consistently lead to the same observed states and rewards. Furthermore, in accordance with the NPG regret lemma, we employed a zero-initialization approach concerning the parameters of the PQC to effectively have an uniform policy at the beginning of training. This means every parameter in the PQC illustrated in Figure 1 was initialized at zero, and since the employed PQC is composed by an initial chain of Hadamard gates, it ensures that policy is in fact uniform and moreover, the variance of the algorithm could solely be attributed to the agent's sampled trajectories.

Our experiments utilized Pennylane's quantum simulator Bergholm et al. [2022] with PyTorch-based automatic differentiation. For replication purposes, our work can be accessed through the following GitHub repository QNPG.

A direct comparison between the Born and Softmax policies for the Cartpole environment is available in Figures 2**(a)** and 2**(b)**. Notably, the Softmax agents exhibit superior and more consistent performance compared to their Born counterparts. This advantage is attributed to β and the ability to regulate the policy's greediness, an ability the Born policy lacks Jerbi et al. [2021]. Both policies demonstrate negligible performance variation across different optimizers. However, slight advantages for the GQNPG algorithm in the Born policy could be observed, although these may be a result of statistical variances. A key observation is that gradient preconditioning, regardless of using quantum or classical FIMs, yields similar results. Such result indicate that in this context, updates in state space could be as effective as updates in policy space. That is, the quantum FIM obtained from infinitesimal distances between quantum states is as efficient as the classical FIM which is obtained from infinitesimal distances between policies directly.

Figures 3**(a)** and 3**(b)** depict the performance registered in the Acrobot environment for the Born and Softmax policies, respectively. The Acrobot environment with three actions and a slightly more complex reward function becomes a more complex environment to be solved compared to Cartpole. Such complexity difference implies a more clear separation in optimizer performance compared to the Cartpole. It can be immediately observed that in this case, for both policies, not every variant of natural optimizers performed better than

the standard Adam. However, there is a clearer separation between the performance associated to the classical NPG and GQNPG optimizers. In this setting, the classical NPG has more evidently better convergence even though for this environment there is not a clear condition in which the environment is considered solved. Thus, the asymptotic behavior is used here to attribute that the classical NPG algorithm necessitates slightly fewer episodes to reach an asymptote in the cumulative reward. Moreover, both optimizers seem to agree in the same policy after 500 episodes. Furthermore, it is more clear as well that in the Born policy, the GQNPG algorithm with $\varphi = 0.5$ performs better than the NPG algorithm with $\varphi = 0.5$. In this setting, however, the same conclusion can be reached for the Softmax policy even though the matrix inequalities can not be guaranteed, observed as before. It is curious to observe that in this scenario, the unregularized NPG with $\varphi = 0.5$ after a great learning period of around 200 episodes seems to saturate and perform worse than the Adam optimizer.

The results obtained experimentally shined a light at the need to test PQC-based policies with different natural optimizers in even more complex environments characterized by multiple state-action spaces and reward functions to be able to further conclude about the efficacy of quantum FIM based natural policy gradient algorithms.

V. COMPARATIVE ANALYSIS FOR THE ESTIMATION OF INFORMATION MATRICES

In this section, we draw a comparison in terms of the resources needed to compute quantum and classical FIM's. The chosen metric to characterize the resources is the number of quantum measurements or quantum circuit executions required to estimate the information matrices. This way, a sample complexity analysis can be made and a possible separation between the two natural gradients assessed. Sample complexity in this context has a specific meaning. It corresponds to the total number of quantum circuit executions and not to the total number of episodes needed to solve an environment, as in standard RL notation.

A. SAMPLE COMPLEXITY OF ESTIMATING CLASSICAL FIM

Recall that the classical FIM is represented as the outer product of the gradient of the log policy averaged through the sampled trajectories,

$$F = \mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi(\cdot|s,\theta)}\left[\nabla_\theta \log\pi(a|s,\theta)\nabla_\theta \log\pi(a|s,\theta)^T\right] \tag{40}$$

where d^{π_θ} is the state distribution under the policy π_θ. Since the gradient of log policy is needed, the sample complexity is actually dependent on the type of policy employed. Let us start discussion with the Born policy.

FIM - BORN POLICY

The Born policy is represented as a probability distribution over a partition V_a of computational basis states, as presented in Section III. Notice, however, that complexity depends on this partition. In its most general form we could $\pi(a|s,\theta) = \langle P_a\rangle_{s,\theta} = \sum_{v\in V_a}\langle P_v\rangle_{s,\theta}$ with $V_a = \frac{2^n}{|A|}$. This is similar to the representation of the Born policy employed in the Acrobot environment in Section IV, with the exception that the number of actions is not even and the partition does not perfectly correspond to $\frac{2^n}{|A|}$. Nevertheless, note that in the Cartpole environment the policy is even more simple than before since a single-qubit is considered. The log policy gradient can be expanded in this case using parameter-shift rules, as follows,

$$\begin{aligned}\partial_\theta \log\pi(a|s,\theta) &= \sum_{v\in V_a}\frac{\partial_\theta\langle P_v\rangle_{s,\theta}}{\langle P_v\rangle_{s,\theta}} \\ &= \sum_{v\in V_a}\frac{\langle P_v\rangle_{s,\theta+\frac{\pi}{2}} - \langle P_v\rangle_{s,\theta-\frac{\pi}{2}}}{\langle P_v\rangle_{s,\theta}}\end{aligned} \tag{41}$$

It is known that for an ϵ-approximation to the probability $\mathcal{O}(\epsilon^{-2})$ circuit executions are needed Schuld and Petruccione [2021]. Ignoring the approximation error, the total number of independent quantum circuits needed to estimate the gradient is three. Since the probability itself was already estimated for the estimation of the policy, the partial derivative can indeed can be reduce to two quantum circuit calls. Each projector is a linear expectation value depending on $\frac{2^n}{|A|}$ partitions. However, since the policy is estimated from finite shots, the number of actions does not influence the number of quantum circuit calls and can be neglected. Therefore, since the FIM is a $k\times k$ matrix for $\theta\in\mathbb{R}^k$, we need $\mathcal{O}(2k^2)$ quantum circuit executions. Recall that in practice we can use the symmetry of the matrix to further reduce the the number of calls.

FIM - SOFTMAX POLICY

Assume, for simplicity, that the same projectors as in the Born policy are considered as action's numerical preferences, but are otherwise irrelevant. Recall the log policy gradient expansion:

$$\nabla_\theta \log\pi(a|s,\theta) = \nabla_\theta\langle P_a\rangle_{s,\theta} - \sum_{a'\in A}\pi(a'|s,\theta)\nabla_\theta\langle P_{a'}\rangle_{s,\theta} \tag{42}$$

Thus, the Softmax policy depends on the total number of actions $|A|$ to estimate the derivative w.r.t a single parameter. Thus, using parameter-shift rules for estimating the partial derivatives of projectors as above, for $\theta\in\mathbb{R}^k$, we need $\mathcal{O}(2|A|k^2)$ quantum circuit executions, in the worst case.

B. SAMPLE COMPLEXITY OF ESTIMATING QUANTUM FIM

Recall that the quantum FIM obtained from the infinitesimal distances between quantum states as represented in Equation (16) depends on the quantum state only. Thus, it can be immediately concluded that the sample complexity of estimating quantum FIM will not be dependent on the policy and thus on the total number of possible actions associated with

an environment. Importantly, an entry of the quantum FIM, $\mathcal{F}_{ij}$ can be obtained from the estimation of four independent overlaps, as proposed in Meyer [2021] shifting the respective parameters i, j:

$$\begin{aligned}\mathcal{F}_{ij} = -\frac{1}{2}\Big(&\left|\langle\psi(\boldsymbol{\theta}) \mid \psi\left(\boldsymbol{\theta} + (\boldsymbol{e}_i + \boldsymbol{e}_j)\,\tfrac{\pi}{2}\right)\rangle\right|^2 \\ &- \left|\langle\psi(\boldsymbol{\theta}) \mid \psi\left(\boldsymbol{\theta} + (\boldsymbol{e}_i - \boldsymbol{e}_j)\,\tfrac{\pi}{2}\right)\rangle\right|^2 \\ &- \left|\langle\psi(\boldsymbol{\theta}) \mid \psi\left(\boldsymbol{\theta} - (\boldsymbol{e}_i - \boldsymbol{e}_j)\,\tfrac{\pi}{2}\right)\rangle\right|^2 \\ &+ \left|\langle\psi(\boldsymbol{\theta}) \mid \psi\left(\boldsymbol{\theta} - (\boldsymbol{e}_i + \boldsymbol{e}_j)\,\tfrac{\pi}{2}\right)\rangle\right|^2\Big)\end{aligned} \tag{43}$$

where e_j is the unit vector along the θ_j axis. Thus, for $\theta \in \mathbb{R}^k$, we need $\mathcal{O}(4k^2)$ quantum circuit executions to estimate the quantum FIM. It seems that the estimation of the quantum FIM may be significantly cheaper compared to that of the classical FIM in the context of using a Softmax policy since every possible action must be taken into account to estimate the classical FIM. However, since the quantum FIM produces updates directly in state-space instead of policy-space, such a difference in sample complexity can be neglected in terms of the actual ability in solving the environment as discussed in Section IV.

Environment	Policy	FIM	QFIM
Cartpole	Born	65	504
	Softmax	65	504
Acrobot	Born	65	504
	Softmax	65	504

TABLE 2. Comparison of FIM and QFIM values for different environments and policies

Table 2 summarizes the number of quantum circuit calls (ignoring the number of shots, for simplicity) required to estimate both the FIM and QFIM for the environments and policies considered in the numerical experiments. The number of quantum circuit calls was estimated using Pennylane's *tracker* functionality. To ensure a fair comparison, only a single time step interaction with the environment was considered instead of full episodes. This was done because episodes can have different lengths, which would imply a varying number of quantum circuit calls, complicating the analysis.

Table 2 clearly demonstrates that the QFIM is more expensive to estimate in the environments considered previously. However, notice that both the Born and Softmax policies have the same number of quantum circuit calls to estimate the FIM, which does not show dependence on the action space. This is true in this case because for the Softmax policy, we are considering the same projectors as in the Born formulation but with a Softmax post-processing activation, introducing the greediness control parameter. Thus, we can conclude that the estimation of the empirical FIM does not depend on the action space if we either consider a Born policy or a Born policy with a Softmax post-processing activation. However, in general, the Softmax policy will be composed of $|A|$ expectation values of arbitrary Hermitian observables encoding the numerical preference of each action. In this setting, the number of quantum circuit calls is expected to depend on the number of actions. Let us consider the most general case in which we have $|A|$ *non-commuting* observables. Table 3 shows the number of quantum circuit calls required to estimate the FIM and QFIM as a function of the number of actions. The number of qubits in this setting is fixed at $n = 4$ to keep the same circuit with the same number of parameters as previously.

$\lvert A\rvert$	Policy	FIM	QFIM
2	Born	65	504
	Softmax	130	504
3	Born	65	504
	Softmax	195	504
4	Born	65	504
	Softmax	260	504
5	Born	65	504
	Softmax	325	504
6	Born	65	504
	Softmax	390	504
7	Born	65	504
	Softmax	455	504
8	Born	65	504
	Softmax	520	504

TABLE 3. Comparison of FIM and QFIM resources for different action spaces. The number of qubits in this setting is fixed at $n = 4$ to keep a fair comparison with previous experiments.

Table 3 clearly demonstrates that the number of quantum circuit calls to estimate the FIM increases with the number of actions for the Softmax policy, as expected. Moreover, note that for $n = 4$, the circuit is composed of $k = 32$ parameters, and for $|A| = 8$, the sample complexity of estimating the FIM is already superior relative to the QFIM. This is a clear indication that the QFIM can indeed be more efficient to estimate depending on the number of parameters and actions. Therefore, there can be real-world scenarios where the user has access only to a limited number of resources, enabling the QFIM to be more efficient to estimate than the FIM. Nonetheless, the actual utility of the QFIM in solving the environment still needs to be further investigated.

VI. CONCLUSION

In this paper, we reported a series of experiments aiming at comparing the effectiveness of natural policy gradients preconditioned by the quantum Fisher Information Matrix (FIM) with those preconditioned by the traditional classical FIM. Our findings indicate that considering a quantum FIM preconditioning leads to a larger approximation error. However, when utilizing the square roots of the information matrices, the square root of the quantum FIM could compensate the approximation error with the gradient vector norm which leads to a reduction in regret relative to its classical counterpart. Note however, that this advantage may not always translate into near-optimal policy. This hypothesis was tested in standard control benchmark settings, confirming that the preconditioning of the quantum FIM with its square root inverse leads to better sample efficiency over the square

root of the classical FIM preconditioning. Conversely, using the full inverse for quantum FIM preconditioning did not significantly outperform the classical approach. It is important to note that our sample complexity analysis revealed that unlike the classical FIM, the quantum FIM's estimation is not affected by the size of the action space in a given environment, which presents a notable distinction between the two. Further investigation is necessary, particularly in environments with large action spaces since these are not easily solved with current quantum technologies, to fully determine the practical efficacy of quantum natural policy gradients. This will be a focus of future research, along with the investigation of approximations of quantum FIM Beckey et al. [2022], Stokes et al. [2020]. The role of the quantum and classical FIM in the trainability of PQC-based policies is also a promising avenue for future research.

ACKNOWLEDGEMENTS

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within grants LA/P/0063/2020, UI/BD/152698/2022 and project IBEX, with reference PTDC/CC1-COM/4280/2021

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. The Journal of Machine Learning Research, 22(1):98:4431–98:4506, January 2021. ISSN 1532-4435.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A Brief Survey of Deep Reinforcement Learning. IEEE Signal Processing Magazine, 34(6):26–38, November 2017. ISSN 1053-5888. .

Jacob L. Beckey, M. Cerezo, Akira Sone, and Patrick J. Coles. Variational Quantum Algorithm for Estimating the Quantum Fisher Information. Physical Review Research, 4(1):013083, February 2022. ISSN 2643-1564. .

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M. Sohaib Alam, Guillermo Alonso-Linaje, B. AkashNarayanan, Ali Asadi, Juan Miguel Arrazola, Utkarsh Azad, Sam Banning, Carsten Blank, Thomas R. Bromley, Benjamin A. Cordier, Jack Ceroni, Alain Delgado, Olivia Di Matteo, Amintor Dusko, Tanya Garg, Diego Guala, Anthony Hayes, Ryan Hill, Aroosa Ijaz, Theodor Isacsson, David Ittah, Soran Jahangiri, Prateek Jain, Edward Jiang, Ankit Khandelwal, Korbinian Kottmann, Robert A. Lang, Christina Lee, Thomas Loke, Angus Lowe, Keri McKiernan, Johannes Jakob Meyer, J. A. Montañez-Barrera, Romain Moyard, Zeyue Niu, Lee James O'Riordan, Steven Oud, Ashish Panigrahi, Chae-Yeun Park, Daniel Polatajko, Nicolás Quesada, Chase Roberts, Nahum Sá, Isidor Schoch, Borun Shi, Shuli Shu, Sukin Sim, Arshpreet Singh, Ingrid Strandberg, Jay Soni, Antal Száva, Slimane Thabet, Rodrigo A. Vargas-Hernández, Trevor Vincent, Nicola Vitucci, Maurice Weber, David Wierichs, Roeland Wiersema, Moritz Willmann, Vincent Wong, Shaoming Zhang, and Nathan Killoran. PennyLane: Automatic differentiation of hybrid quantum-classical computations, July 2022.

Rajendra Bhatia. Matrix Analysis, volume 169 of Graduate Texts in Mathematics. Springer, New York, NY, 1997. ISBN 978-1-4612-6857-4 978-1-4612-0653-8. .

Samuel Yen-Chi Chen, Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, and Hsi-Sheng Goan. Variational quantum circuits for deep reinforcement learning. IEEE Access, 8:141007–141024, 2020.

El Amine Cherrat, Snehal Raj, Iordanis Kerenidis, Abhishek Shekhar, Ben Wood, Jon Dee, Shouvanik Chakrabarti, Richard Chen, Dylan Herman, Shaohan Hu, Pierre Minssen, Ruslan Shaydulin, Yue Sun, Romina Yalovetzky, and Marco Pistoia. Quantum Deep Hedging, March 2023.

Tobias Haug and M. S. Kim. Optimal training of variational quantum algorithms without barren plateaus, June 2021.

Sofiene Jerbi, Casper Gyurik, Simon Marshall, Hans Briegel, and Vedran Dunjko. Parametrized Quantum Policies for Reinforcement Learning. In Advances in Neural Information Processing Systems, volume 34, pages 28362–28375. Curran Associates, Inc., 2021.

Sofiene Jerbi, Arjan Cornelissen, Māris Ozols, and Vedran Dunjko. Quantum policy gradient algorithms, December 2022.

Sham M Kakade. A Natural Policy Gradient. In Advances in Neural Information Processing Systems, volume 14. MIT Press, 2001.

Johannes Jakob Meyer. Fisher Information in Noisy Intermediate-Scale Quantum Applications. Quantum, 5: 539, September 2021. .

Nico Meyer, Daniel D. Scherer, Axel Plinge, Christopher Mutschler, and Michael J. Hartmann. Quantum Natural Policy Gradients: Towards Sample-Efficient Reinforcement Learning, August 2023a.

Nico Meyer, Daniel D. Scherer, Axel Plinge, Christopher Mutschler, and Michael J. Hartmann. Quantum Policy Gradient Algorithm with Optimized Action Decoding, May 2023b.

Murphy Yuezhen Niu, Sergio Boixo, Vadim N. Smelyanskiy, and Hartmut Neven. Universal quantum control through deep reinforcement learning. npj Quantum Information, 5 (1):1–8, April 2019. ISSN 2056-6387. .

Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach (4th Edition). Pearson, 2020. ISBN 978-0-13-461099-3.

Maria Schuld and F. Petruccione. Machine Learning with Quantum Computers. Springer, Cham, Switzerland, second edition edition, 2021. ISBN 978-3-030-83098-4.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. Physical Review A, 99(3):032331, March 2019. ISSN 2469-9926, 2469-9934. .

John Schulman, Sergey Levine, Philipp Moritz, Michael I.

Jordan, and Pieter Abbeel. Trust Region Policy Optimization, April 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017b.

André Sequeira, Luis Paulo Santos, and Luis Soares Barbosa. Policy gradients using variational quantum circuits. Quantum Machine Intelligence, 5(1):18, April 2023. ISSN 2524-4914. .

Andrea Skolik, Sofiene Jerbi, and Vedran Dunjko. Quantum agents in the gym: A variational quantum algorithm for deep q-learning. Quantum, 6:720, 2022.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum Natural Gradient. Quantum, 4:269, May 2020. ISSN 2521-327X. .

Ajay Subramanian, Sharad Chitlangia, and Veeky Baths. Reinforcement learning and its connections with neuroscience and psychology. Neural Networks, 145:271–287, January 2022. ISSN 0893-6080. .

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning - an Introduction. Adaptive Computation and Machine Learning. MIT Press, 1998. ISBN 978-0-262-19398-6.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In Advances in Neural Information Processing Systems, volume 12. MIT Press, 1999.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3):229–256, May 1992. ISSN 1573-0565. .

Xingzhi Zhan. 1. Inequalities in the Löwner Partial Order. In Xingzhi Zhan, editor, Matrix Inequalities, Lecture Notes in Mathematics, pages 1–15. Springer, Berlin, Heidelberg, 2002. ISBN 978-3-540-45421-2. .

APPENDIX. TABLES FOR ENVIRONMENTS DESCRIPTION AND PQC'S

Environment	State	Action	Reward function	Horizon	Termination criteria
Cartpole	4 features	2 actions $A = \{0, 1\}$	+1 per time step	200 time steps	Reach horizon or out of bounds
Acrobot	4 features	3 actions $A = \{0, 1, 2\}$	-1 + height	500 time steps	Reach goal or horizon

TABLE 4. Characterization of the environments considered in the numerical experiments.

Environment	Policy	Layers	Observables	Batch Size
CartPole	Born	4	$\{P_0, P_1\}$	10
	Softmax	4	$\{P_0, P_1\}$	10
Acrobot	Born	5	$P_a = \sum_{b\in\{0,1\}^n}^{int(b) \mod 3=a} \|b\rangle\langle b\|$	10
	Softmax	5	$P_a = \sum_{b\in\{0,1\}^n}^{int(b) \mod 3=a} \|b\rangle\langle b\|$	10

TABLE 5. Characterization of the PQC's considered in the numerical experiments. P_i indicates the projector in the computational basis in decimal. For the Cartpole environment a single-qubit was measured and the probability of each basis state associated to an action. In the Acrobot environment, the action assignmment was made using $int(b) \mod 3 = a$ for a particular basis state b.