Reinforcement Learning based Actor Critic and Policy Agent for Optimized Quantum Sensor Circuit Design

Temitope Adeniyi
Dept. of Electrical Engineering and Computer Science
Cleveland state University
Cleveland, OH, USA
t.adeniyi@vikes.csuohio.edu

Sathish Kumar
Dept. of Electrical Engineering and Computer Science
Cleveland state University
Cleveland, OH, USA
s.kumar13@csuohio.edu

2024 IEEE International Conference on Quantum Computing and Engineering (QCE) | 979-8-3315-4137-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/QCE60285.2024.00146

***Abstract*—This paper explores the application of reinforcement learning (RL) algorithms to the domain of quantum sensor circuit design. We focus on the design of four qubits quantum circuits that can efficiently generate states with optimal sensitivity, which maximizes the Quantum Fisher Information (QFI). We formulate the design task as a Markov Decision Process, implementing and comparing two distinct RL strategies: pure policy gradient and actor-critic agents to optimize a Ramsey sequence for maximum QFI. The results demonstrate that both agents are able to find quantum circuit designs with an optimal QFI value of 1. However, the agents differ in the gate sequence of the final Quantum sensor circuit configurations that indicate the optimal state. The policy gradient outperforms the actor-critic method in convergence time and the number of gate sequences for the encoding circuit.**

***Index Terms*—Quantum Sensing, Quantum Metrology, Optimization**

I. Introduction

Quantum sensing exploits quantum mechanical phenomena to achieve measurements that surpass classical limits [1], [2]. Unlike classical sensors, which are fundamentally constrained by the Heisenberg uncertainty principle, quantum sensors can achieve higher precision through the manipulation and control of quantum states. This capability is underpinned by Quantum Fisher Information (QFI), a metric that quantifies the ultimate precision achievable in quantum parameter estimation. QFI plays a pivotal role in quantum sensing by guiding the design of optimal measurement strategies that minimize uncertainties and maximize sensitivity. Quantum sensors have a wide array of applications, ranging from gravitational wave detection [3]–[5] to imaging [6] including atomic clocks [7], [8]. Quantum sensors engage in interactions with the environment such that environmental variables or phenomena become encoded within their quantum states. This encoding occurs through quantum entanglement and superposition where the quantum sensors effectively become entwined with the environmental factors they are measuring [9]. Through this entanglement, quantum sensors can discern subtle changes in their surroundings with high sensitivity, allowing for the precise measurement of physical quantities.

The heart of a quantum sensor is its circuit, which coordinates the behavior of qubits to interact with an external parameter. The influence exerted on a quantum sensor by a physical phenomenon can be expressed through the application of a suitable unitary transformation, leading to a modification in the sensor's quantum phase. A well-designed circuit maximizes the quantum sensor's sensitivity by preserving quantum states and enhancing their response to the physical quantity of interest [10]. Circuit design is thus critical, as it directly impacts the efficacy and efficiency of a quantum sensor. However, the complex and delicate nature of quantum states presents significant challenges in designing circuits that can maintain coherence and accurately translate environmental interactions into measurable quantum states.

In this research paper, we focus on designing an optimal Ramsey interferometry to estimate a phase shift. Here, the term "optimality" is defined as achieving the highest Quantum Fisher Information (QFI) for a given parameter of interest within a quantum system. The QFI is a metric in quantum metrology quantifying the precision with which a parameter, can be estimated [11].

To be able to design a optimal Ramsey interferometer quantum circuit, we use reinforcement learning (RL). RL algorithms has the ability to learn to make decisions by interacting with an environment, receiving feedback in the form of rewards, and using this information to improve future decisions [12]. Our work extends beyond conventional variational optimization approaches, as RL affords the flexibility to adapt to complex quantum environments without explicit knowledge of their underlying dynamics. Through a combination of exploration, exploitation, and reward-driven learning, we use RL agents to autonomously navigate the landscape of quantum circuit design, honing their strategies and discover gate sequences and configurations that yield optimal sensor performance.

Despite the potential of RL in this domain, there is a paucity of research systematically comparing different RL strategies

This work was supported by National Science Foundation Grant No. OMA 2231377

979-8-3315-4137-8/24/$31.00 ©2024 IEEE
DOI 10.1109/QCE60285.2024.00146

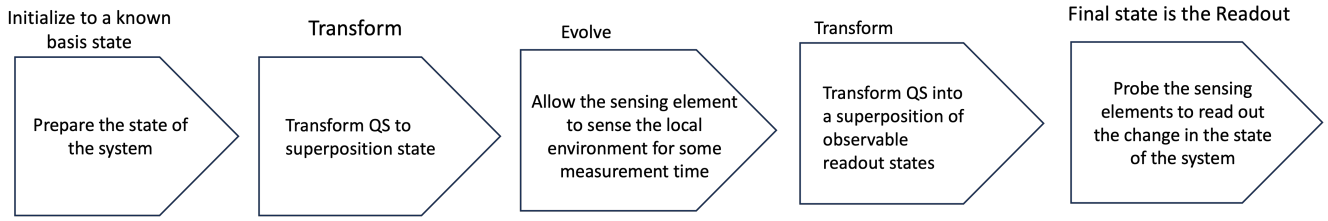

Fig. 1. A quantum sensing process

for quantum circuit design. Moreover, the peculiarities of quantum systems pose unique challenges to traditional RL methods. Thus, there is a need for a comprehensive study that not only compares the effectiveness of various RL approaches in this context but also adapts these strategies to accommodate the intricacies of quantum sensing. This paper presents a comparative study of two RL approaches—policy gradients [13], and actor-critic methods [14], [15] in the design of quantum sensing circuits. We adapt these algorithms to account for the non-classical nature of quantum information processing and evaluate their performance in a simulated quantum sensing environment.

The selection of these specific reinforcement learning (RL) algorithms was guided by several key considerations. Policy gradient methods directly optimize the policy by computing gradients with respect to the expected return. This feature is crucial for maximizing QFI, as it allows the RL agent to continuously adjust and improve the quantum gate sequences to enhance the sensitivity of the quantum sensor. By directly optimizing the policy, these methods can efficiently navigate complex quantum state spaces and discover gate sequences that lead to higher QFI. Also, both policy gradient and actor-critic methods are well-suited for handling stochastic policies, which is important in quantum systems where uncertainty and probabilistic behaviors are inherent. The actor-critic approach, in particular, offers a balanced trade-off between policy optimization and value function estimation. The actor proposes actions, while the critic evaluates them. This balanced approach helps ensure that the gate chosen at each step is the most optimal action to maximize the QFI.

Our contributions include the design of a specialized simulation environment and Markov Decision Process (MDP) for quantum sensor circuit design. Our paper also presents adaptations of the pure policy gradient, and actor-critic RL agents to operate within the quantum domain. Finally, our paper provides guidelines for selecting and implementing RL methods in quantum technologies.

The rest of the manuscript is as follows; Section II explains the related work. Section III gives a background for the quantum sensing design problem, and describes the measure of optimality used in the study and how a quantum sensor can be modelled as a quantum circuit. Section IV describes the methodology for the RL algorithm and both RL agents used in the study. Section V explains the environment and experimental setup. Section VI gives the result of the study, and Section VII discusses the result. Section VIII notes the future work recommendations and finally section IX concludes the paper

II. Related Works

The field of quantum sensing, marked by its complex integration of quantum mechanics and computational strategies, has seen notable advances through the application of reinforcement learning (RL). This section reviews critical contributions from various researchers who have explored different aspects of quantum metrology and sensor optimization using RL and other computational techniques.

Xiao et al., in [18], investigate the use of deep reinforcement learning (DRL) to improve parameter estimation in quantum sensing across various scenarios, including both time-dependent and independent conditions, under both noise-free and noisy environments. Their research introduces a linear time-correlated control ansatz, inspired by physical principles, which integrates a precisely defined reward function to speed up the training of DRL networks. This approach notably enhances the generation of quantum control signals. The robustness and efficiency of the DRL method are confirmed through extensive simulations, showing marked improvements, especially in time-dependent parameter estimation and adaptability to parameter deviations. The study also integrates LSTMCell to better understand sequential quantum states, improving stability and proposes the use of classical shadow and generative neural quantum states to overcome challenges like the necessity for full-state tomography in quantum sensors.

Kaubruegger et al. [19] explore strategies for quantum metrology within the Bayesian framework for SU(2) quantum interferometry, aiming to optimize quantum sensor configurations for precise multi-parameter estimation using N-atom sensors. Their research demonstrates that sensors with limited entanglement capabilities can significantly outperform traditional non-entangled sensors and approach optimal performance levels. The study also discusses practical implementations using current programmable quantum technologies, employing specifically tailored entangled states and measurements. Christain and Thomas et al., [17] discuss an experimental programmable quantum sensor using trapped ions, employing low-depth, parametrized quantum circuits tailored for specific sensing tasks. The sensor significantly outperforms traditional spin-squeezing methods, achieving a sensing precision near the quantum limits, with precision up to 1.45 ± 0.01 of the funda-

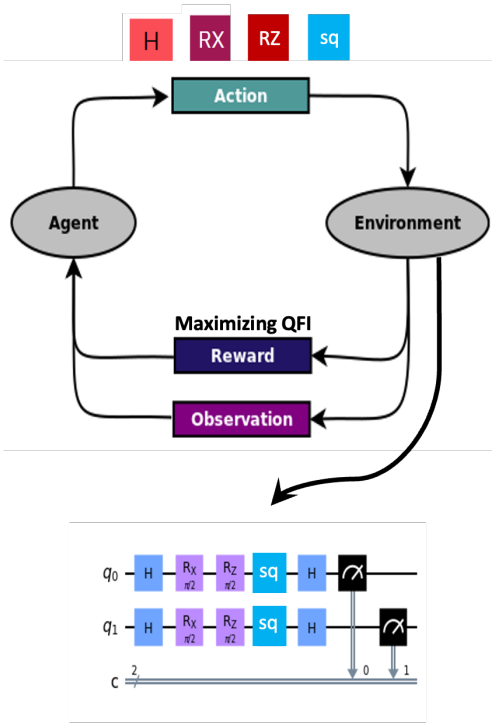

Fig. 2. RL architecture for the quantum sensor design methodology employed in this research.

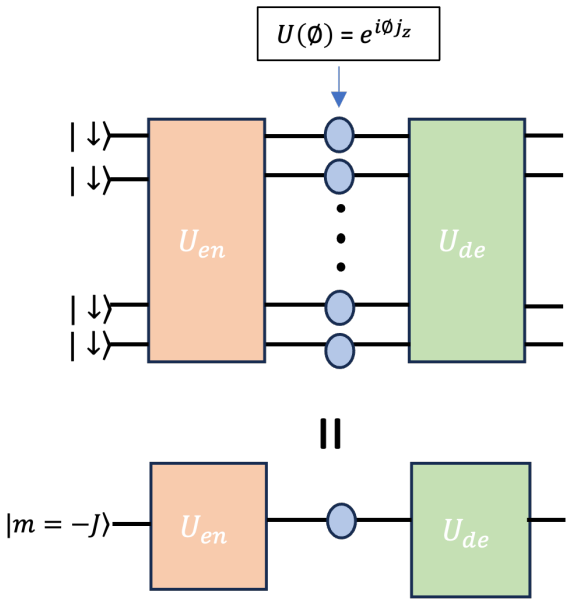

Fig. 3. The figure shows a four qubit quantum sensor circuit model. The encoder U_{en} and decoder U_{de} transform the quantum state of the system before and after the phase encoding. The image also shows the One-partite quantum sensor circuit model in which the atoms are controlled uniformly. One-partite control is realized by circuits composed of collective application of the quantum gates contained in U_{en} and U_{de}

mental limit and enhancing stability by reducing the required averages for a given Allan deviation by a factor of 1.59 ± 0.06. The research further introduces on-device quantum-classical feedback for self-calibration, allowing the sensor to optimize adaptively without prior device or environmental knowledge.

The paper [20] presents an approach to enhance measurement precision in quantum sensors through machine optimization. It specifically addresses the trade-off between state preparation and sensing time in quantum metrology by proposing a method where entanglement generation and sensing occur simultaneously. This optimization leverages designed sequences of rotations and twist-and-turn dynamics to navigate the constraints of limited coherence time, aiming to maximize the Quantum Fisher Information (QFI).

III. Background

A. Quantum Sensing Design Problem

In a general quantum sensing protocol, we aim to detect an external parameter, represented by a complex-valued time function $\lambda(\tau)$. The typical approach involves initially setting the quantum system to a well-defined quantum state, then subjecting it to the parameter, which alters this state. The system is then allowed to evolve over a period τ, after which we measure the altered state to gather data about the parameter (this measurement collapses the wavefunction into an eigenvector of the selected observable). This procedure is repeated multiple times to accumulate more detailed information about the parameter.

In this study, we consider a quantum system characterized by two distinct states, $|0\rangle$ and $|1\rangle$, which we postulate as eigenvectors of the sensor's inherent Hamiltonian. These states correspond to energies E_0 and E_1, respectively, with an energy difference $\Delta E = E_1 - E_0$ between them [21]. Additionally, we assume that the magnitudes of E_0, E_1, and ΔE are significantly greater than the perturbations in the Hamiltonian caused by $\lambda(\tau)$. This assumption is generally reasonable for the majority of quantum sensing scenarios.

The Hamiltonian H of a quantum sensor, instrumental in defining the system's evolution, can be generalized for a system affected by an external parameter $\lambda(\tau)$ [1], [21] as described in (1):

$$\widehat{H}(\tau) = \widehat{H}_{\lambda}(\tau) + \widehat{H}_0 + \widehat{H}_{control}(\tau), \tag{1}$$

where $H_{\lambda}(\tau)$ is the component of the Hamiltonian directly interacting with $\lambda(\tau)$ and varies with time τ, $\widehat{H}_0 = \begin{pmatrix} E_0 & 0 \\ 0 & E_1 \end{pmatrix}$ is the internal hamiltonian of the system in the absence of $\lambda(\tau)$ and $\widehat{H}_{control}(\tau)$ is used for tuning the separation of the energy levels and implementing unitary operations with quantum gates as required to carry out a sensing protocol. The sensitivity of the quantum sensor to variations in $\lambda(\tau)$ is predicated on the specific form of $H_{\lambda}(\tau)$ and its commutation relations with other parts of the Hamiltonian.

Our quantum sensing circuit design problem entails finding a sequence of quantum gates that prepares a system in an initial state that evolves under a Hamiltonian influenced by $\lambda(\tau)$. The task is to optimize the circuit so that the system's

final state encodes information about $\lambda(\tau)$ with maximum sensitivity. This optimization problem is emblematic of quantum metrology, where the precision of parameter estimation is fundamentally limited by the laws of quantum mechanics. A typical scenario is the Ramsey sequence, a type of qunatum interferometry, where a system with $|0\rangle$ and $|1\rangle$ differing in energy by $\Delta E = \hbar\omega$ is initially prepared in state $|0\rangle$ and put in a superposition state $|\psi_0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$with a $\frac{\pi}{2}$ operator. The system is allowed to evolve freely under the influence of an unknown field for time τ, yeilding a state $|\psi(\tau)\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle + e^{i\omega\tau}|1\rangle\right)$. After another $\frac{\pi}{2}$ operation, the final state, $|\psi_f(\tau)\rangle = \frac{1}{2}(1 + e^{i\omega\tau})|0\rangle + \frac{1}{2}(1 + e^{i\omega\tau})|1\rangle$ can then be measured.

The protocol for quantum sensing can be summarised as three stages: preparation of the initial quantum state, interaction of the state with the external parameter $\lambda(\tau)$ to encode its information, and measurement of the state to infer $\lambda(\tau)$. The Ramsey sequence mentioned above serves as a foundational protocol for this process, especially in atomic clocks and interferometry.

B. QFI as a measure of Optimality

In quantum metrology, the Quantum Fisher Information (QFI) serves as an important metric for assessing the optimality of a measurement scheme [23]–[26]. QFI is the quantum counterpart of classical Fisher information which quantifies the amount of information that an observable provides about an unknown parameter [27]–[29].

QFI measures the sensitivity of a quantum state to changes in a parameter λ, making it a fundamental tool for assessing the effectiveness of quantum encoding states in parameter estimation tasks. High QFI indicates greater sensitivity to parameter changes, leading to more precise and accurate measurements. The general definition of QFI applicable to any quantum state represented by the density matrix $\rho(\theta)$ [29] is given by (2)

$$I(\theta;\rho_\theta) = \mathrm{Tr}[J_\theta^2 \rho_\theta] \tag{2}$$

where J_θ is the symmetric logarithmic derivative of $\rho(\theta)$ with respect to θ

In designing an optimal quantum sensor, the selection of the most effective initial quantum encoding state is crucial and this choice will be guided by identifying the state that exhibits the highest QFI. In our research, we focus on optimizing the initial quantum encoding state to maximize QFI, thus enhancing the sensor's performance. This involves selecting quantum states that can most effectively utilize quantum mechanics to improve measurement precision. For instance, in Ramsey interferometry, a widely used technique for phase estimation, maximizing QFI equates to achieving the highest precision of phase measurements [30], [31].

Equation (2) mathematically defines the QFI for a quantum state ρ and a parameter $\lambda(\tau)$.

$$F_Q(\lambda(\tau)) = 4(\langle\psi'(\lambda(\tau))|\psi'(\lambda(\tau))\rangle - |\langle\psi'(\lambda(\tau))|\psi(\lambda(\tau))\rangle|^2), \tag{3}$$

with $\psi'(\lambda(\tau))$ being the derivative of the quantum state with respect to $\lambda(\tau)$. The Quantum Cramér-Rao Bound (QCRB) sets the ultimate limit on the precision of estimating the parameter $\lambda(\tau)$ [22] and is given by (4)

$$(\Delta\lambda(\tau))^2 \geq \frac{1}{MF_Q(\lambda(\tau))}, \tag{4}$$

where denotes $(\Delta\lambda(\tau))^2)$ the variance in the estimation of $\lambda(\tau)$, M represents the number of independent measurements, and F_Q is the Quantum Fisher Information (QFI).

In general, we expect that the circuit that maximizies QFI along a direction, say z, would be a modified GHZ state in the z direction:

$$|\psi\rangle = \frac{|\uparrow\rangle^{\otimes N} + e^{i\phi}|\downarrow\rangle^{\otimes N}}{\sqrt{2}} \tag{5}$$

C. Modelling a Quantum Sensor as a Quantum Circuit

To demonstrate the effectiveness of our design approach, we implement a basic quantum sensor circuit with thee desired number of qubits. This circuit has a well-known optimal solution, which we aim to rediscover. Our goal is to generate what is known as the N00N state, as it maximizes the Quantum Fisher Information (QFI), a crucial measure in quantum metrology.

First, we start with all qubits in the $|0\rangle$. Then, we execute a generalized Ramsey sequence. This sequence helps us measure the accumulation of a phase difference between the collective behavior of the qubits and a stable local oscillator, crucial for timing synchronization. The phase accumulation process is related to the Ramsey sequence's generator, denoted by $\mathcal{G}$ which is basically the Pauli-Z operator.

To encode and later decode the phase, we use two unitary operations first an encoding operation $U_{en}(\theta)$ preparing a state $|\psi\rangle$ from the initial product state $|\downarrow\rangle^{\otimes N}$ of N particles, and secondly a decoding operation $U_{de}(\vartheta)$. The decoding operation often involves applying a sequence of gates that effectively reverses the initial encoding process, thus making the accumulated phase relative to each qubit's initial state measurable. These operations involve sequences of basic gate operations, like single-qubit rotations and multi-qubit entanglement gates. Specifically, we rotate the qubits around the x and y axes by angles θ_x and θ_z using the Rx, Rz gates respectively. Additionally, we introduce a squeezing operation $S(r,\phi)$ which alters the uncertainty relationship between certain properties of the quantum system, enhancing sensitivity in one aspect while increasing uncertainty in another.

The sequence can be represented by the unitary operator U, given by (6)

$$U(\theta,\phi) = R_z(\phi)U_{en}(\theta)R_x\left(\frac{\pi}{2}\right)e^{i\mathcal{G}\tau}R_x\left(\frac{\pi}{2}\right)U_{de}(\vartheta)R_z(\phi) \tag{6}$$

where $R_x\left(\frac{\pi}{2}\right)$ and $R_z(\theta)$ are rotation gates around the x- and z-axis, respectively, $e^{i\mathcal{G}\tau}$ denotes the evolution of the system under the Hamiltonian H for an interaction time τ. $U_{en}(\theta)$ and $U_{de}(\vartheta)$ are the entangling and decoding circuits, respectively, with control parameters θ and ϑ

It's worth noting that our design considers a circuit as a one-partite quantum sensor [19], Each gate or unitary operation acts respectively on a single quantum system consisting of four qubit. This setup lets us measure a collective parameter, such as phase, across all qubits simultaneously, making our sensor more efficient and sensitive.

IV. Methodology

A. RL Algorithm Overview

In designing optimal quantum sensors, we employ Reinforcement Learning (RL) algorithms to iteratively improve the performance of our sensor designs. RL provides a framework for training agents to make sequential decisions in uncertain environments through trial and error, guided by a reward signal. RL enables the optimization of control parameters to maximize the sensitivity and precision of the sensor. Here, we maximize the QFI, as it directly influences the sensor's ability to accurately estimate the parameter λ of interest.

The RL algorithm operates within a Markov Decision Process (MDP), defined by a tuple (S, A, P, R) [32] where S is the set of states representing the quantum system's configurations denoted as ψ. A is the set of actions corresponding to unitary or quantum gate operations denoted as U. P is the state transition probability function, describing the evolution of the quantum state based on the chosen action. R is the reward function based on the Quantum Fisher information denoted as F_Q, it provides feedback to the agent based on its actions and the resulting state transitions [33]. It reflects the sensitivity of the quantum state to parameter changes at each step of the circuit's evolution. At each time step t, the agent observes the current state $\psi_t \in \psi$, selects an action $U_t \in U$ based on its policy, and receives a reward $F_{Qt} \in F_Q$ from the environment. After receiving the reward, the environment transitions to a new state ψ_{t+1}. If ψ_{t+1} is a terminal state, the agent receives a final reward $f(\psi_{t+1})$. We define the terminal state as the state having the highest F_Q. Equation (7) gives the trajectory of the MDP as a sequence of random variables [34]

$$(F_{Q0}, \psi_0, U_0), (F_{Q1}, \psi_1, U_1), ..., (F_{Qt+1}, \psi_{t+1}, U_{t+1}) \quad (7)$$

The policy function π defines the probability of taking action U in state ψ. We express it as:

$$\pi(U|\psi) = P(U_t = U|\psi_t = \psi) \quad (8)$$

. where $U_t = U$ is applying a unitary operation at time t, and $\psi_t = \psi$ is the quantum state at time t. For a policy ψ, the goal of the agent is to learn an optimal policy π^* that maximizes the expected cumulative reward over time. The optimal policy π^* can be described by (9):

$$\pi^* = \arg\max_{\pi} E[F_{Qt}|\psi_t = \psi, \pi] \quad (9)$$

where F_{Qt} is the reward at tome t , and E denotes the expectation. The policy is refined iteratively, where updates may depend on the gradients of expected rewards or the maximization of the action-value function.

The RL algorithm leverages the concept of value functions to evaluate the goodness of states and actions. The state-value function V^π in (10) gives the expected cumulative reward starting from state ψ and following policy π [34].

$$V^\pi(\psi) = E\left[\sum_{k=0}^{\infty} \gamma^k F_{Qt+k+1}|\psi_t = \psi\right] \quad (10)$$

where γ is the is the discount factor, representing the difference in importance between future rewards and immediate rewards, and F_{Qt+k+1} is the reward received after k steps from t.

The action-value function Q^π given in (11) estimates the expected cumulative reward starting from state ψ, taking an action U and thereafter following policy π [34] .

$$Q^\pi(\psi, U) = E[F_{Qt} + \gamma V^\pi(\psi_{t+1}|\psi_t = \psi, U_t = U)] \quad (11)$$

where ψ_{t+1} is the new state after action U is taken. Both V^π and Q^π are central to the RL algorithm's operation as they guide the policy improvement process.

The reward mechanism used in our MDP is designed to guide the agent towards maximizing the quantum sensor's sensitivity to variations in parameters by optimizing the sequence of actions to achieve the highest possible QFI. This setup is a continuous feedback loop, with rewards given based on immediate outcomes (QFI of the state after each action) and potentially larger rewards after successful episodes.

B. Policy Gradient Agent

We implemented a model free policy gradient RL agent that uses the REINFORCE Algorithm [35]. Policy gradient methods constitute a class of algorithms in reinforcement learning that directly optimize the policy—a mapping from states to actions—without requiring a value function. The optimization is performed in the direction of maximizing the expected return in terms of the Quantum Fisher Information (QFI). The REINFORCE algorithm, a Monte Carlo policy gradient method, adjusts the policy parameters θ in the direction that maximizes the expected reward. Given a trajectory $\tau = (F_{0Q}, \psi_0, U_0, ..., F_{Qt-1}, \psi_{t-1}, U_{t-1})$ the update to the parameters at each step after an episode is given by the gradient of the expected return J_θ with respect to θ [13].

$$\Delta\theta = \alpha \sum_{t=0}^{T-1} \nabla_\theta log\pi_\theta(U_t|\psi_t)(G_t - b_t) \quad (12)$$

where α is the learning rate, π_θ is the policy, G_t is the return time from t and b_t is a baseline, often the state value, used to reduce variance. The return G_t is computed in (13) as the sum of rewards from time t discounted by γ at each step:

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} F_{Qk} \quad (13)$$

where $\gamma \in [0, 1]$ denotes a discount factor.

The policy network architecture designed for this task is a neural network with parameters θ that outputs a probability

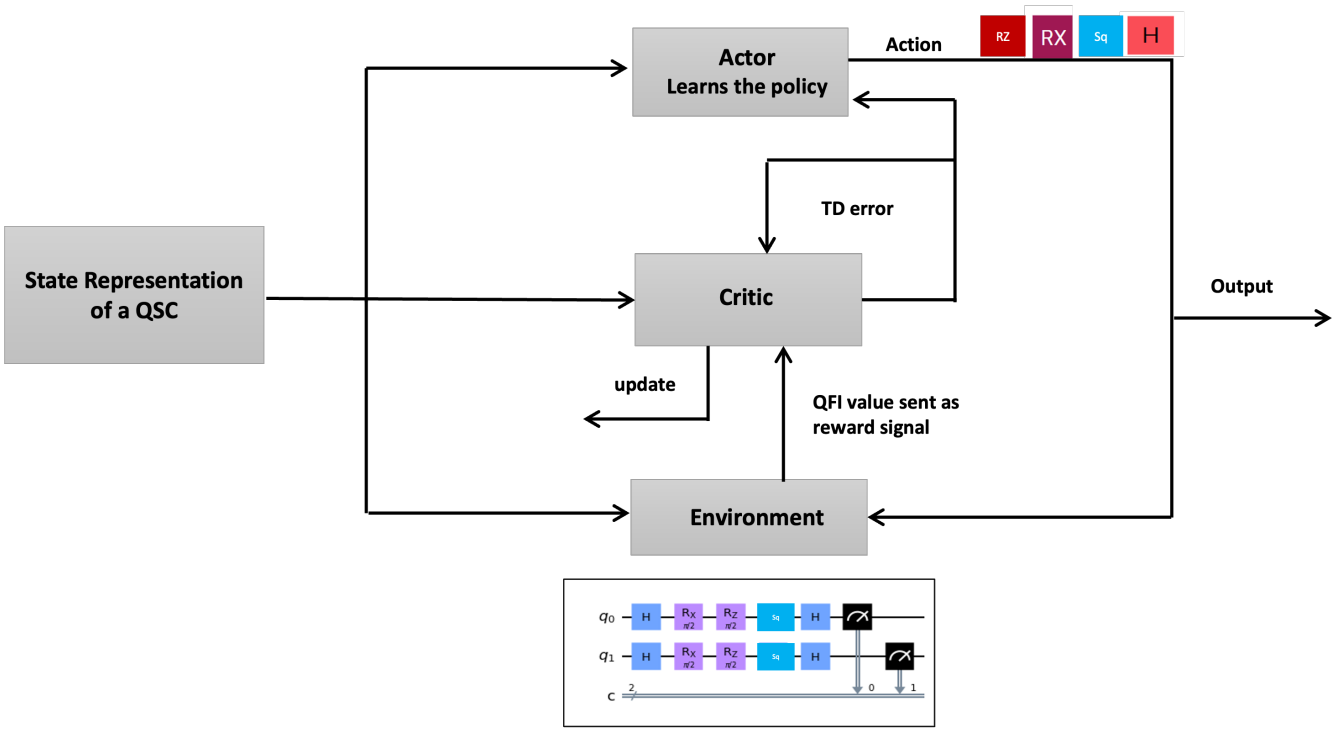

Fig. 4. Architectural Diagram for Actor-Critic Agent.

distribution over the possible quantum gate sequences. The input to the network is the quantum state of the sensor circuit represented in a suitable parametrization, such as the amplitudes and phases of the state vector, while the output layer employs a softmax function to ensure a valid probability distribution over actions:

$$\pi_0\left(U|\psi\right) = softmax(f_\theta\left(\psi\right)) \tag{14}$$

where f_θ is the function computed by the neural network representing the policy.

In our setup, the policy network is trained over episodes, each consisting of a sequence of quantum gate applications leading to the measurement of QFI. Training proceeds by sampling actions from the policy network, applying the corresponding gates to the quantum circuit, measuring the QFI, and using this to compute the gradient of the policy's performance metric. The REINFORCE algorithm's episodic nature is particularly well-suited for this application, as it aligns with the sequential structure of quantum sensor circuit design, where the outcome is only revealed after a complete sequence of gates has been applied. To implement this in practice, we utilize a simulation environment for quantum circuits, interfacing with the policy network to allow for gradient-based optimization of quantum gate sequences.

V. Environment and Experimental Setup

A. Actor Critic Agent

Actor-Critic methods in reinforcement learning combine the strengths of policy-based and value-based approaches, employing two neural network models an actor and a critic [14]–[16]. The Actor proposes actions U which are unitary transformations or quantum gate operations applied to the quantum state. These actions are selected based on a policy that is represented by a probability distribution over the actions, conditioned on the current state of the quantum sensor. The actor network outputs these probabilities, which are shaped by a softmax layer ensuring a valid probability distribution. This network adjusts its parameters to increase the probability of actions that lead to higher rewards, effectively learning to optimize F_Q. Conversely, the critic, evaluates the actions taken by the actor by estimating the value function of the state-action pairs. This estimation helps determine how good the chosen actions are, based on the expected return. The critic thus provides the actor with feedback on the quality of its decisions, guiding the learning process.

The framework operates by approximating both the policy $\pi\left(U|\psi;\theta\right)$ and the value function $V(\psi;w)$ where θ and w is are the parameters of the actor and critic networks, respectively. The objective of the actor is to maximize the expected return by following its policy, and it is updated using the policy gradient given in (15):

$$\nabla_\theta J\left(\theta\right) = \mathbb{E}_\theta\left[\nabla_\theta \log \pi\left(U|\psi;\theta\right) A\left(\psi, U\right)\right] \tag{15}$$

where $A\left(\psi, U\right)$ is the advantage function, representing the benefit of taking action U in a state ψ over following the current policy. The advantage function can be computed as:

$$A\left(\psi, U\right) = Q\left(\psi, U; w\right) - V\left(\psi; w\right) \tag{16}$$

Fig. 5. The figure shows the reward per episode, actor loss per update, the critic loss per update, and the epilson per step for the actor-critic agent for 3000 episode. High and consistent rewards per episode in the Reward per Episode plot indicate that the agent has learned a highly effective policy, achieving the maximum reward of 1 in almost every episode, with only occasional drops. The fluctuating actor and critic losses are typical during training, reflecting ongoing policy refinement and value function estimation improvements. The Epsilon per Step plot shows the expected exponential decay, illustrating a successful transition from exploration to exploitation

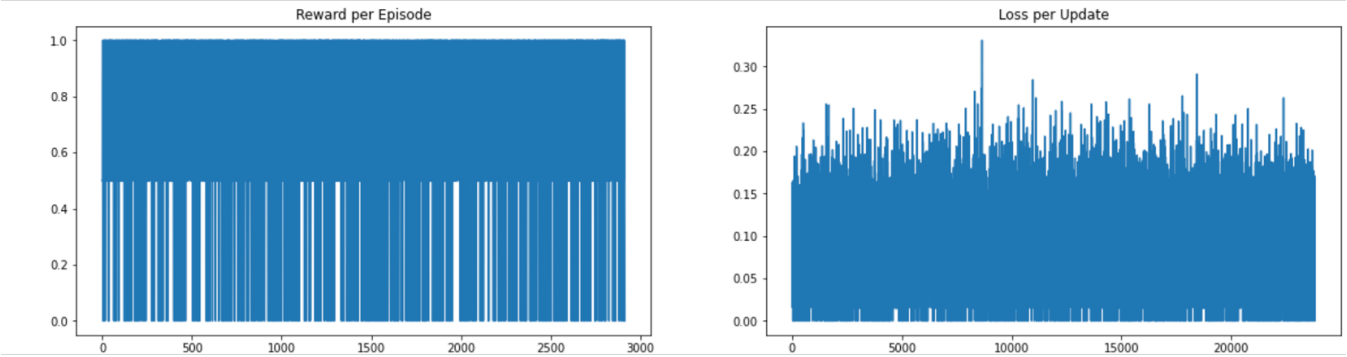

Fig. 6. The graph displays the training metrics of a policy gradient reinforcement learning agent, showing Reward per Episode and Loss per Update. Initially, the rewards are highly variable, indicating the agent's exploration phase, but they become consistently high over time, demonstrating the successful learning of an effective policy. The Loss per Update plot exhibits typical fluctuations due to the high variance in gradient estimates inherent in policy gradient methods, with a general stabilization trend indicating policy convergence

with $Q(\psi, U; w)$ being the action-value function approximated by the critic. The critic's goal is to accurately estimate the value function, which is often done by minimizing the temporal difference (TD) error [17] δ given by (7):

$$\delta = F_Q + \gamma V(\psi^{'}; w) - V(\psi; w) \tag{17}$$

where F_Q is the immediate reward, γ is the discount factor and $\psi^{'}$ s the next state following action U. These networks are parameterized by deep neural networks capable of capturing the complex relationships inherent in quantum systems.

Training involves iteratively updating both networks based on the sampled experiences composed of sequences of gate applications. After each action, the critic updates its value function using the observed reward and the estimated value of the next state, computing the TD error. The actor then updates its policy in the direction suggested by the critic's TD error, effectively using the critic's assessment to adjust its policy gradient. This This formalized training process includes minimizing the loss $L(w) = \delta^2$ for the critic and adjusting θ for the actor using the gradient $\nabla_\theta J(\theta)$ weighted by δ This approach allows for more stable and efficient training by leveraging immediate feedback on the actor's decisions.

The simulation environment for our quantum sensing circuit is crafted to mimic the dynamics of when the system is subject to an external parameter λ, influencing the Hamiltonian governing the system's evolution. This environment, a digital analog of a quantum system, allows for the implementation and testing of reinforcement learning (RL) algorithms in optimizing quantum sensor designs. The Hamiltonian H of the system is represented as:

$$H = \lambda H_\lambda + H_0, \tag{18}$$

where H_λ corresponds to the component of the Hamiltonian that interacts with the external parameter λ, and H_0 denotes the system's inherent Hamiltonian in the absence of λ. The unitary evolution over time τ due to this Hamiltonian is given by:

$$U(\tau) = e^{-iH\tau}, \tag{19}$$

where τ represents the evolution time, and i is the imaginary unit. The quantum sensing circuit simulation environment allows for the application of quantum gates, represented as unitary operations, to manipulate the qubits' states. The environment is designed around a quantum system characterized

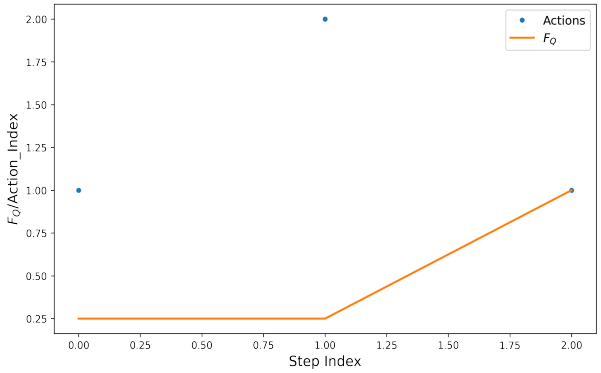

Fig. 7. The figures show the line plot for the actor-critic QFI values. The actions vs steps are also plotted as circles on the graph. The optimal QFI was achieved in 3 steps 0-2. An action index 1 at the first time step indicates a Ry gate, an action index 2 at the second time step indicates a Sq gate and an action index 1 at the third time step indicates another Ry gate applied

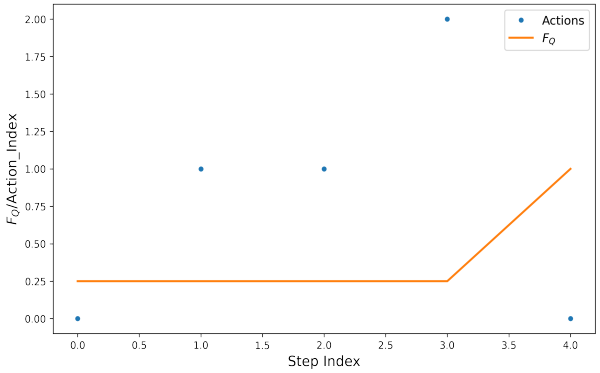

Fig. 8. The figure show the line plot for the policy gradient QFI values observed over five times steps from 0-4, the actions are plotted as circles on the graph. At step index 0, action index 0 = Rx gate was applied, at step index 1 and 2, action index 1 = Ry gate was applied, at step index 3, action 2 = Sq gate and finally at step index 4, action 0 = Rx was applied

by a set of states ψ and a predefined set of unitary operations or actions [Rx, Ry, Rz, Sq] representing the possible quantum gate operations that can be applied. The initial quantum state ψ_0 is set at the beginning of each simulation run, and the system undergo transformations via the unitary operations. These transitions are driven by the application of these unitary transformations, and the resultant new state of the system is computed using matrix operations on the state vector. The reward is the product of F_Q and a boolean indicating whether the F_Q exceeds a predetermined threshold or the maximum number of steps is reached. This reward signals the effectiveness of the action sequence in enhancing the sensor's parameter sensitivity.

We employ two types of RL agents—pure policy gradient and actor-critic agents—each with distinct configurations. A policy gradient agent that utilize a neural network to directly model the policy $\pi(U|\psi;\theta)$ mapping the current state of the quantum system to a probability distribution over possible actions (quantum gates) and an Actor-Critic Agent comprising of two neural networks. The actor-network models the policy $\pi(U|\psi;\theta)$ similar to the policy gradient agent. The critic network estimates the value function $V(\psi;w)$ providing feedback to the actor on the expected return of state-action pairs.

Training done on a CPU involves running numerous episodes where the agents apply sequences of actions to the quantum circuit and observe the resultant F_Q. The networks are updated iteratively based on experiences sampled from a replay buffer, employing strategies like stochastic gradient descent facilitated by the Adam optimizer.

The performance of the RL agents is evaluated against several metrics. Quantum Fisher Information F_Q is the primary metric, measuring the sensitivity of the quantum state to the external parameter with higher QFI indicating greater sensitivity. We also considered the convergence Time The number of episodes required for the RL agent to converge to a policy that consistently yields high F_Q values. The robustness was assessed by introducing variations in the initial states and quantum noise, evaluating the agent's ability to maintain high F_Q under these perturbations. Computational Efficiency was also measured in terms of the computational resources and time required to train the agents to a satisfactory level of performance. These metrics collectively provide a comprehensive evaluation of the RL agents' efficacy in designing quantum sensing circuits, contributing to the broader goal of enhancing quantum metrology through automated, intelligent design strategies.

VI. Result

Our investigation focused on evaluating the efficacy of reinforcement learning (RL) techniques—specifically, policy gradient and actor-critic methods—in optimizing quantum sensing circuits to maximize the Quantum Fisher Information (F_Q). The learning performance of both agents was assessed over 3,000 training episodes for a qubit configuration Q=4 quantum circuit. Figures 7 and 8 illustrate that both agents exhibited an increasing trend in F_Q over the training episodes, reflecting an enhancement in the circuit's sensitivity. The actions taken by the agents are denoted by circles on the graphs, where the y-axis values represent the specific actions at various time steps during the test phase, corresponding to the indices of the actions available in the environment. The line plot depicts the F_Q values observed at each time step during the test phase, with both agents achieving a maximum F_Q of 1. Notably, the policy gradient agent commenced with a F_Q of 0.5, progressively optimizing the encoding states to attain the maximum F_Q.

Detailed analysis of the actions taken by each agent reveals distinct strategies. The actor-critic agent, as shown in Figure 9, followed a sequence of Ry-Sq-Ry gates to achieve the optimal F_Q in three steps (0-2). Specifically, an action index of 1 at the first time step denotes an Ry gate, an index of 2 at the second step denotes an Sq gate, and an index of 1 at the third step denotes another Ry gate. This sequence shows the actor-critic agent's ability to reach the optimal F_Q with minimal gate operations, highlighting its operational efficiency.

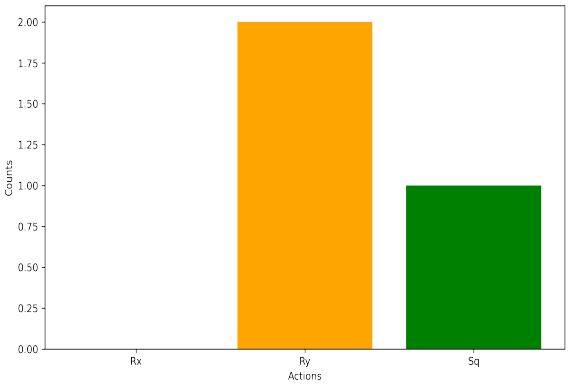

Fig. 9. The figures show the distribution of the actions taken by the actor-critic agent, three actions were taken by the agent and the action follows a sequence Ry-Sq-Ry for the final quantum state with the most optimal QFI of 1

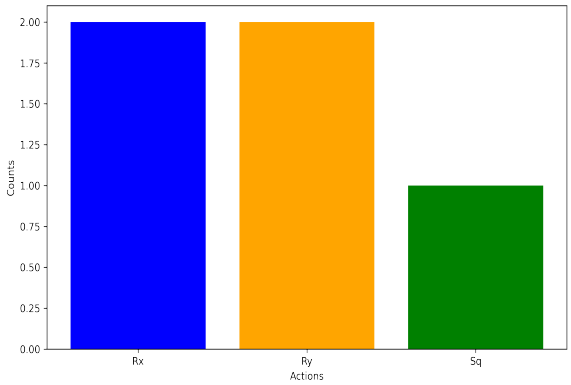

Fig. 10. The figures show the distribution of the actions taken by the policy agent, a total of five actions were taken and the action follows a sequence Rx-Ry-Rx-Sq-Ry for the final quantum state with the most optimal QFI of 1

Conversely, the policy gradient agent's strategy, depicted in Figure 10, involved a sequence of Rx-Ry-Rx-Sq-Ry gates over five steps to achieve the optimal F_Q. At step index 0, an action index of 0 corresponds to an Rx gate; at indices 1 and 2, an action index of 1 corresponds to Ry gates; at index 3, an action index of 2 corresponds to an Sq gate; and at index 4, an action index of 1 corresponds to another Ry gate. Although this sequence involved more steps, the agent was able to iteratively refine the encoding state to achieve the desired F_Q.

A comparative analysis reveals that the convergence time, T_{conv}, was shorter for the policy gradient agent, taking 105.4 seconds compared to 129.8 seconds for the actor-critic agent, indicating a more rapid learning process. However, when evaluating the efficiency based on the number of quantum gates used to achieve the F_Q threshold of 1, the actor-critic method was more efficient, requiring only three gates compared to the five gates required by the policy gradient method. This indicates that while the policy gradient method demonstrated faster convergence, the actor-critic method achieved optimal F_Q with fewer gate operations, emphasizing its gate efficiency. In summary, the actor-critic agent exhibited an advantage in achieving the optimal F_Q with fewer gates, showcasing a more efficient quantum gate sequence, albeit with a slightly longer convergence time. Conversely, the policy gradient agent demonstrated quicker convergence but required more gates to reach the optimal F_Q, reflecting a trade-off between convergence speed and gate efficiency. These findings highlight the distinct advantages and limitations of each RL method in optimizing quantum sensing circuits, suggesting that the choice of method may depend on the specific requirements for convergence speed or gate efficiency.

Also, the Husimi Q-function plot provided in Figures 11 and 12 for final state QSC from both RL agents are identical and it shows a continuous gradient from low values in the center to high values at the top and bottom edges, indicating a state that is broadly distributed in phase space rather than being sharply localized. The symmetry and coherence observed in the plot suggest that the quantum state is a coherent superposition of multiple basis states, indicating entanglement among the qubits. In relation to the QFI, the characteristics of this Husimi plot imply that the quantum sensor has a robust, coherent state that is well-suited for high-precision measurements. The spread in the Husimi plot suggests a balance between sensitivity and noise robustness, essential for optimizing the QFI and enhancing the sensor's performance in detecting small parameter shifts with high accuracy

VII. Discussion

The results from our comparative study on the application of reinforcement learning (RL) methods for quantum sensing circuit design reveal a landscape of potentials and challenges. The practical implications of these findings for quantum circuit design are significant. RL methods can autonomously identify optimal gate sequences, enhancing quantum sensor sensitivity, thus reducing design time and resource expenditure. The increased efficiency and specificity of circuits optimized via RL methods could lead to quantum sensors with higher precision, opening new avenues in quantum metrology.

The simplicity and direct policy optimization feature of policy gradient methods make them accessible and relatively easy to implement, offering a solid baseline for RL applications in quantum systems. The policy gradient method demonstrated faster convergence in our study, highlighting its potential for rapid learning. However, its performance can sometimes be hindered by high variance in complex, high-dimensional quantum state spaces.

In contrast, actor-critic methods integrate the benefits of policy-based and value-based approaches, resulting in more efficient learning and fewer required gates to achieve the optimal F_Q. Their structure allows for more feedback on chosen actions, directly influencing policy updates. Nonetheless, the complexity of actor-critic architectures introduces challenges

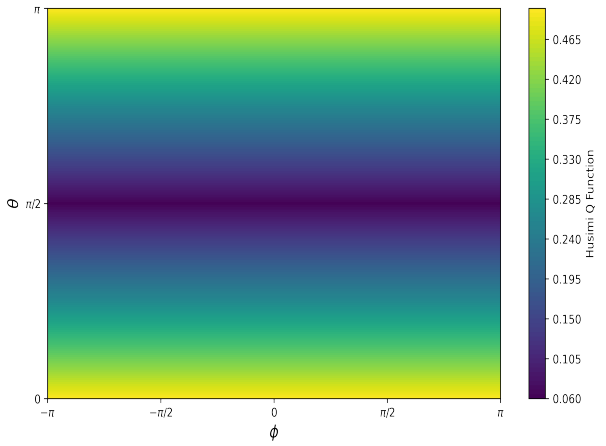

Fig. 11. Husimi Q Function Distribution of the final state achieved by the actor-critic agent

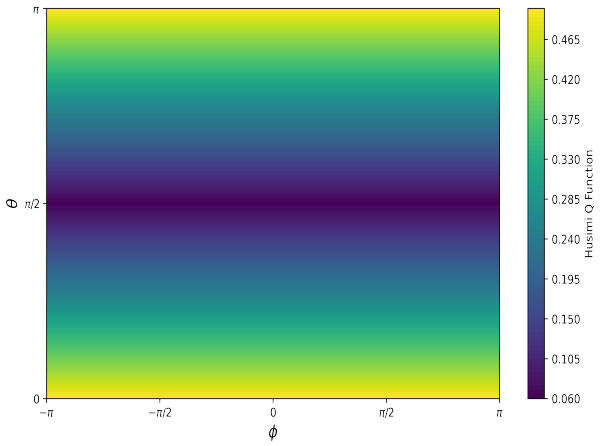

Fig. 12. Husimi Q Function Distribution of the final state achieved by the policy gradient agent

in tuning RL parameters and computational overhead, potentially limiting their use in resource-constrained scenarios.

Choosing between policy gradient and actor-critic methods involves several trade-offs. While actor-critic methods are more complex, their efficiency in learning and circuit design may justify the additional complexity in scenarios where precision and performance are paramount. Computational resources available may dictate the feasible choice, with simpler policy gradient methods being more accessible for limited-resource environments. The specific objectives of the quantum circuit design task—whether maximizing sensitivity, minimizing noise impact, or optimizing F_Q as in this study—can also influence the selection of the RL method.

In conclusion, the exploration of RL methods in quantum sensing circuit design presents a promising frontier for enhancing quantum technologies' capabilities. Understanding the advantages, limitations, and practical implications of these methods provides a foundation for better decision-making in the application of RL to quantum circuit optimization tasks. The actor-critic agent's ability to achieve optimal F_Q with fewer gates shows its operational efficiency, while the policy gradient agent's faster convergence demonstrates its potential for rapid learning. These insights will be crucial in guiding future research and applications.

VIII. Future Work

Future research will focus on advancing the sophistication of reinforcement learning (RL) methods for quantum sensing by incorporating hybrid models and extending their applicability to a broader range of quantum computing challenges. Key areas of investigation will include enhancing the robustness of RL agents against various types of environmental noise and other perturbative factors that affect quantum systems. This will involve rigorous testing and adaptation to ensure reliable performance in the presence of such disturbances.

Furthermore, research efforts will be directed towards scaling these techniques to more complex scenarios, including few-partite quantum sensor models and multiparameter sensing. By addressing these advanced challenges, we aim to significantly increase the practical relevance and effectiveness of RL methods in intricate quantum systems. This extension is anticipated to provide deeper insights into optimizing quantum sensor performance under realistic operational conditions, potentially leading to breakthroughs in quantum metrology and other applications.

Overall, the goal is to develop RL approaches that not only enhance the precision and efficiency of quantum sensors but also ensure their robustness and scalability in real-world environments. This will involve a multidisciplinary effort, combining insights from quantum physics, machine learning, and noise mitigation strategies, to push the boundaries of what is achievable in quantum sensing and computing.

IX. Conclusion

Our investigation into the application of reinforcement learning (RL) methods for quantum sensing circuit design has revealed significant advancements in optimizing quantum Fisher Information (QFI) and enhancing sensor sensitivity. By comparing the policy gradient and actor-critic methods, we demonstrated that both approaches are effective in learning efficiency and achieving optimal QFI values of 1. Each method exhibits unique strengths: the policy gradient method converges faster, while the actor-critic method requires fewer quantum gates, demonstrating its operational efficiency.

This study highlights the transformative potential of RL in automating and optimizing quantum circuit design. These methods streamline the design process and uncover non-intuitive gate sequences, thereby enhancing the capabilities of quantum sensors. The actor-critic framework, with its detailed feedback mechanism and dual network architecture, and the policy gradient method, with its simplicity and rapid convergence, both emerge as powerful tools for addressing the complexities of quantum systems.

The implications of these findings extend beyond theoretical advancements. Efficiently designed quantum sensors could

bring about significant improvements in precision measurements across various fields. For instance, in gravitational wave detection, the enhanced sensitivity of these sensors could lead to more accurate and earlier detections. In medical imaging, quantum sensors could improve the resolution and accuracy of diagnostic tools. Environmental monitoring could benefit from the ability to detect minute changes in physical parameters, leading to better-informed decisions in climate science and resource management.

Our research lays the groundwork for further exploration of RL techniques in quantum technologies. Future studies could focus on integrating hybrid models, enhancing robustness against environmental noise, and scaling to more complex quantum systems. Advancing RL methods in these directions will contribute to the development of next-generation quantum sensors that are not only more efficient but also more resilient and versatile in real-world applications.

In conclusion, the deployment of RL methods in quantum sensing circuit design represents a promising frontier in quantum technology. Our findings emphasize the potential of these techniques to significantly improve the performance and applicability of quantum sensors, driving innovations that could have a profound impact on science and technology.

ACKNOWLEDGMENT

Thanks to the individuals in the group working on projects for the National Science Foundation Grant No. OMA 2231377 for valuable insights.

REFERENCES

[1] C. Degen, F. Reinhard, and P. Cappellaro, "Quantum sensing," Rev. Mod. Phys. 89, 035002, July 2017.
[2] A.Steinberg, A light touch. Nature 463, 2010, pp. 890–891.
[3] J. Aasi et al.,"Enhanced sensitivity of the LIGO gravitational wave detector by using squeezed states of light," Nat. Photonics 7, 613 2013.
[4] B. P. Abbott et al. "Observation of Gravitational Waves from a Binary Black Hole Merger," (LIGO Scientific and Virgo Collaborations), Phys. Rev. X 6, 041015, 2016.
[5] M. Tse et al., "Quantum-Enhanced Advanced LIGO Detectors in the Era of Gravitational-Wave Astronomy," Phys. Rev. Lett. 123, 231107, 2019.
[6] M. A. Taylor and W. P. Bowen, "Quantum metrology and its application in biology," "Phys. Rep. 615, 1 2016.
[7] C. W. Chou, D. B. Hume, T. Rosenband, and D. J. Wineland, "Optical Clocks and Relativity,"Science 329, 1630, 2010.
[8] T. Bothwell, C. J. Kennedy, A. Aeppli, D. Kedar, J. M. Robinson, E. Oelker, A. Staron, and J. Ye, Nature, "Resolving the gravitational redshift across a millimetre-scale atomic sample," (London) 602, 420, 2022.
[9] P. Busch, "The Role of Entanglement in Quantum Measurement and Information Processing". International Journal of Theoretical Physics 42, 2003, pp937–941.
[10] G. Goldstein et al., "Environment-Assisted Precision Measurement," Phys. Rev. Lett. 106, 140502, April 2011
[11] A. Sone, M. Cerezo, J. L. Beckey and P. J. Coles, "Generalized measure of quantum Fisher information," Phys. Rev. A 104, 062602, December 2021.
[12] F. Woergoetter and B. Porr, "Reinforcement learning," Scholarpedia, 3(3):1448, 2008.
[13] J. Peters, "Policy gradient methods," Scholarpedia, 5(11):3698, 2010
[14] J. Peters, S. Schaal, "Natural Actor-Critic," Neurocomputing, Volume 71, Issues 7–9, 2008, pp 1180-1190
[15] V. R. Konda, J. N. Tsiksiklis, "Actor-Critic Algorithms", Advances in neural information processing systems 12, 1999
[16] V. R. Konda and V. S. Borkar. "Actor-critic like learning algorithms for Markov decision processes,". SIAM Journal on Control and Optimization, 38(1) :94-123, 1999.
[17] R. H. Crites, A. G. Barto, "An Actor/Critic Algorithm that Equivalent to Q-Learning," MIT Press, Cambridge, MA, USA, 1994
[18] T. Xiao, J. Fan, and G. Zeng, "Parameter estimation in quantum sensing based on deep reinforcement learning," npj Quantum Inf 8, 2 January 2022.
[19] R. Kaubruegger, A. Shankar, D. Vasilyev, P. Zoller, "Optimal and Variational Multi-Parameter Quantum Metrology and Vector Field Sensing," American Physical Society, PRX Quantum 4, 020333, June 2023
[20] S. A. Haine and J. J Hope, "Machine-Designed Sensor to Make Optimal Use of Entanglement-Generating Dynamics for Quantum Sensing," Phys. Rev. Lett. 124, 060402, 2020.
[21] C. Egerstrom, "A Mathematical Introduction to Quantum Sensing," Unpublished.
[22] A. S. Holevo, "Probabilistic and Statistical Aspects of Quantum Theory", North-Holland Series in Statistics and Probability (North-Holland Publishing Company, Amsterdam, 1982).
[23] V. Giovannetti, S. Lloyd, and L. Maccone, "Advances in quantum metrology," Nat. Photonics 5, 222, 2011.
[24] R. Demkowicz-Dobrzanski, J. Kołody ´ nski, and M. Gu¸ ´ ta, "The elusive Heisenberg limit in quantum-enhanced metrology," Nat. Commun. 3, 1063, 2012.
[25] L. Pezzè, A. Smerzi, M. K. Oberthaler, R. Schmied, and P. Treutlein, "Quantum metrology with nonclassical states of atomic ensembles," Rev. Mod. Phys. 90, 035005, 2018.
[26] J. J. Meyer, J. Borregaard, and J. Eisert, "A variational toolbox for quantum multi-parameter estimation," npj Quantum Inf. 7, 89, 2021.
[27] M. Hayashi, "Quantum Information Theory," Mathematical Foundation, 2nd ed. Springer, New York, 2016.
[28] J. Liu, H. Yuan, X.-M. Lu, and X. Wang, "Quantum Fisher information matrix and multiparameter estimation,", J. Phys. A: Math. Theor. 53, 023001 2020.
[29] A. Sone, M. Cerezo, J. L. Beckey, and P. J. Coles, "Generalized measure of quantum Fisher information," Phys. Rev. A, 104 6 062602, 2023, pp 13.
[30] X. Li, J. Cao, Q. Liu, M. K. Tey, L. You, "Multi-parameter estimation with multi-mode Ramsey interferometry," New J. Phys. 22 043005, 2022
[31] R. Kaubruegger, D. V. Vasilyev, M. Schulte, and K. Hammerer and P. Zoller. "Quantum Variational Optimization of Ramsey Interferometry and Atomic Clocks," Phys. Rev. X 11, 041045, 2021.
[32] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of Markov reward processes," IEEE Transactions on Automatic Control, VOL. 46, NO. 2, FEBRUARY 2001
[33] D. Bertsekas and J. Tsitsiklis, " Neuro-Dynamic Programming. Athena Scientific," The MIT Press, Cambridge, MA, 1996
[34] R.S. Sutton, D. McAllester, S. Singh and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems", 12, 1999.
[35] R.J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning,". Mach Learn 8, 229–256 1992.