Proximal Policy Optimization in Uncoordinated and Distributed Multi-Agent Resource Allocation in Cognitive Radio Network

Ankita Tondwalkar
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester,USA
at3235@g.rit.edu

Andres Kwasinski
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester,USA
axkeec@rit.edu

***Abstract*—This paper studies the learning performance of the Proximal Policy Optimization (PPO) algorithm in the challenging use case of uncoordinated and distributed multi-agent (UDMA) resource allocation in a cognitive radio network (CRN) via an underlay dynamic spectrum access (DSA) paradigm. The challenge is to establish whether PPO will converge faster to the best performance policies than UDMA-DQL in the non-stationary environment. The simulation results show that the UDMA-PPO achieves faster learning performance than UDMA table-based Q-learning and UDMA deep Q-learning (DQL) and is robust to the non-stationary environment. The UDMA-PPO achieves better performance with about 70% less training steps compared to the UDMA-Table and achieves similar performance with 25% less training steps than the UDMA-DQL algorithm.**

***Index Terms*—cognitive radios, uncoordinated distributed multi-agent PPO, underlay dynamic access**

I. Introduction

To sustain the increasing demand for wireless applications, it is important for wireless networks to efficiently use the radio spectrum. The dynamic spectrum access (DSA) and spectrum sharing is important from the point of view of efficient spectrum utilization, since conventional static radio spectrum allocation results in spectrum under-utilization [1], [2], [3].The work [4] uses cognitive radios (CRs) in a multi-agent setting to efficiently utilize radio spectrum because of their ability to adapt and learn from the dynamic environment. The model-free reinforcement learning (RL) approach naturally aligns with the CR paradigm, making it a suitable machine learning candidate for resource allocation. However, the use of RL in a general wireless network is a challenging problem statement.

In a distributed and uncoordinated network setup, the transmission from one CR affects the environment perceived by other CRs (acting as interference), resulting in a non-stationary environment. The multi-agent interaction coupled with the non-stationary environment and the interplay of agents due to their transmissions may not always lead to the best policy selection by the agents. For coordinated multi-agent RL set up, the agents can learn to find the best solution (optimal policy), but the most challenging scenario is when the agents are distributed and uncoordinated as a result of a non-stationary environment. In [5], we presented the UDMA-DQL technique with the combined effect of a deep neural network with learning in exploration phases and a Best Reply Process with Inertia for the gradual learning of the best policy. The work in [5], analytically showed that the UDMA-DQL algorithm can achieve convergence with probability one when learning is allowed for an arbitrarily long time (as is prevalent in RL training) in a non-stationary environment. Policy gradient methods [6] are known to achieve significantly faster convergence than DQN algorithms along with improving the reward in a multi-agent stationary environment. In this paper, our aim is to determine whether PPO will converge in a non-stationary environment and evaluate its learning performance against UDMA-DQL and UDMA-Table in terms of learning time (number of training steps) required to find the optimal solution.

The work in [7] presents a multi-user CR framework for sensing and power allocation to achieve maximum throughput using the actor-critic approach. The assumption in [7] is that the CRs operate in a coordinated fashion, where the fusion centre (FC) aggregates the sensing results of the CRs to estimate the vacant/occupied status of each channel. In [8], the authors present a RADDPG (Resource Allocation Deep Deterministic Policy Gradient) algorithm aimed at jointly optimizing the interference threshold and the power distribution. The work in [9] proposes a centralized power allocation technique for a cellular network using DQL. The work in [7], [8] and [9] models a multi-agent interplay where there is coordination among the agents or learning at the central node and does not deal with the more complex scenario as the one considered in our work with a non-stationary environment. The work in [10] uses a single agent PPO-based algorithm that allows a CR to realize continuous power control and share the spectrum with PN. The work in [11] uses PPO to address the allocation of resources for unmanned aerial vehicles (UAV) with the assumption of favorable channel conditions and quasistatic user locations.In [12], adaptive MIMO transmission is coupled with PPO to maximize spectral efficiency subject to bit error rate (BER) performance in a non-

ISBN 978-3-903176-73-7 © 2025 IFIP

stationary environment. In [13], the authors use a centralized learning with decentralized execution for multiple base stations with a single UE in an effort to improve the sum and maximum user throughput. Although many works [10],[11],[13] have studied PPO convergence in a stationary environment, our RL approach is distributed, resulting in an uncoordinated and dynamic interaction between the CRs.

Our main contribution is to confirm that the multi-agent actor-critic PPO algorithm when applied in a non-stationary environment arising from its use for multi-agent distributed and uncoordinated resource allocation in a CRN converges to the equilibrium policies faster than other competing RL approaches. The main contributions that are presented in this work are as follows:

- To the best of our knowledge, our work is the first to present a multi-agent PPO approach for uncoordinated and distributed resource allocation in a non-stationary CRN.
- The simulations demonstrate performance advantage of UDMA-PPO in the challenging case of a non-stationary environment because of multiple uncoordinated active CR agents that are learning through a shared wireless environment.
- The experimental results also validate the stable and comparatively faster convergence to optimal solution behavior of the UDMA-PPO compared to the baseline approaches (UDMA-Table and UDMA-DQL) and a robust performance in a non-stationary environment.

The organization of this paper is as follows. The system model is described in Section II. In Section III, we discuss in detail the PPO-based distributed and uncoordinated multi-agent resource allocation. The results and analysis are given in Section IV, followed by the conclusion in Section V.

II. System Model and Problem Formulation

The system illustrated in Fig. 1, consists of a primary network (PN) and a secondary network (SN), both operating in a common geographical area and sharing the same radio spectrum band via the underlay DSA mechanism. The

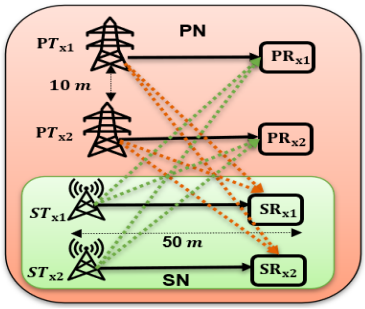

Fig. 1: Network model for primary and secondary network.

ultimate goal for an SN is to be able to share with PN the radio spectrum it owns, to facilitate efficient spectrum management. Therefore, we use the underlay DSA model where both networks can transmit simultaneously subject to the SN interference being within the acceptable limits. For illustration purposes, we only show a primary and secondary transmitter-receiver pair in Fig. 1. However, the PN consists of M access points (APs) and the SN is formed by N CR links that perform autonomous learning, i.e., the CR nodes are distributed and uncoordinated during resource allocation and there is no information exchange between the two networks.

Since we use underlay DSA, the transmit power for the CRs does not exceed the threshold, which otherwise may cause interference to the PN. The interference from a CR negatively affects the Signal-to-Interference-plus-Noise ratio (SINR) and thereby the throughput at PN links. For the system model as shown in Fig.1, the SINR for the i^{th} PN link is, $\gamma_i^{(p)}$, given by (1), where $P_i^{(p)}$ is the transmit power at the i^{th} primary link transmitter $(i = 1, 2, ..M)$, $P_j^{(s)}$ is the transmitter power at the j^{th} SN link, $G_{ij}^{(pp)}$ is the channel gain between the i^{th} PN AP and the j^{th} PN receiver, $G_{ij}^{(ps)}$ is the gain from j^{th} transmitting SU to the receiver in the i^{th} primary link and σ^2 is the background additive white Gaussian noise (AWGN) power. To ensure high performance in wireless systems, we assume that the PN employs an Adaptive Modulation and Coding (AMC) technique that adapts the modulation scheme and the channel coding rate (together known as AMC mode) based on the SINR of the transmission link. The ultimate aim for the CRs is to achieve the largest possible SINR to reach the largest possible throughput. The SINR for the i^{th} SN link is evaluated as:

$$\gamma_i^{(p)} = \frac{G_{ii}^{(pp)} P_i^{(p)}}{\sum_{j \neq i}^{M} G_{ij}^{(pp)} P_j^{(p)} + \sum_{j=1}^{N} G_{ij}^{(ps)} P_j^{(s)} + \sigma^2}, \tag{1}$$

$$\gamma_i^{(s)} = \frac{G_{ii}^{(ss)} P_i^{(s)}}{\sum_{j \neq i}^{N} G_{ij}^{(ss)} P_j^{(s)} + \sum_{j=1}^{M} G_{ij}^{(sp)} P_j^{(p)} + \sigma^2}, \tag{2}$$

where $P_i^{(s)}$ is the transmit power at the i^{th} SN link, $G_{ij}^{(ss)}$ is the channel gain between the i^{th} transmitting SN cognitive radio and the j^{th} receiving SN node, and $G_{ij}^{(sp)}$ is the channel gain between the i^{th} SN CR and a j^{th} PN receiver.

$$S_i[0] = \tilde{G}_{in}^{(ps)} P_n^{(p)}[0] + \sum_{\substack{j=1 \\ j \neq n}}^{M} \tilde{G}_{ij}^{(ps)} P_j^{(p)}[0] \tag{3}$$

$$S_i[1] = \tilde{G}_{in}^{(ps)} P_n^{(p)}[1] + \sum_{\substack{j=1 \\ j \neq n}}^{M} \tilde{G}_{ij}^{(ps)} P_j^{(p)}[1] + I_{ss} \tag{4}$$

However, the underlay DSA mechanism limits the transmit power of the CRs in the SN. Owing to the direct correspondence between the throughput and SINR, the CRs evaluate their transmission effect on the PN by estimating the relative

change in the throughput at its nearest PN link according to the two-stage approach explained in detail in [14]. In the first i.e., listening stage, each CR listens and monitors the power signal at the i^{th} SU transmitter given by (3) where, $\tilde{G}_{in}^{(ps)} P_n^{(p)}[0]$ is the received power for the nearest PN transmitter (subscript n represents the nearest link to the i^{th} CR, out of all the active transmissions) and $\tilde{G}_{ij}^{(ps)} P_j^{(p)}[0]$ is the received power from the remaining active users in the PN. During the second i.e., probing stage, every CR sends a probing signal with the same power that interferes with the PN. The observed power at the i^{th} SU transmitter during this stage is given by (4) where I_{ss} is the combined interference from the SN and $P_j^{(p)}[1]$ is the adapted power for the j^{th} PN link.

$$I_{ss} = \sum_{\substack{j=1 \\ j \neq i}}^{M} \tilde{G}_{ij}^{(ss)} P^{(s)} \tag{5}$$

The signal power, during both the listening and probing stage consists of cross-channel gain information between the i^{th} SN transmitter and its PN receiver $G_{in}^{(ps)}$. In most of the high-performing wireless systems, it is assumed that the PN employs an Adaptive Modulation and Coding (AMC) technique [15] [16] that alters the modulation scheme and the coding rate (together known as the AMC mode) based on the SINR of the transmission link. Therefore, for interference caused by the probing signal, every PN receiver sends a control signal to its transmitter to adjust the transmit power and the coding rate. In a fully autonomous network, it is not practical to assume knowledge of all parameters for precise CR cross-channel gain calculations. The work [14] uses an artificial neural network to calculate the transmission effect of CRs on PN without knowing the cross-channel gains.

The underlay DSA operates in a way that it allows the SN to share the same radio band as the PN, provided that the interference limit condition is met. With the same analogy, the limit is now set on the relative throughput change at their nearest PN link (understood hereafter to be the one that is received with largest power). Additionally, there is no exchange of information between the two networks, and the CRs cannot access the PN's feedback channel, as this would be a violation of the separation between the two networks (if nodes from the two networks share access to a common control channel, they form in effect a single network with nodes of different type). This CR operation makes for a distributed and uncoordinated scenario with the PN and between each other.

As is standard in most wireless communications, the PN and SN nodes allocate resources. Therefore, the problem at hand is the distributed and uncoordinated resource allocation for cognitive agents in the SN. For the purpose of this work which focuses on evaluating learning performance of UDMA-PPO, where each CR independently learns its best resource allocation decision without any prior knowledge of the non-stationary wireless conditions, we focus on the allocation of transmit power only (since the problem under study does not change if we consider additional transmission parameters). The ultimate goal for every CR in the SN is to find the transmit power that yields the largest possible throughput without violating the limit on the relative throughput change on their nearest PN link. This issue is characterized by CRs as learning agents within the system of uncoordinated multi-agent interplay, where the action of one learning agent may affect the state of the wireless environment for other learning agents (transmission of one CR can be a source of interference to the rest of the SN and PN). The lack of coordination between the two networks and even within the SN further convolutes the already challenging power allocation problem for cognitive agents. The multi-agent interaction results in a non-stationary environment that influences the CRs learning by introducing noisy feedback on the transmit powers and presumably causing non-convergent behavior. As will be seen, the UDMA-PPO technique effectively addresses these challenges, ensuring faster learning to an optimum (when such an optimum can be defined).

III. PPO for Uncoordinated and Distributed Multi-agent Resource Allocation

The underlying assumption in our work is that the agents have no prior knowledge about surrounding conditions and there is no exchange of information between the PN and the SN or between the CRs. This issue can be addressed by autonomous learning of the environment by the CRs. The need for autonomous learning of the environment conditions can be thought of as the need for a model-free approach which can be achieved by employing the Observe-Orient-Decide-Act framework [17]. The model-free RL does not explicitly model the environment dynamics, but learns a policy (mapping from states to actions) or a value function (which estimates expected cumulative discounted rewards) or both directly from the interactions with the environment. This approach of learning aligns with the Observe-Orient-Decide-Act framework and hence we use RL to solve the distributed and uncoordinated power resource allocation for multiple agents.

To apply RL as the solution to our problem, we define the action space, state space and reward function as follows:

Action space: In a distributed and uncoordinated multi-agent scenario, even though the dynamics between the state change is unknown, it is known that it will be strongly affected by actions taken by the agents. The action space is defined as the set $\mathcal{A} = \{0, P_0, P_1, \ldots, P_L\}$ of possible discrete transmit powers (where '0' indicates no transmission). We assume without loss of generality that all the CRs have the same action space as each agent chooses an action independently of the other agents in our uncoordinated and distributed setup.

State space: The different conditions that the environment could be in can be thought of as a set of states. The state space is the set of all possible states $s_t \in \mathcal{S}$ that the environment can be in. The state space for our scenario consists of two states, S_0 and S_1, where state S_0 corresponds to the conditions when the interference limit for underlay DSA is met, and state S_1 corresponds to the case when the condition is not met. As mentioned previously, the underlay DSA limit is characterized

based on the effect, the SN interference has on the PN through the relative change of throughput. The PN network in our setup transmits using AMC and therefore, can absorb up to a limit an increase in combined background noise and interference without any noticeable decrease in average throughput [14]. This property of AMC is shown in Fig. 2, shows the relative average throughput change as a function of background noise power for a network transmitting with AMC. Since there is no exchange of any information between the two networks, for the PN, the interference from SN is indistinguishable from an increase in the background noise power. However, since the background noise power is usually constant, Fig. 2 effectively is showing PN relative average throughput change hereafter denoted as $T_{\%}$ as a function of the increase in interference from the SN. This relative average throughput change in the PN, can be expressed as a function of the interference from the SN through the approximate expression [14],

$$T_{\%} = \frac{-1}{\xi} \log_2 \left(1 + \Gamma * 10^{I/10}\right), \tag{6}$$

where ξ is the PN spectral efficiency in absence of any SN transmission, Γ is the signal-to-noise ratio (SNR) gap factoring the practical coding and transmission, and I is the interference caused by SN on the PN. As described in [14], each CR manages its transmission effect on the PN by evaluating the relative throughput change at the PN link that is received with the highest power (usually the nearest link but not necessarily always the case due to random fading phenomena). With ϵ as the limit on the relative throughput change, the state space which is not global for the environment, but one that is local to each of the CRs is defined as:

$$s_t = \begin{cases} S_0, & \text{if } |T_{\%}^{(i)}| \leq \epsilon, \\ S_1, & \text{else .} \end{cases} \tag{7}$$

Reward: The need to maximize each CRs throughput is represented by the reward function and is defined as a function of the state and the action taken to transition to a new state:

$$R_t^{(i)}(a_t, s_t) = \begin{cases} 10^{T_i}, & \text{if environment state is } S_0, \\ 0, & \text{if environment state is } S_1, \end{cases} \tag{8}$$

where T_i is the throughput on the i^{th} SN link. This reward function was aimed to augment the difference between rewards for actions that correspond to low SINRs.

The problem described in Section II is addressed with the PPO framework that enables the agent to learn and adapt to its surrounding conditions. Proximal Policy Optimization [18] is a policy-based algorithm comprising two neural networks: one, with parameters denoted as θ and called the "actor", models the policy, and a second, with parameters denoted as ϕ and called the "critic", models the value function. Just like any other RL algorithm, PPO also aims to learn a policy $\pi_\theta(a|s)$ that maximizes the expected cumulative reward. A policy is a mapping from the state space to the action space. At each time step t, the agent interacts with the environment and collects trajectories <*current state, next state, action taken, rewards*> to update its policy and value function. The training of the PPO agent uses an advantage function Adv_t as given by (9) which measures in a particular state s_t how much worse or better an action a_t is compared to the expected action value for that state under the current policy. It is computed as the sum of the discounted future rewards minus the value function at the current state.

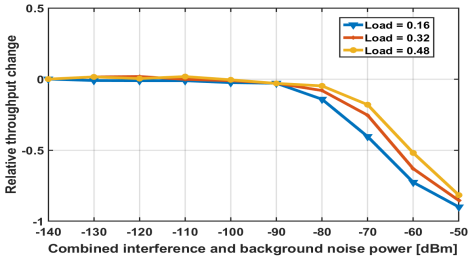

Fig. 2: Relative throughput change vs background noise power in the PN.

$$Adv_t = \sum_{t=0}^{\infty} \gamma^t r_t - V_\phi(s_t) \tag{9}$$

where $\gamma \in [0, 1]$ is the discount factor that manages the balance between short-term and long-term rewards. The value function $V_\phi(s)$ estimates the expected total reward from a given state s_t. If $Adv_t > 0$, it implies that taking a_t in state s_t yields a higher reward than the expected reward whereas $Adv_t < 0$ indicates that the a_t performed worse than the expected return. When $Adv_t = 0$, the action a_t is at par with the expected reward. In PPO, the actor and critic loss each serve a different purpose in optimizing the policy. The critic loss function is updated by minimizing the mean squared error between the estimated and the target value as,

$$L_{critic}^{(t)}(\phi) = \frac{1}{2}(V_\phi(s_t) - (r_t + \gamma V_\phi(s_{t+1})))^2 \tag{10}$$

The policy is updated using the PPO objective (actor loss) function, which aims to maximize the expected advantage while avoiding large policy changes. The actor loss function for a single time step is given as,

$$L_{actor}^{(t)}(\theta) = \min\left(r_t(\theta)\, Adv_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) Adv_t\right) \tag{11}$$

$$r_t(\theta) = \pi_\theta(a_t|s_t) / \pi_{\theta_{old}}(a_t|s_t) \tag{12}$$

where ϵ is a hyperparameter that controls the clipping range within $[1-\epsilon, 1+\epsilon]$, $r_t(\theta)$ is the ratio of the updated policy probabilities to the old policy probabilities.

$$S[\pi_\theta(a_t|s_t)] = -\sum_a \pi_\theta(a_t|s_t) \log \pi_\theta(a_t|s_t) \tag{13}$$

To further improve the stability of training, an entropy regularization term can also be included in the objective function. The

entropy $S[\pi_\theta(a_k|s_k)]$ measures the uncertainty or randomness in the policy distribution and is given by (13). We have not used entropy regularization in our PPO framework so the overall objective function will focus only on the critic and actor loss. The critic loss trains the critic network which approximates $V_\phi(s)$. The critic loss in (10) helps the critic to provide accurate value function estimation. The actor loss optimizes the actor (policy) network by updating the action probabilities using a temperature-controlled distribution over action values. The actor loss is the clipped surrogate objective function as shown in (11).

TABLE I: Training parameters for PPO

Parameter	Value
Number of actor and critic layers	3
Neurons for actor and critic layer	{32,64,64}
Activation function for all layers	Relu
Optimizer	Adam
Actor learning rate α_a	0.0003
Critic learning rate α_c	0.0002
Discount factor γ	1- 0.0001
Temperature τ	12
Temperature decay rate ρ	1.0030
Clipping parameter ϵ	0.1

Instead of the standard entropy regularization used in the PPO framework, we use the Boltzmann exploration that influences the action selection using a temperature controlled distribution over the action values. The temperature parameter τ balances the trade-off between exploration and exploitation. For each action a_t in state s_t, the probability of taking that action is:

$$P(a_t|s_t) = \frac{exp(Adv_t/\tau_t)}{\sum_{\mathcal{A}}(exp(Adv_t/\tau_t)}, \tag{14}$$

where $\tau_t = \tau/\rho$. The temperature τ balances the trade off between exploration and exploitation. Higher the τ, higher is the exploration as actions have similar probabilities. Lower the τ, higher is the exploitation as the agent increasingly chooses actions with positive advantage function. The decay rate ρ controls the rate at which the temperature decreases, reducing exploration over time and allowing the agents to exploit actions with higher advantage.

IV. Results and Discussion

Our primary focus in this work is to evaluate the learning performance of the UDMA-PPO algorithm in a non stationary environment. The performance is evaluated as the number of training steps required to learn equilibrium policies faster than the UDMA-DQL and UDMA-Table approaches. All experiments are based on Monte Carlo simulations. Each simulation instantiates a network set up where a PN shares a 180 kHz radio band with SN through underlay DSA. The PN consists of nine access points (APs) with each AP at a distance of 200 m from its adjacent APs. Of the nine APs only seven APs that are chosen at random for each Monte Carlo run transmit to the receiver which is randomly located within the coverage area of the corresponding AP. The SN consists of two point-to-point links with each of the two CR transmitters randomly located within the PN grid and the two receivers located anywhere within 50 m of their corresponding transmitters. This setup models a network of small radio devices (the SN) that share spectrum with an incumbent network (the PN) of larger devices. For underlay spectrum sharing, the limit on the change in relative throughput in PN links is set to 5%.

Algorithm 1 Proximal Policy Optimization Training

Input: Policy network $\pi_\theta(a, s)$, Value network $V_\phi(s)$, learning rate α, discount factor γ, clipping parameter ϵ, environment env
Output: Policy function π_θ, Value function V_ϕ
Initialize θ and ϕ parameters
for each Monte Carlo run **do**
  Initialize environment env and get initial state s_0
  for each update step $k = 1, 2, \ldots, K$ **do**
    Collect set of trajectories <*current state, next state, action taken, rewards*> by running policy π_θ
    Compute rewards r_t
    Compute advantage function Adv_t
    $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$
    Compute the surrogate objective (actor loss): $L^{(t)}_{actor}(\theta) = min(r_t(\theta)Adv_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)Adv_t)$
    Compute critic loss: $L^{(t)}_{critic} = \frac{1}{2}(V_\phi(s_t) - (r_t + \gamma V_\phi(s_{t+1}))^2$
    Update actor network using step 10
    Update critic network using step 11
  end for
end for

The transmitted signal is affected by the penetration, shadowing and path loss [19] in both the PN and SN where the received signal power P_{Rx}, the transmitted signal power P_{Tx}, the transmitter-receiver distance in km d, and the shadowing loss S are related as $P_{Rx} = P_{Tx} - 128.1 - 37.6 \log d - 10 - S$. The shadowing loss is modeled as a zero-mean Gaussian random variable with 6 dB standard deviation, the penetration loss is constant at 10 dB and P_{Rx} and P_{Tx} are measured in dBs. The AWGN power for all the links is set to -130 dBm. In the simulation, we adopted the LTE AMC scheme (consisting of fifteen possible modes made from the combination of different channel coding rates with one of QPSK, 16-QAM or 64-QAM modulation schemes) [20]. Furthermore, we assume that PN transmitters can adapt transmit power in the range of -20 to 40 dBm by employing the iterative power allocation algorithm in [21] and CRs operate with an action space, consisting of fourteen transmit power levels between -10 and 20 dBm equally spaced by 2.5 dBm, with no transmission (transmit power zero). We implement the actor and critic deep neural network framework, with input being the environment

state for the actor as well as the critic and output being the policy π_θ for actor network and expected return (cumulative future rewards) V_ϕ for the critic network respectively. The PPO algorithm is shown in Algorithm 1 and the training parameters are listed in Table I. In order to have a fair comparison with UDMA-DQL and UDMA-Table, we use the same modified reward function for the PPO framework. The reward function maintains the characteristic of uncoordinated and distributed design and results in a global optimum solution. This reward is defined as:

$$R_t^{(i)}(a_t, s_t) = \begin{cases} 10^{\sum_{j=1}^{N} T_j}, & \text{if environment state is } S_0, \\ 0, & \text{if environment state is } S_1, \end{cases} \quad (15)$$

Even though the reward function in (15) assumes that the CRs can inform of their throughput to all other CRs, we would like to highlight this departure from the no co-ordination setup is only to instantiate the particular solution of a global optimum without affecting the nature of our algorithm. We evaluate the learning performance of PPO in a non-stationary and uncoordinated environment by measuring the percentage of times the UDMA-PPO finds optimal solution after running for a number of training steps. Fig. 3 demonstrates the learning performance

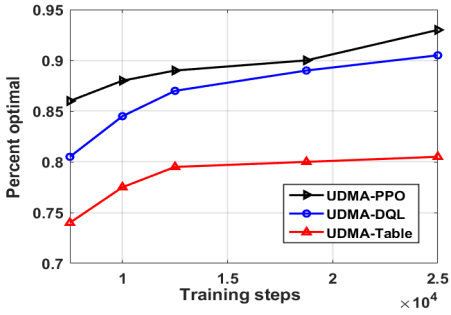

Fig. 3: Learning performance comparison between UDMA-PPO, UDMA-DQL and UDMA-Table.

by comparing the UDMA-PPO, UDMA-DQL and UDMA-Table curves. Each point in the curve is the percentage of times over 100 Monte Carlo runs that the RL algorithm found the optimum solution after learning for a number of training steps. The optimum solution was determined by performing an exhaustive search over all possible action combinations for all the CRs ($|\mathcal{A}|^N = 196$ combinations in our case) for the network setup in each Monte Carlo run. In order to have an unbiased comparison, the UDMA-DQL and UDMA-Table algorithms were realized for the same number of training steps as the UDMA-PPO. While it is known that the DQL learns faster than the table-based approach, work in [5] presented this performance advantage in a complex setting of a non-stationary environment. This paper studies the learning advantage of PPO in the same challenging setting of a distributed and uncoordinated network that results in a non-stationary environment and presents a comparative analysis with work in [5]. The Fig. 4 shows that the UDMA-PPO is capable of learning faster with fewer training steps than the baseline approaches. The UDMA-DQL succeeds in addressing the challenges of a non-stationary multi-agent radio environment through the combined effect of exploration phases and a Best Reply Process for gradual learning of the optimal solution. These additions are needed for UDMA-DQL to converge to the global best equilibrium states but comes at the cost of increased learning time (although it is still faster than UDMA-Table) [5].

Unlike UDMA-DQL, it is important to note that the UDMA-PPO addresses the challenges of a non-stationary environment solely with its policy update design, without the need for adding a specific mechanism. This leads to UDMA-PPO learning faster than UDMA-Table and UDMA-DQL. Now keeping the same settings, when allowed for an arbitrarily long learning time, the UDMA-PPO continues to achieves better results. In Fig. 4, it can be seen that for the same number of

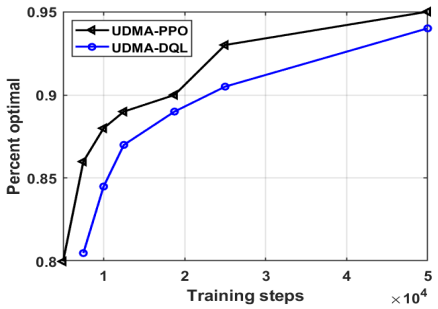

Fig. 4: Convergence to optimal policy as a function of training steps

$50K$ training steps, UDMA-PPO achieves an optimal policy selection accuracy of 95% compared to the 94% accuracy of the UDMA-DQL. That UDMA-PPO inherently can address the non-stationarity of the environment can be attributed to the policy update design. The clipping parameter ϵ in (11), clips the policy ratio to remain within the range of $[0.9, 1.1]$ and ensures that the updated policy does not deviate too far from the old policy, maintaining the exploration-exploitation trade off. Since the advantage function is closely related to the rewards (9), the clipping parameter ϵ, moderates the effect advantage values (and by extension noisy rewards resulting from non-stationary environment) and the actor loss function (11) have on the policy updates during learning. The clipping of the actor loss function effectively damps the relatively large changes in the rewards that may arise from the non-stationary nature of the environment, reducing the non-stationary nature of the environment and presenting it to the learning algorithm more as a pseudo-stationary environment. Thus clipping prevents large policy updates in response to noisy advantages and ensures that noisy rewards arising from non-stationary environment do not dominate the learning process.

V. Conclusion

The results demonstrate superior learning performance of the PPO algorithm in a distributed and uncoordinated multi-agent network which, to the best of our knowledge, is the

first work to use PPO in a non-stationary CRN environment. The multi-agent scenario is a result of the cognitive radios that share with the PN, the radio spectrum through underlay dynamic spectrum access. Each cognitive node independently learns its transmit power without violating the relative throughput change threshold at its nearest PN link while achieving the highest possible throughput. Because the independently-set transmit power from each cognitive radio translates into interference to the PN and other SN links, the key interest herein is assessing the performance of PPO in a non-stationary environment. The results demonstrate that uncoordinated and distributed multi-agent PPO (UDMA-PPO) is able to learn faster than the equivalent table-based Q-learning (UDMA-Table) and deep Q-learning (UDMA-DQL), achieving better performance with about 70% less training steps compared to UDMA-Table and similar performance with 25% less training steps than UDMA-DQL. Moreover, we observed that UDMA-PPO is inherently capable of learning in a non-stationary environment, without the need of adding improvements to the standard implementation, and we explain that this property is because the clipping that is done within the advantage function-based actor loss has a damping effect on the variability of rewards from the non-stationary environment.

REFERENCES

[1] E. Hossain, D. Niyato, and Z. Han, *Dynamic spectrum access and management in cognitive radio networks*. Cambridge university press, 2009.

[2] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE communications surveys & tutorials*, vol. 11, no. 1, pp. 116–130, 2009.

[3] M. Song, C. Xin, Y. Zhao, and X. Cheng, "Dynamic spectrum access: from cognitive radio to network radio," *IEEE Wireless Communications*, vol. 19, no. 1, pp. 23–29, 2012.

[4] F. Shah-Mohammadi and A. Kwasinski, "Fast learning cognitive radios in underlay dynamic spectrum access: Integration of transfer learning into deep reinforcement learning," in *2020 Wireless Telecommunications Symposium (WTS)*. IEEE, 2020, pp. 1–7.

[5] A. Tondwalkar and A. Kwasinski, "A Deep Q-Learning Algorithm with Guaranteed Convergence for Distributed and Uncoordinated Operation of Cognitive Radios," *IEEE Access*, vol. 13, pp. 19 678–19 693, 2025.

[6] Sutton, R and Barto A , Reinforcement learning: An introduction, MIT Press, 2018.

[7] D. C. Nguyen, D. J. Love, and C. G. Brinton, "Intelligent spectrum sensing and resource allocation in cognitive networks via deep reinforcement learning," in *ICC 2023, IEEE International Conference on Communications*. IEEE, 2023, pp. 4603–4608.

[8] N. Mishra, S. Srivastava, and S. N. Sharan, "RADDPG: Resource allocation in cognitive radio with deep reinforcement learning," in *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2021, pp. 589–595.

[9] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep Q learning approach," in *ICC 2019-2019 IEEE International Conference On Communications (ICC)*. IEEE, 2019, pp. 1–6.

[10] F. Zishen, X. Xianzhong, S. Zhaoyuan, and C. Xiping, "Proximal policy optimization based continuous intelligent power control in cognitive radio network," in *IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp. 820–824.

[11] K. Sun, J. Yang, Y. B. Li, J, and S. Ding, "Proximal policy optimization-based hierarchical decision-making mechanism for resource allocation optimization in uav networks," *Electronics*, vol. 14, no. 4, 2025.

[12] X. Lin, A. Liu, C. Han, X. Liang, Y. Sun, G. Ding, and H. Zhou, "Intelligent adaptive mimo transmission for nonstationary communication environment: A deep reinforcement learning approach," *IEEE Transactions on Communications*, 2025.

[13] A. Doshi and J. G. Andrews, "Distributed proximal policy optimization for contention-based spectrum access," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 2021, pp. 340–344.

[14] F. Shah-Mohammadi, H. H. Enaami, and A. Kwasinski, "Neural network cognitive engine for autonomous and distributed underlay dynamic spectrum access," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 719–737, 2021.

[15] J. Yang, N. Tin, and A. K. Khandani, "Adaptive modulation and coding in 3G wireless systems," in *Proceedings IEEE 56th Vehicular Technology Conference*, vol. 1. IEEE, 2002, pp. 544–548.

[16] C. Tarhini and T. Chahed, "On capacity of ofdma-based IEEE 802.16 WiMAX including adaptive modulation and coding (amc) and inter-cell interference," in *2007 15th IEEE Workshop on Local & Metropolitan Area Networks*. IEEE, 2007, pp. 139–144.

[17] M. Joseph III, "Cognitive radio: An integrated agent architecture for software defined radio," *Ph. D. Thesis, KTH Royal Inst. Technology*, 2000.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[19] 3GPP, "Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 3GPP TR 36.814 v9.0.0, 2010.

[20] 3GPP , "Physical layer procedures," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 3GPP TR 36.213 v9. 2.0, 2010.

[21] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE transactions on Communications*, vol. 47, no. 6, pp. 884–895, 1999.