Imperfect CSI Based Intelligent Dynamic Spectrum Management Using Cooperative Reinforcement Learning Framework in Cognitive Radio Networks

Amandeep Kaur, *Student Member, IEEE* and Krishan Kumar, *Member, IEEE*

Abstract—The rapid development of wireless traffic pushed the wireless community to research different solutions towards the efficient utilization of the available radio spectrum. However, a recent study shows that most of the dynamically allocated spectrum bands (radio frequency resources), experience significant underutilization as cognitive radio (CR) technology still lacks intelligence. An intelligence in CRs can be incorporated with machine learning algorithms. Further, the perfect channel state information (CSI) is hardly obtained and CSI imperfections play a crucial role in Dynamic Spectrum Management. Thus, for efficient utilization of available spectrum, a decentralized Multi-Agent Reinforcement Learning based resource allocation scheme has been proposed. A robust resource allocation scheme is proposed which integrates machine learning and CR technology into a sophisticated multi-agent system (MAS). Moreover, assisted with cloud computing which provides a huge amount of storage space, reduces operating expenditures, and provides wider flexibility of cooperation. Hence, to foster the performance of the proposed scheme, a cooperative framework in MAS is introduced which enhances the performance of the proposed scheme in terms of network capacity, outage probability, and convergence speed. Numerical results verify the effectiveness of the proposed scheme and show the non-negligible impact of imperfect CSI, thus highlighting the importance of robust designs that maintains users' QoS in practical wireless networks.

Index Terms—Cognitive radio (CR) networks, cloud computing, cooperative multi-agent system, reinforcement learning, spectrum management, imperfect CSI

1 Introduction

In Digital Era, wireless technology has become an essential part of everyday lives from online bill payments to government offices. The increasing demand for data rate and Quality of Service (QoS) for current wireless applications such as video calling, voice over internet protocol, online gaming, and video streaming creates challenges for the existing wireless community [1]. Cisco Virtual Networking Index forecasts that the mobile data traffic will grow with the Compound Annual Growth Rate (CAGR) of 46 percent, reaching 77.5 Exabytes per month by 2022 [2]. According to Federal Communications Commission report [3], most of the allocated spectrum bands are underutilized and there exist spectrum holes called white spaces. Cognitive Radio technology is envisioned as a potential solution to mitigate the problem of the underutilized spectrum by accessing spectrum bands which are not utilized by Primary Users (PUs) without interference [4], [5]. Generally, CR technology is widely adopted as *intelligent technology* that utilizes its cognition ability to autonomously adapt its transmission parameters according to the radio environment. Even so, there exist some problems in CR networks such as high computational time and complexity [6]. To reduce these problems, CRs have to adopt intelligent learning with autonomous decision-making capabilities by interacting with the environment. Such type of intelligence is provided by machine learning techniques which paved a path for intelligent spectrum management [7].

Up to now, most of the research problems in CR networks have applied machine learning techniques to solve various issues such as dynamic channel selection [8], routing, and spectrum sensing [9]. One of the favorable state-of-the-art machine learning techniques is Reinforcement Learning (RL) that aimed to build up the solution to the problem on a trial-and-error basis [10]. The most widely adopted RL algorithm in CR networks is Q-Learning (Q-L) [11]. Most of the work existing in literature focuses on Q-L and its variants such as Transfer Actor-Citric Learning [12] and distributed inter-cell interference coordination accelerated Q-L [11]. Although RL provides favorable outcomes in CR networks but it depends on network characteristics such as single-agent, multi-agent, centralized, decentralized, cooperative, and non-cooperative. Nguyen *et al*. focused on energy-efficient power allocation problems in a multi-agent environment with a non-cooperative framework that shows the selfish dynamics of CRs [13]. However, authors in [14] proposed RL based cooperative resource allocation scheme in which multi-agents have been feeding their learned Q-values in common Q-table in a centralized manner which improves the performance in

• *The authors are with the Department of Electronics and Communication Engineering, National Institute of Technology, Hamirpur, Himachal Pradesh 177005, India. E-mail: adeep5524@yahoo.com, krishan_rathod@nith.ac.in.*

Manuscript received 25 Apr. 2020; revised 19 Aug. 2020; accepted 21 Sept. 2020. Date of publication 24 Sept. 2020; date of current version 4 Apr. 2022. (Corresponding author: Amandeep Kaur.)
Digital Object Identifier no. 10.1109/TMC.2020.3026415

1536-1233 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

terms of capacity and energy efficiency. But it also increases signaling overhead and memory requirements. Using the heuristic idea of inexpensive cloud storage reduces the operational expenditures and provides a large storage space to save the experience earned that is previously ignored or trashed [15]. Due to the conflicts among cooperative and non-cooperative frameworks, it is more suitable to address the problem of resource allocation in a decentralized cooperative fashion which enhances the performance with cooperative communication with reduced signaling overhead.

Most of the existing works consider resource allocation problems with the assumption that complete instantaneous Channel State Information (CSI) is available which is impossible to obtain in practical systems [16], [17]. To cope up with the unknown environment, a method was proposed by modeling CSI imperfections as Gaussian error [18] to analyze the impact of imperfect CSI on resource allocation in CR networks. Further, authors in [19] modeled CSI imperfections using ellipsoidal approximations for joint resource allocation in multi-relay Orthogonal Frequency Division Multiple Access (OFDMA) networks. The joint optimal chunk assignment, transmission link selection and power allocation problem forms mixed integer programming problem, which was solved with dual maximization based joint resource allocation scheme and greedy based resource allocation scheme under imperfect CSI. In greedy based resource allocation scheme, power is allocated to chunks using water-filling method with low complexity. Using an idea of dynamic programming, a joint resource scheduling problem was investigated for deadline constrained transmission with arithmetic mean estimation and geometric mean estimation [17]. In [20], the authors proposed a Lagrange dual decomposition approach to solve different resource allocation and optimization problems to facilitate proper treatment to imperfect CSI. Moreover, an online learning based transmission optimization problem with channel estimation errors was investigated in [16] for delay-sensitive data. Inspired by the Q-L algorithm, which uses ε-greedy action selection strategy, two computationally efficient scheduling algorithms with and without uncertainty bound of CSI imperfections are proposed. The proposed algorithms can tackle the imperfect CSI issue and improve system performance in terms of energy efficiency, transmission delay, and overflow probability. However, with the increase in the number of power levels, the algorithm leads to a trap in the curse of dimensionality which increases the required computations and thus leads to slower convergence.

To best of our knowledge, there is no existing solution for resource allocation assisted by cloud computing with the cooperative framework in CR networks using machine learning techniques. This motivates the exploration of the novel Cloud Assisted Multi-Agent RL (CA-MARL) scheme to solve the problem of resource allocation with imperfect CSI consideration, aiming at deciding how the appropriate resources are allocated among CRs with the cooperative framework. The proposed cooperative scheme involves multi-BS cooperation and analyzes local information of the environment that can improve the performance by decreasing interference and congestion problems significantly. Moreover, cloud storage to save the experience earned in the form of Q-values is employed to handle the curse of dimensionality during the learning process. RL is the most prominent machine learning technique as it has model-free schemes, which facilitate its usage in dynamic and complex networks. The key contributions of this paper are as follows:

- The work proposed and investigates imperfect CSI based decentralized resource allocation scheme which can deal with channel estimation errors
- The robust resource allocation scheme is proposed with the consideration of imperfect CSI that aims to maximize network capacity. The CSI imperfections are modelled using ellipsoidal approximation. To achieve a robust solution, the idea of multi-agent reinforcement learning is adopted to solve the resource allocation problem
- Novel cloud-assisted cooperative multi-agent reinforcement learning-based scheme is proposed that allows each agent to update its learning information under complete information about other agents' past strategies using historical information available in the cloud without increasing cooperation overhead
- The real-time environment is set up with Software Defined Radio (SDR) based hardware testbed to measure effective channel gain of all links for each resource which is utilized for robust resource allocation with cloud-assisted multi-agent reinforcement learning

The rest of the paper is organized as follows. Section 2 presents the system model and resource allocation problem formulation in CR networks. The problem of resource allocation under MARL based Q-L scheme is presented in Section 3. In Section 4, cloud-assisted cooperative multi-agent reinforcement learning-based scheme for resource allocation with its convergence is proposed. Testbed implementation setup along with the achieved experimental results and simulated numerical results are presented in Section 5 and finally, the paper concludes in Section 6.

2 System Model and Problem Formulation

The considered system model consists of the primary network and the CR network components. The primary network consists of Primary Users (PUs) who have a license to operate in a certain spectrum band. The operations of the PUs are controlled through Primary Base Stations (PBS). Due to their priority of PUs, the presence of unlicensed users should not affect the performance of PUs. The CR network have CRs associated with Cognitive Base Stations (CBSs) which share the same licensed spectrum band with PUs. The architecture of considered cloud-assisted CR networks is illustrated in Fig. 1. It consists of four main components:

- A cloud with high storage capacities
- Base Stations (PBSs and CBSs) with wireless access functions
- Co-existing PUs and CRs (as secondary users)
- Backhaul links to deliver the learned experience by interacting with the environment from BS to cloud for Multi-BS cooperation

The architecture of the considered cloud-assisted CR networks consists of Primary Base Stations (PBSs) which are underlaid with Cognitive Base Stations (CBSs). Let $\mathbf{U} =$

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

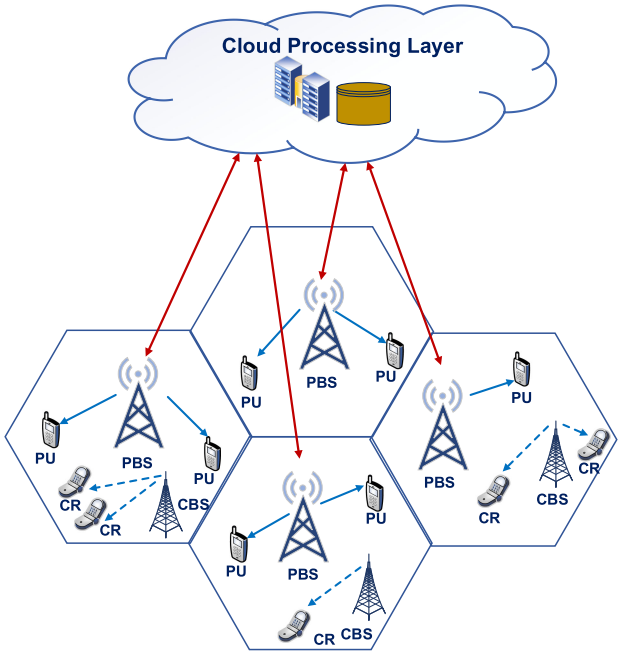

Fig. 1. Cloud assisted CR networks architecture.

$\{1, 2, \ldots, U\}$ denote the set of indices of PBSs, $\mathbf{B} = \{1, 2, \ldots, B\}$ denote the set of indices of CBSs, and $\mathbf{N} = \{1, 2, \ldots, N\}$ denote the set of orthogonal resources with total bandwidth W. The resource allocation process for CBSs is achieved through the cooperation of PBSs in a decentralized way. There are two types of users present in the network: PUs and CRs. The PUs are denoted with indices $\mathbf{M} = \{1, 2, \ldots, M\}$ and CRs are denoted with indices $\mathbf{K} = \{1, 2, \ldots, K\}$ that are associated with PBSs and CBSs respectively. The PBSs and CBSs are linked with the cloud to facilitate data processing and cloud computing. Let $a^n_{b,k} \in \{0, 1\}$ represents resource allocation with binary indicator. The value of $a^n_{b,k} = 1$ if the resource n is allocated to k^{th} CR associated with b^{th} CBS. To allocate the resources efficiently, the instantaneous CSI of each resource should be evaluated and send back to PBS. The CSI of each resource is represented by effective channel gain. However, it is impossible to obtain full feedback information on effective channel gain [17] and only the estimated value of effective channel gain can be obtained.

Let $g^n_{u,m}, g^n_{b,k}(h^n_{u,m}, h^n_{b,k})$ represents estimated (actual) effective gains for m^{th} PU and u^{th} PBS and over the link between k^{th} CR and b^{th} CBS on the resource n. The CSI imperfections can be deterministically modeled using ellipsoidal approximations as follows

$$g^n_{u,m} = h^n_{u,m} \times 10^{\frac{-\upsilon^n_{u,m}}{10}} : \left|\upsilon^n_{u,m}\right| \leq \theta, \forall u \in U, \forall n \in N, \forall m \in M \tag{1}$$

$$g^n_{b,k} = h^n_{b,k} \times 10^{\frac{-\upsilon^n_{b,k}}{10}} : \left|\upsilon^n_{b,k}\right| \leq \theta, \forall b \in B, \forall k \in K, \forall n \in N, \tag{2}$$

where $\upsilon^n_{u,m}, \upsilon^n_{b,k}$ represents channel estimation error vector defined as

$$\upsilon^n_{u,m} = h^n_{u,m} - g^n_{u,m}$$
$$\upsilon^n_{b,k} = h^n_{b,k} - g^n_{b,k},$$

where $\upsilon^n_{u,m}, \upsilon^n_{b,k}$ is assumed to be bounded in an ellipsoidal uncertainty region, θ.

The Signal to Interference and Noise Ratio (SINR) for k^{th} CR occupying resource n is given as

$$\xi^n_{b,k} = \frac{g^n_{b,k} P^n_{b,k}}{\sigma^2 + I^n_k}, \tag{3}$$

where $P^n_{b,k}$ is the transmission power of b^{th} CBS allocated to k^{th} CR on the resource n, σ^2 is the noise power, and $g^n_{b,k}$ represents the channel gain between k^{th} CR and b^{th} CBS on the resource n. The term I^n_k represents encountered interference calculated as

$$I^n_k = \underbrace{\sum_{m=1}^{M} P^n_{u,m} h^n_{u,m}}_{\text{Interfernce due to PUs}} + \underbrace{\sum_{i=1, i\neq k}^{K} \sum_{l \in \mathbf{B}/\{b\}} P^n_{l,i} h^n_{l,i}}_{\substack{\text{Interfernce due to other CRs} \\ \text{served by other CR}-\text{BSs}}},$$

where $P^n_{u,m}$ is the transmission power of u^{th} PBS allocated to m^{th} PU on resource n, $g^n_{u,m}$ denotes channel gain between u^{th} PBS allocated to m^{th} PU on resource n. Note that the index i represents other CRs served by other CBSs.

Similarly, the SINR achieved by m^{th} PU associated with u^{th} PBS utilizing resource n is found as follows,

$$\xi^n_{u,m} = \frac{g^n_{u,m} P^n_{u,m}}{\sigma^2 + I^n_m}. \tag{4}$$

The term I^n_m represents encountered interference evaluated as

$$I^n_m = \sum_{i=1}^{K} \left(\sum_{b \in \mathbf{B}} P^n_{b,i} h^n_{b,i} \right).$$

Based on CSI feedback as effective channel gain in (1) and (2), (3) and (4) can be rewritten as

$$\xi^n_{b,k} = \frac{\left(h^n_{b,k} 10^{\frac{-\upsilon^n_{b,k}}{10}} \right) P^n_{b,k}}{\sigma^2 + I^n_k} \tag{5}$$

$$\xi^n_{u,m} = \frac{\left(h^n_{u,m} 10^{\frac{-\upsilon^n_{m,u}}{10}} \right) P^n_{u,m}}{\sigma^2 + I^n_m}. \tag{6}$$

The capacity for all CRs associated with b^{th} CBS is expressed as

$$C^n_{b,k} = \sum_{k=1}^{K} \sum_{n=1}^{N} a^n_{b,k} W \log_2(1 + \xi^n_{b,k}). \tag{7}$$

The capacity for all PUs associated with u^{th} PBS is expressed as

$$C^n_{m,u} = \sum_{m=1}^{M} \sum_{n=1}^{N} a^n_{m,u} W \log_2(1 + \xi^n_{m,u}), \tag{8}$$

where W is system bandwidth and $a^n_{m,u}$ represents binary indicator similar to $a^n_{b,k}$. The total achievable network capacity in time slot t is given as

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

$$C_{NW} = C^n_{b,k} + C^m_{m,u}, \forall k \in \mathbf{K}, \forall m \in \mathbf{M}. \quad (9)$$

The allocation of resources and transmission power with the objective of capacity maximization subjected to maintain QoS of PUs and CRs under following constraints is formulated as follows

$$\max_{a^n_{m,u}, a^n_{b,k}, p^n_{m,u}, p^n_{b,k}} C_{NW} \quad \text{s.t.} \quad (10)$$

$$\text{C1}: a^n_{b,k} \in \{0,1\}, \forall n \in \mathbf{N}, \forall b \in \mathbf{B}, \forall k \in \mathbf{K}$$
$$\text{C2}: a^n_{m,u} \in \{0,1\}, \forall m \in \mathbf{M}, \forall u \in \mathbf{U}, \forall n \in \mathbf{N}$$
$$\text{C3}: C^n_{b,k} = \sum_{n=1}^{N} a^n_{b,k} W \log_2(1 + \xi^n_{b,k}) \geq C^*_k, \forall k \in \mathbf{K}$$
$$\text{C4}: C^n_{m,u} = \sum_{n=1}^{N} a^n_{m,u} W \log_2(1 + \xi^n_{m,u}) \geq C^*_m, \forall m \in \mathbf{M}$$
$$\text{C5}: P^n_{b,k} \leq a^n_{b,k} p_{b,\max}, \forall b \in \mathbf{B}, \forall k \in \mathbf{K}, \forall n \in \mathbf{N}$$
$$\text{C6}: P^n_{u,m} \leq a^n_{m,u} p_{u,\max}, \forall u \in \mathbf{U}, \forall m \in \mathbf{M}, \forall n \in \mathbf{N}$$
$$\text{C7}: \sum_{k=1}^{K} \sum_{n=1}^{N} a^n_{b,k} P^n_{b,k} I^n_m \leq I^{th}_m, \forall m \in \mathbf{M}$$
$$\text{C8}: \left|\upsilon^n_{u,m}\right| \leq \theta, \forall u \in \mathbf{U}, \forall m \in \mathbf{M}, \forall n \in \mathbf{N}$$
$$\text{C9}: \left|\upsilon^n_{b,k}\right| \leq \theta, \forall b \in \mathbf{B}, \forall k \in \mathbf{K}, \forall n \in \mathbf{N}.$$

The constraint C1 and C2 represent a binary indicator for resource allocation. The constraints C3 and C4 are capacity constraint which ensures QoS for CR and PU is above the thresholds C^*_k and C^*_m respectively. C5 and C6 indicate the maximum allowed power for each CBS and PBS during the allocation of resources with $a^n_{b,k} = 1$ and $a^n_{m,u} = 1$ respectively. C7 represents a constraint on interference to PU where I^{th}_m is interference threshold of the m^{th} PU. Finally, the constraint C8 and C9 represent bounded uncertain parameters for PUs and CRs respectively. The problem of resource allocation in (10) is tackle with MARL based Q-L scheme to reach the ultimate goal of achieving maximum network capacity while maintaining QoS of PUs and CRs.

3 Resource Allocation using MARL Framework

The goal of this section is to tackle the optimization problem formulated in Section 2 using the cooperative MARL framework. The resource allocation problem is mapped into RL components that are state, action, and reward. Moreover, it becomes challenging to determine exact state transition probability in model-based learning schemes due to the complex network dynamics. It is not trivial to form a state transition model beforehand. For this reason, RL becomes suitable for dynamic spectrum management in cooperative CR networks [21]. Here, each PBS plays a role of learning agent, which interacts and observe network state $s(t)$ and its associated actions $a(t)$ during the time slot t. In the next time slot, the reward is generated based on the action performed and state transition occurs from the current state to the next state. The state, action, reward and transition functions are defined as follows

- *State:* The environment state at a certain time slot t is defined as

$$s(t) = (k, m, I^n_{b,k}, I^n_{u,m}) \quad (11)$$

The information is acquired from the cloud and PBSs. It is assumed that each PBS is aware of CBSs operating in its coverage area. A state is represented as four tuple information which includes the number of PUs and CRs present in the network and estimated CSI as SINR represented with binary indicators as $I^n_{b,k} \in \{0,1\}$ and $I^n_{m,u} \in \{0,1\}$. Each binary indicator represents 1 when received SINR is greater than or equal to the SINR threshold and 0 elsewhere.

- *Action:* The action $a_u(t) = (a^n_{b,k}, p_{u_{b,k}})$ is defined as the allocation of available resources with transmission power, CBS allocated to its associated CRs. To deal with channel estimation errors, the transmission power can be determined as

$$p_{u_{b,k}} = \min\left\{\frac{\xi^n_{b,k} \times \left(\sigma^2 + I^n_k\right)}{\left(h^n_{b,k} 10^{\frac{-\upsilon^n_{b,k}}{10}}\right)}, p_{b,max}\right\}, \quad (12)$$

where $p_{b,max}$ represents the maximum transmission power of CBS.

- *Reward:* The reward function is defined as network capacity obtained by performing action $a(t)$ in a state $s(t)$ defined as follows

$$R(s,a) = C_{NW}(s,a), \quad (13)$$

where $C_{NW}(s,a)$ is achievable network capacity defined in (10) and the reward function is obtained if C1 to C9 are satisfied.

- *Transition Function:* The transition function $T(s,a,s')$ is evaluated under the policy $\pi \in \Pi$ while moving from current state $s(t)$ to next state $s(t+1)$ as a result of resource allocation as action by PBS at time slot t is given as follows

$$T(s,a,s') = \Pr(s(t+1) = s' | s(t) = s, a(t) = a). \quad (14)$$

If each PBS $u \in \mathbf{U}$ learns independently and acts selfishly to select a strategy $\pi_u(s_u)$ to maximize its discounted reward function, it creates certain interference and congestion problems due to lack of information of other PBSs strategies. Thus, it is necessary to include cooperation among PBSs to select the appropriate strategy and to avoid interference and congestion problems. Therefore, cloud-assisted multi-agent RL with a cooperative framework provides a considerable solution. The strategies of all PBSs taken in $(t-1)$ time slots are stored as historical information in the cloud. Let $u \in \mathbf{U}$ denotes a set of PBSs participating for cooperation. For this case, each PBS has complete information about other PBS's strategies as,

$$\pi_{-u} = (\pi_1, \ldots . \pi_{u-1}, \pi_{u+1}, \ldots . \pi_U), \forall u \in \mathbf{U}.$$

Thus, the total expected discounted reward obtained under complete knowledge of other PBSs strategies is defined

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

as follows

$$\max_{\pi_u \in \Pi_u} \left\{ E\left[\sum_{t=0}^{\infty} \gamma^t C_u(s_u^t, \pi_u(s_u^t), \pi_{-u}(s_u^t)) \right] \right\}, \tag{15}$$

where Π_u is a set of available strategies to PBS u and γ represents discount factor. The strategy $\pi_u(s_u, a_u)$ represents the probability of PBS u for selecting an action a_u in a particular state s_u. The term $\pi_u(s_u)$ represents the strategy taken by PBS u to select an action at state s_u. Another term $\pi_{-u}(s_u)$ with $-u$ index represents the strategy taken by other PBSs for the state s_u. Thus, the total discounted reward of PBS u under complete information of other PBSs strategies π_{-u} is given as

$$\begin{aligned} V_u(s_u, \pi_u, \pi_{-u}) &= E[C_u(s_u, \pi_u(s_u), \pi_{-u}(s_u))] \\ &+ \gamma \sum_{s'_u \in S_u} T(s_u, a, s'_u) \pi_u(s_u), \pi_{-u}(s_u) \\ &\times V_u(s'_u, \pi_u, \pi_{-u}), \end{aligned} \tag{16}$$

where $E(.)$ represents expectation operator under the joint distribution of a_{-u} given by π_{-u} in particular state s defined as

$$\begin{aligned} &E[C_u(s_u, \pi_u(s_u), \pi_{-u}(s_u))] \\ &= \sum_{(a_u, a_{-u}) \in A} C_u(s_u, a_u, a_{-u}) \prod_{j \in \mathbf{U}/\{u\}} \pi_j(s_j, a_j). \end{aligned}$$

The term s_u represents the state at a time slot t and s'_u represents the state at a time slot $(t+1)$. Each PBS aims to achieve the best strategy π_u^* for each environment state that satisfies Bellman's optimality equation [22]. Thus, the total discounted reward defined in (16) under best strategy can be rewritten as

$$\begin{aligned} V_u(s_u, \pi_u^*, \pi_{-u}^*) &= E\big[C_u(s_u, a_u, \pi_{-u}^*(s_u))\big] \\ &+ \gamma \sum_{s'_u \in S_u} T(a_u, \pi_{-u}^*(s_u)) \\ &\times V_u(s'_u, \pi_{-u}^*, \pi_{-u}^*), \end{aligned} \tag{17}$$

where

$$\begin{aligned} &E\big[C_u(s_u, a_u, \pi_{-u}^*(s_u))\big] \\ &= \sum_{(a_u, a_{-u}) \in A} C_u(s_u, a_u, a_{-u}) \prod_{j \in \mathbf{U}/\{u\}} \pi_j^*(s_j, a_j). \end{aligned}$$

Further, the optimal Q-value of PBS u reached in a recursive way defined as the sum of the current reward and future expected rewards given as

$$\begin{aligned} Q_u^*(s_u, a_u) &= E\big[C_u(s_u, a_u, \pi_{-u}^*(s_u))\big] \\ &+ \gamma \sum_{s'_u \in S_u} T_{s_u, s'_u}(a_u, \pi_{-u}(s_u)) V_u(s'_u, \pi_u^*, \pi_{-u}^*) \\ &= E\big[C_u(s_u, a_u, \pi_{-u}^*(s_u))\big] \\ &+ \gamma \sum_{s'_u \in S_u} T_{s_u, s'_u}(a_u, \pi_{-u}(s_u)) \max_{\tilde{a}_u \in A_u} Q_u^*(s'_u, \tilde{a}_u). \end{aligned} \tag{18}$$

The optimal Q-value $Q^*(s, a)$ can be found recursively with the information tuple $\langle s, a, r, s' \rangle$. Thus, the general update rule for Q-L is expressed as

$$\begin{aligned} Q_u^{t+1}(s_u, a_u) &= (1 - \alpha^t) Q_u^t(s_u, a_u) \\ &+ \alpha^t \left\{ \begin{array}{l} \sum_{a_{-u} \in A_{-u}} \left[C_u(s_u, a_u, a_{-u}) \prod_{j \in \mathbf{U}/\{u\}} \pi_j^t(s_j, a_j) \right] \\ + \gamma \max_{\tilde{a}_u \in A_u} Q_u^t(s'_u, \tilde{a}_u) \end{array} \right\}. \end{aligned} \tag{19}$$

where $\alpha \in [0, 1)$ is the learning rate which determines how often the learning agent updates its Q-value. The value $\alpha = 0$ represents no learning whereas $\alpha = 1$ represents new knowledge completely replaces the old one.

The trade-off between exploration and exploitation is also an important issue during the learning process. In exploration, it tries to explore new policies and does not apply previously known policies while exploitation seizes already explored policies. To balance the trade-off between exploration and exploitation, actions are selected using ε-greedy selection [23], which chooses equally among available actions during exploration. One drawback of this approach is that the probability of selecting the worst action is equal to the probability of selecting the best one. Another approach is to use Boltzmann distribution in which action probability is varied as a graded function of Q-value and hence select the best strategy with high probability. At time slot t, each PBS u selects an action a_u in the state s_u with probability

$$\pi_u^t(s_u, a_u) = \frac{e^{Q_u^t(s_u, a_u)/\Gamma}}{\sum_{\tilde{a}_u \in A_u} e^{Q_u^t(s_u, \tilde{a}_u)/\Gamma}}, \tag{20}$$

where Γ represents temperature. The high value Γ leads to the selection of action with equal probability while lower value leads to induce a large difference in their Q-value.

4 Proposed Cloud Assisted Cooperative Multi-Agent Q-Learning Paradigm

As discussed in Section 3, it is noticed that the network capacity obtained as a reward function depends on the strategies of other BSs. However, with the inclusion of cooperation, the exchange of information about learned strategies between participating PBSs increases computational complexity as well as overhead. Therefore, it is necessary to include cooperation while taking care of increasing complexity and overhead. Thus, the CA-MARL mechanism provides strategies of other PBSs as historical information. Hence, it exploits the historical information to obtain the best strategy without increasing computational complexity as well as overhead.

Let $\tau^t(s_u, a_{-u}) = \prod_{j \in \mathbf{U}/\{u\}} \pi_j(s_j, a_j)$ represents the collection of strategies of all other PBSs in a time slot t which is used to determine $Q_u^{t+1}(s_u, a_u)$ in next time slot. It forms historical information of all PBSs which selects action a_{-u} according to $\pi_{-u}(s_u)$ in time slot $(t-1)$. Each PBS selects its strategy based on its previous strategy $\pi_u^{t-1}(s_u, a_u)$ as well as about other PBSs strategies available at the time slot t in the form of historical information. This means that if PBS u changes its strategy, then it will induce changes in the strategy of other PBSs in the next time slot. As a result, every PBS experiences impact of historical information available at the time slot t. Therefore, the impact factor can be

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

estimated from local historical information available in the time slot t which is defined as follows

$$\tau_u^t(s_u, a_{-u}) = \tau_u^{t-1}(s_u, a_{-u}) + \mu_u\left[\pi_u^t(s_u, a_u) - \pi_u^{t-1}(s_u, a_u)\right], \tag{21}$$

where μ_u represents a positive scalar. Therefore, the update rule in (19) is modified in a way that each PBS updates its Q-value under complete information of other PBSs strategies with Q-L scheme such that

$$Q_u^{t+1}(s_u, a_u) = (1-\alpha^t)Q_u^t(s_u, a_u) + \alpha^t\left\{\begin{array}{l}\sum_{a_{-u}\in A_{-u}}\left[\tau_u^t(s_u, a_{-u}) \times C_u(s_u, a_u, a_{-u})\right] \\ +\gamma \max_{\tilde{a}_u \in A_u} Q_u^t(s'_u, \tilde{a}_u)\end{array}\right\}, \tag{22}$$

For each network state $s \in S$, the action taken by PBSs independently in a decentralized way. To proceed with multi-BSs learning, it is assumed that the policy of different PBSs does not alter significantly under similar environment states. According to this assumption, each PBS is aware of the spectrum band allocation policy of other PBSs without increasing cooperation overhead through the use of historical information available in the cloud, if it encounters the same environment. The decentralized CA-MARL based resource allocation scheme with a cooperative framework in CR networks is illustrated in Algorithm 1. The algorithm is repeated until converges to evaluate the best resource allocation strategy.

Algorithm 1. CA-MARL: **C**loud **A**ssisted cooperative **M**ulti-**A**gent **R**einforcement **L**earning Algorithm

1: **Initialize:**
2: Let $t = 0$
3: **for** all $s \in S_u, a \in A_u$ **do**
Initialize spectrum band allocation strategy,$\pi^t(s,a)$, $Q^t(s,a)$ value,$\tau_u^t(s_u, a_{-u})$ and $\mu_u \succ 0$
4: **end for**
5: Evaluate state at time step $t, s = s(t)$
6: **while (True) do**
Select an action according to $\pi^t(s,a)$ in (20)
if ($C.1$ to $C.9$ are satisfied) **then**
$R(s,a) = C_{NW}(s,a)$ is achieved
else
$R_u(s,a) = 0$
end
7: Update $Q^{t+1}(s,a)$ according to (22)
8: Update $\pi^{t+1}(s,a)$ according to (20)
9: Update $\tau_u^t(s_u, a_{-u})$ according to (21)
10: Set $t = t+1$ and next state become current state as $s_t = s_{t+1}$
11: **end while**

4.1 Convergence of Proposed Scheme

The proposed CA-MARL based resource allocation scheme converges if conditions of general Q-L process are satisfied. The proof relies on the following assumptions.

Assumption 1. *Every pair $(s,a) \in S \times A$ often visit infinitely as current state and action.*

Assumption 2. *The learning rate, α decreases over time such that $\sum_{\alpha=0}^{t} \alpha_t = \infty$ and $\sum_{\alpha=0}^{t} (\alpha_t)^2 \prec \infty$.*

The convergence proof is based on the following *Lemma* proved by Szepesvari and Littman [24]. Let H represents pseudo-contraction operator and Θ be the space of all Q-values such that mapping occurs on complete metric space $\Theta \to \Theta$ given as

$$HQ_u^t(s_u, a_u) = \sum_{a_{-u}\in A_{-u}}\left[\tau_u^t(s_u, a_{-u}) \times C_u(s_u, a_u, a_{-u})\right] + \gamma \max_{\tilde{a}_u \in A_u} Q_u^t(s'_u, \tilde{a}_u). \tag{23}$$

Lemma. *Under Assumptions 1 and 2, for the mapping $\mathrm{H}: \Theta \to \Theta$, there exist $0 \prec \delta \prec 1$ and $\zeta^t \geq 0$ converging to zero with probability (w.p.) 1 such that $||HQ^t - HQ*|| \leq \delta||Q^t - Q*|| + \zeta^t$ for all $Q^t \in \Theta$ and $Q* = E[HQ*]$, then the iterative process defined as $Q^{t+1} = (1-\alpha^t)Q^t + \alpha^t(HQ^t)$ converges to $Q*$ w.p. 1.*

Proof.

$$\begin{aligned}
&\|HQ - HQ'\| = \max_{u\in\mathbf{U}} \max_{s_u\in S_u} \max_{a_u\in A_u} |HQ_u(s_u,a_u) - HQ'(s_u,a_u)| \\
&= \max_{u\in\mathbf{U}} \max_{s_u\in S_u} \max_{a_u\in A_u} \left|\begin{array}{l}\sum_{a_{-u}\in A_{-u}}\left[\left\{\tau_u(s_u,a_{-u}) - \tau'_u(s_u,a_{-u})\right\} \times C_u(s_u,a_u,a_{-u})\right] \\ +\gamma[\max_{\tilde{a}_u\in A_u} Q_u(s'_u,\tilde{a}_u) - \max_{\tilde{a}_u\in A_u} Q'_u(s'_u,\tilde{a}_u)]\end{array}\right| \\
&\leq \max_{u\in\mathbf{U}} \max_{s_u\in S_u} \max_{a_u\in A_u} \left|\sum_{a_{-u}\in A_{-u}}\left[\left\{\tau_u(s_u,a_{-u}) - \tau'_u(s_u,a_{-u})\right\} \times C_u(s_u,a_u,a_{-u})\right]\right| \\
&+ \max_{u\in\mathbf{U}} \max_{\tilde{a}_u\in A_u} \gamma\left|Q_u(s'_u,\tilde{a}_u) - Q'_u(s'_u,\tilde{a}_u)\right| \\
&\leq \max_{u\in\mathbf{U}} \max_{s_u\in S_u} \max_{a_u\in A_u} \left|\sum_{a_{-u}\in A_{-u}}\left[\left\{\tau_u(s_u,a_{-u}) - \tau'_u(s_u,a_{-u})\right\} \times C_u(s_u,a_u,a_{-u})\right]\right| \\
&+ \gamma\|Q - Q'\|.
\end{aligned}$$

□

Consider the first term of the above equation

$$\sum_{a_{-u}\in A_{-u}}\left[\left\{\tau_u(s_u,a_{-u}) - \tau'_u(s_u,a_{-u})\right\} \times C_u(s_u,a_u,a_{-u})\right],$$

which is calculated from (20). Thus,

$$\pi_u^t(s_u,a_u) = \frac{e^{Q_u^t(s_u,a_u)/\Gamma}}{\sum_{\tilde{a}_u\in A_u} e^{Q_u^t(s_u,\tilde{a}_u)/\Gamma}},$$

When Γ is sufficiently large, then

$$e^{Q_u(s_u,a_u)/\Gamma} = 1 + \frac{Q_u(s_u,a_u)}{\Gamma} + \Omega\left(\frac{Q_u(s_u,a_u)}{\Gamma}\right),$$

where Ω is a polynomial of order $O((\frac{Q_u(s_u,a_u)}{\Gamma})^2)$.

Thus

$$\sum_{\tilde{a}\in A_u} e^{Q_u(s_u,\tilde{a})/\Gamma} = a_u + 1 + \sum_{\tilde{a}\in A_u}\left[\frac{Q_u(s_u,\tilde{a})}{\Gamma} + \Omega\left(\frac{Q_u(s_u,\tilde{a})}{\Gamma}\right)\right].$$

It can be proved that

$$\pi_u(s_u,a_u) = \frac{1}{a_u+1} + \frac{1}{a_u+1}.\frac{Q_u(s_u,a_u)}{\Gamma} + \rho\left(\frac{Q_u(s_u,\tilde{a})}{\Gamma}\right)_{\tilde{a}}, \tag{24}$$

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

where $\rho((\frac{Q_u(s_u,\tilde{a})}{\Gamma})_{\tilde{a}})$ is polynomial of smaller order than $O((\frac{Q_u(s_u,a_u)}{\Gamma})_{\tilde{a}})$.

Similarly,

$$\pi'_u(s_u,a_u)=\frac{1}{a_u+1}+\frac{1}{a_u+1}.\frac{Q'_u(s_u,a_u)}{\Gamma}+\rho\left(\frac{Q'_u(s_u,\tilde{a})}{\Gamma}\right)_{\tilde{a}}. \tag{25}$$

From (24), (25) and with the assumption of large value of Γ

$$\left|\sum_{a_{-u}\in A_{-u}}\left[\left\{\tau_u(s_u,a_{-u})-\tau'_u(s_u,a_{-u})\right\}\times C_u(s_u,a_u,a_{-u})\right]\right| \leq \frac{1-\gamma}{a_u+1}\left|Q_u(s_u,a_u)-Q'_u(s_u,a_u)\right|.$$

This implies

$$\begin{aligned}\|HQ-HQ'\| &\leq \max_{u\in\mathbf{U}}\max_{s_u\in S_u}\max_{a_u\in A_u}\frac{1-\gamma}{a_u+1}\left|Q_u(s_u,a_u)-Q'_u(s_u,a_u)\right| \\ &+\gamma\|Q-Q'\| \\ &\leq \frac{1-\gamma}{a+1}\|Q-Q'\|+\gamma\|Q-Q'\|=\frac{\gamma a+1}{a+1}\|Q-Q'\|,\end{aligned}$$

where $a=\min_{u\in\mathbf{U}}a_u$ thus,

$$\frac{\gamma a+1}{a+1}\prec 1,$$

which concludes the convergence of the proposed scheme.

4.2 Complexity Analysis of Proposed Scheme

The main objective of the proposed algorithm is to reduce the cooperation overhead in a decentralized environment. However, this is achieved by adding some extra cost, in terms of cloud space, to the learning process itself. The complexity analysis of the proposed decentralized CA-MARL scheme is presented as:

- *Space.* As the proposed CA-MARL scheme is decentralized, each PBS require $O|\mathbb{S}||\mathbb{A}|$ to store Q-table in order to obtain optimal Q-value. The Q-table obtained by each PBS is stored in the cloud. Thus, the total space required in the cloud is $O|\mathbb{S}||\mathbb{A}||\mathbf{U}|$. However, the state considers binary indicators to represent the received SINR is above or below the threshold. It reduces the infinite increase in the number of states. The proposed scheme will converge if each state-action pair is often visited as its current state and action.
- *Cooperation Overhead.* The cooperation overhead of the proposed scheme is measured in terms of the number of policies exchanged between learning agents to obtain an optimal policy. Each learning agent store it's learned optimal policies in the form of Q-table in the cloud which acts as historical information for other agents in the next time slots. Thus, there is no direct exchange of policies between learning agents in a decentralized environment. In the case, when all the U PBSs obtain their optimal policies in $(t-1)$ time slots then the total available policies in the cloud for cooperation in the time slot t are calculated as $U\times\mathbb{S}\times\mathbb{A}$. Thus, it is very important to have a compact state in the proposed scheme to reduce time and space complexity during cooperation to expedite the convergence.

From a practical point of view, the proposed scheme deployed in a decentralized environment in which learning agents are interacting with the environment to obtain optimal policies. These learned policies are stored in the cloud in the form of Q-tables, executed as historical information, which will improve the execution time by minimizing the communication cost and removing the latency of taking actions in a particular state.

5 Performance Analysis and Evaluation

5.1 Testbed Implementation

In this section, the performance of the proposed decentralized CA-MARL based resource allocation scheme has been demonstrated. The real-time evaluation environment is set up with Universal Software Radio Peripherals (USRPs) which consist of National Instruments (NI) hardware and software. The USRP-2954R is utilized as a hardware platform that can operate in the frequency range of 10 MHz to 6 GHz. Each USRP-2954R is equipped with a Xilinx Kintex-7 FPGA, dual high-speed Digital-to-Analog Converter, and Analog-to-digital Converter. Each USRPs consist of two RF transceivers namely RF0 and RF1 which are programmable with LabVIEW Communication System Design Suite 2.0 (CSDS) software [25]. The network includes PBS with CBS under its coverage and PUs and CRs associate with their respective BSs. SINR is utilized for resource allocation which depends on the acquisition of CSI with the communication protocol of USRPs. The CSI acquisition is performed in two stages: synchronization and acquisition stage. In the synchronization stage, each PBS performs synchronization with CBSs operating under its coverage. A beacon signal is periodically transmitted to compensate for any misalignment in the clock frequency of each node. Further, in the acquisition stage, all CRs and PUs acquire CSI in terms of effective channel gain of all links for each resource and report their CSI to their respective BSs. Here, two PBSs, two CBSs, two PUs, and two CRs form a real-time environment with USRPs hardware as shown in Fig. 2. In addition to this, two host computers to run LabVIEW CSDS 2.0 for baseband processing are used. However, two host computers act as servers for each PBS in a decentralized manner. Finally, the CSI is utilized in python coding to develop the CA-MARL algorithm for resource allocation as shown in Fig. 3. Table I presents the parameters used for testbed implementation. The value of the parameter $\Gamma=(1000/\mathrm{Z})$ where Z represents episodes. As the value of Z increases, the value of Γ reduces. Further, the calibration parameters and specifications of USRP-2954 for testbed implementation are presented in Table 2.

5.2 Experimental Results

In this sub-section, the performance of proposed resource allocation has been demonstrated in terms of average network capacity and convergence speed. The convergence speed measures the number of time steps required to reach optimal policy. The average network capacity obtained during the learning process is plotted in Fig. 4 with perfect and imperfect CSI considerations. It is observed that the time steps taken to

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

Fig. 2. Real-time environment setup.

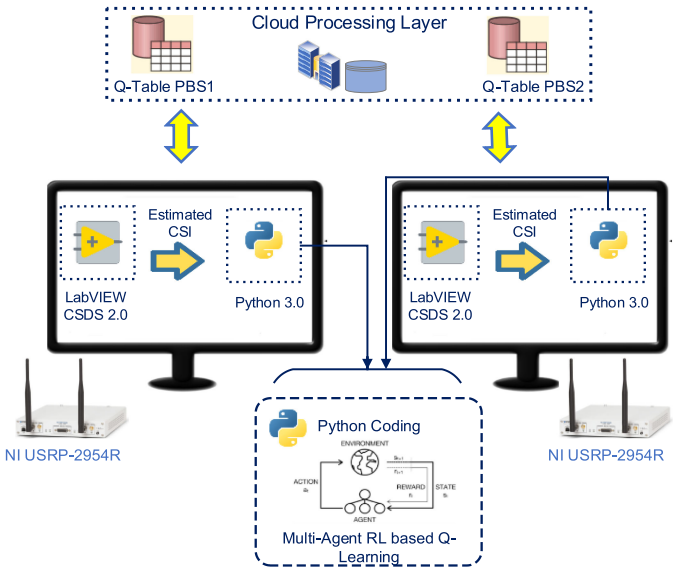

Fig. 3. Cloud assisted implementation topology.

TABLE 1
System Parameters for Testbed

CONFIGURATION PARAMETERS	
Parameters	**Value**
System Bandwidth	10 MHz
PBS Transmission Power, $p_{u,\max}$	43 dBm
CBS Transmission Power, $p_{b,\max}$	20 dBm
LEARNING PARAMETERS	
Learning Rate (α)	0.6
Discount Factor (γ)	0.9

TABLE 2
Parameters and Specifications of USRP-2954 (R)

Acquisition Stage Time Slot	3
Sensing Duration	10 ms
Range of Frequency	10 MHz to 6 GHz
Frequency Step	<1 KHz
Gain Range	0 dB to 31.5 dB
Gain Step	0.5 dB
Maximum instantaneous real-time bandwidth	160 MHz
Maximum I/Q sample rate	200 MS/s

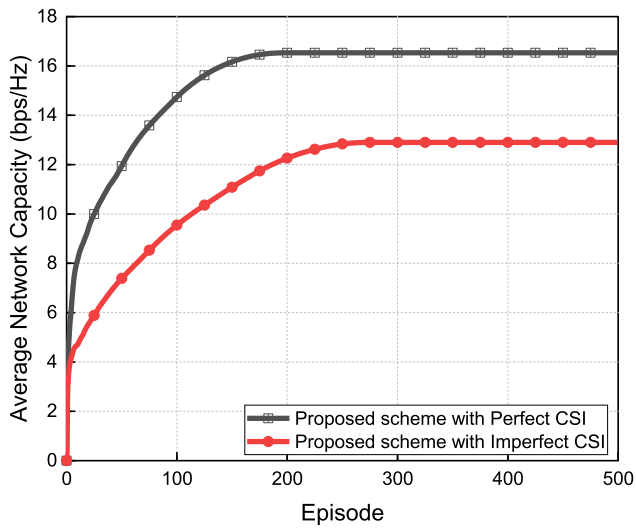

Fig. 4. Average network capacity and convergence under perfect and imperfect CSI.

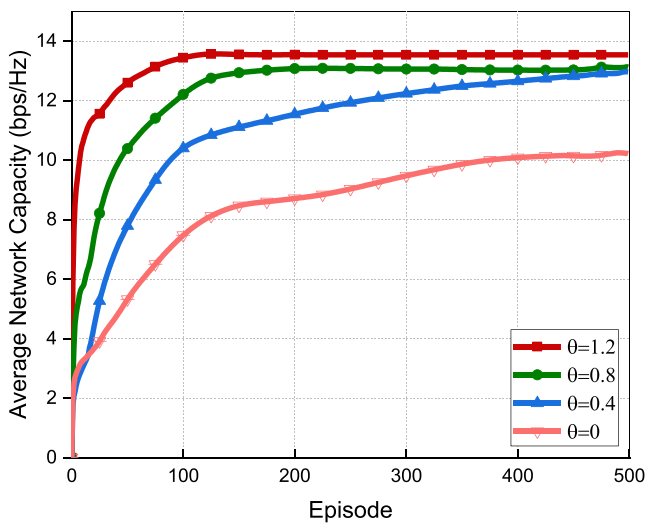

Fig. 5. Convergence with different uncertainty bound θ.

reach optimal policy are 180 and 240 under perfect and imperfect CSI respectively. The Fig. 4 shows that the proposed scheme performed better even with CSI imperfections. This is due to the consideration of uncertainty bound θ, which provides robust design to improve average network capacity.

The plot of the average network capacity against the number of episodes with the varying value of uncertainty bound is presented in Fig. 5. It is observed from Fig. 5 that the performance of the proposed algorithm is degraded significantly under imperfect CSI. Specifically, when the robust design is not considered, the uncertainty bound θ is set to 0, the outage probabilities (probability of the SNR being below threshold) become high, which reduces average network capacity. With the increase of the uncertainty bound θ, the robust design is adopted to decrease the outage probabilities at the cost of more transmission power. When θ is bigger than 1.2, no outage will occur and hence obtains higher network capacity. The convergence of the proposed scheme is also verified under the varying value of the discount factor as shown in Fig. 6. It is observed that average network capacity increases with an increase in the value of the

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

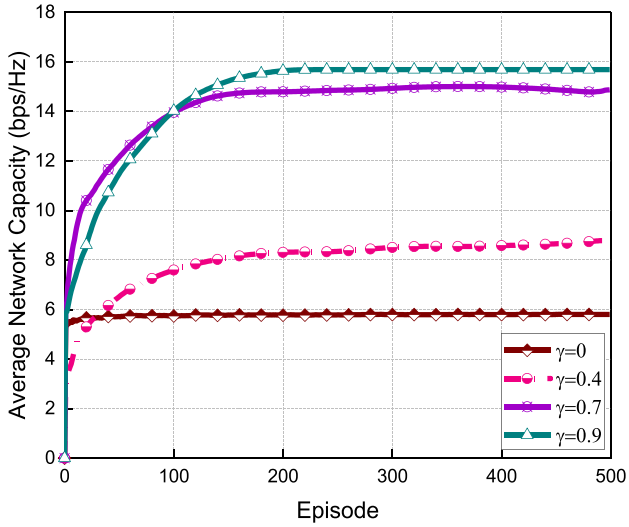

Fig. 6. Average network capacity and convergence under the varying value of the discount factor γ.

discount factor. When $\gamma = 0$, the proposed scheme becomes myopic in which action is selected based on immediate reward. But with an increase in the value of γ, each agent tends to maximize its future rewards instead of immediate reward and hence outperforms than myopic approach. Thus, the value of the discount factor γ is set to be 0.9. Higher the value of the discount factor, the greater the importance of future rewards relative to the current reward. Further, the convergence of the proposed scheme is also verified for the varying value of the learning rate α as shown in Fig. 7. The learning rate α is chosen closer to 1 to speed up the convergence as α represents the extent with which new Q-value overrides the current Q-values. However, it is observed from Fig. 7 that higher the value of α, reduces average network capacity as a reward.

This is because it traps the algorithm in local optimization which reduces reward. Thus, the optimum value of α is set to be 0.6. The experimental results in Figs. 4 to 7 demonstrates the capability of the implemented CA-MARL scheme for resource allocation under the varying value of parameters.

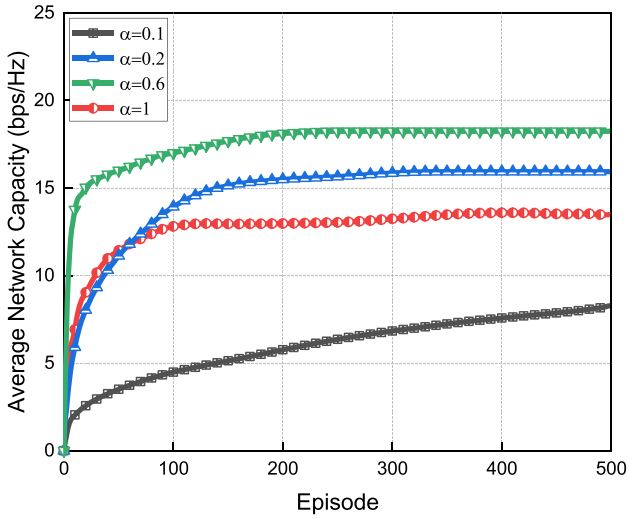

Fig. 7. Average network capacity and convergence under the varying value of learning rate α.

TABLE 3
Simulation Parameters

Parameters	Value
System Bandwidth	10 MHz
Number of Resources	50
PBS Transmission Power, $p_{u,\max}$	43 dBm
CBS Transmission Power, $p_{b,\max}$	20 dBm
Thermal Noise power	-114 dBm/Hz
Total Number of Episodes	500
SINR Threshold, ξ^*	10 dB
PU Interference Threshold, I_m^{th}	5×10^{-12} W

Moreover, it is noticed that the proposed scheme converges even under imperfect CSI within a short duration. This is due to the combination of ellipsoidal approximation along with Reinforcement learning that reduces the difference between the approximated CSI and the perfect one.

5.3 Numerical Results

In this sub-section, the performance of the proposed scheme is verified in terms of average network capacity and outage probability. The simulation environment consists of two PBSs with 10 PUs, 6 CBS with 47 CRs accessing 50 available resources, each of bandwidth 200 kHz. It is assumed that CRs are distributed randomly and uniformly within the coverage of CBS and Each PBS is aware of CBSs operating in its coverage area. Further, it is assumed that all CRs have identical and independent Rayleigh fading channels with channel gain expressed as $G = (d)^{-c}$ where d is the physical distance between the transmitter and receiver and c represents the path-loss factor which is set to 4. Here, each CBS and PBS consider a similar type of data traffic. The QoS requirement of PUs (C_m^*) and CRs (C_k^*) are assumed to be 512 Kbps and 256 Kbps respectively. It is also assumed that the inter-tier interference from CBSs to PUs remains constant. The rest of the simulation parameters are presented in Table 3. Further, the performance of the proposed scheme is compared with other schemes with imperfect CSI considerations such as Greedy based joint resource allocation scheme [19] and Dual Decomposition scheme [20]. Besides, independent Q-L without cooperation is also considered for comparison. The average network capacity obtained under these schemes is demonstrated in Fig. 8. It is observed that the proposed scheme obtains higher network capacity among other schemes. This is because the robust design is considered with uncertainty bound $\theta = 1.6$, due to which outage probability decreases with an increase in transmission power.

Further, network capacity and outage probability with uncertainty bound of CSI imperfections are shown in Figs. 9 and 10. It should be noted that with the increase in uncertainty bound θ, outage probability decreases, and hence network capacity increases. Moreover, in Greedy based joint resource allocation scheme, excessive power is transmitted which reduces outage probability and performs better than the Dual Decomposition scheme. In the proposed scheme, outage probability becomes minimum at $\theta = 1.6$. It should also be noted that ignoring the effect of channel uncertainty will result in high outage probability.

Further, the impact of dynamic network conditions including various users' data rate requirements, and dynamic

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

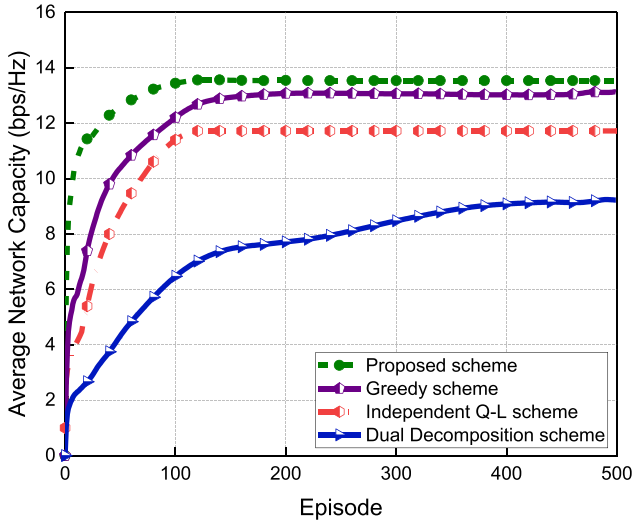

Fig. 8. Average network capacity with different schemes under imperfect CSI.

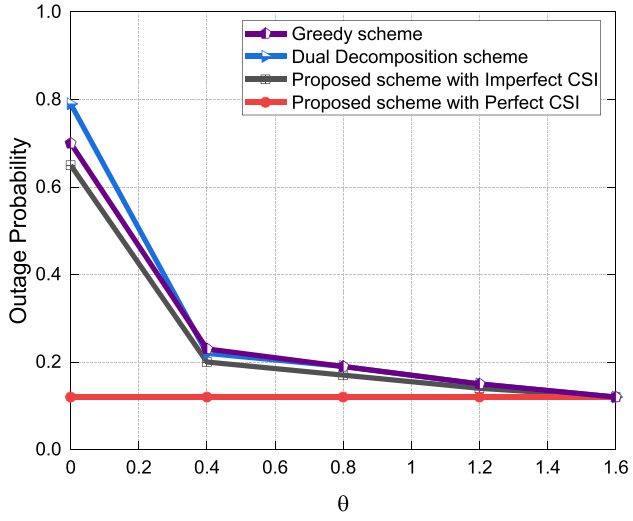

Fig. 10. Outage probability with varying values of uncertainty bound.

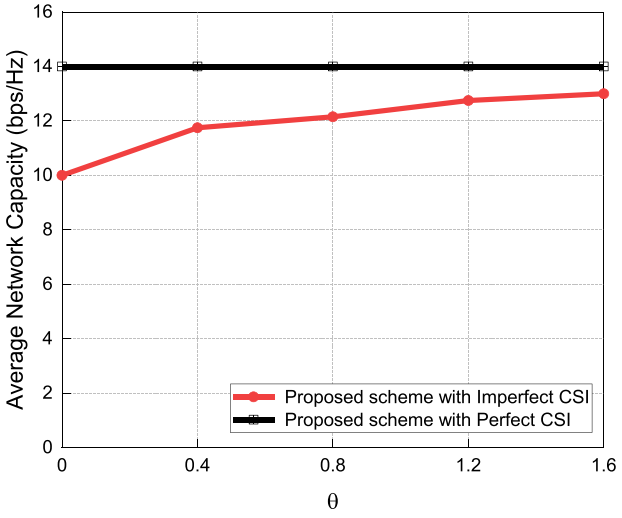

Fig. 9. Average network capacity with different values of uncertainty bound.

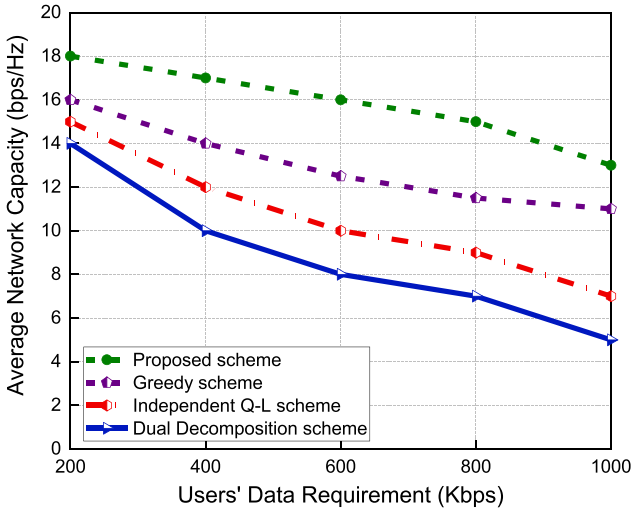

Fig. 11. Average network capacity with variable users' data rate requirement.

channel conditions represented by SINR on the achieved network capacity has been investigated. For the diverse users' data rate requirements, the performance of the proposed CA-MARL scheme has been evaluated. The transmission power of PBSs and CBSs are set as specified in Table 3. Fig. 11 presents the average network capacity achieved as a function of variable users' data rate requirements. It is observed from Fig. 11 that the proposed scheme maintains the average network capacity at the highest level as compared to other schemes despite an increase in the users' data rate demands. It is further noticed that the average network capacity degrades as the users' data rate requirements increases. This is because a higher data rate requires higher transmission power of the CBSs thus degrades the overall average network capacity. In addition, the impact of varying channel conditions on the achieved average network capacity has been studied. The impact of varying SINR on achieved network capacity is plotted in Fig. 12 with the assumption that other simulation parameters remain constant. It is observed from the figure that the proposed scheme outperforms all other schemes even under poor channel conditions. This is due to the cooperative behavior of the proposed scheme that each agent performs its action with complete information of other agents' strategies which improves system performance.

From the presented results, it is observed that the proposed scheme for resource allocation outperforms other schemes in terms of average network capacity and outage probability with a cooperative framework in a decentralized manner. The use of machine learning in the resource allocation scheme results in superior performance as compared to the simple convex optimization in a dual decomposition scheme for resource allocation with imperfect CSI consideration without forming a specific network model. This is a significant factor in highly dynamic CR networks, which cannot be tied to a specific model. Thus, model-free Reinforcement learning, which adopts learning from experience approach is a good fit for the resource allocation problem. Furthermore, Table 4 presents the comparison of experimental and numerical results in terms of average network capacity and convergence speed.

The results are compared based on the optimal (maximum) value reached for average network capacity. It is observed

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

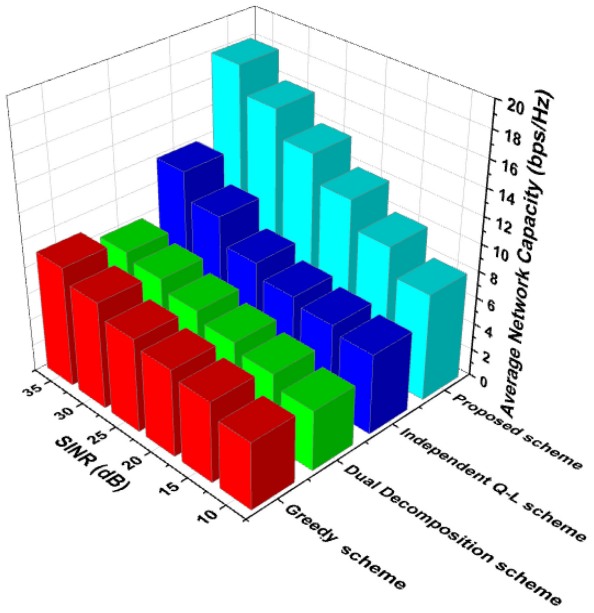

Fig. 12. Average network capacity with varying SINR.

TABLE 4
Comparison of Experimental and Numerical Results

Performance Metric	Numerical	Experimental
Average Network Capacity	18 bps/Hz	16 bps/Hz
Convergence Speed	100	125

that the obtained numerical results record better values than the experimental one. The reason is real-time interference due to the existence of other wireless devices nearby the operating hardware which degrade the achieved experimental results.

6 Conclusion

This paper investigates the resource allocation problem in CR networks with imperfect CSI considerations. Since perfect CSI is hardly obtained. Thus, a robust resource allocation scheme that efficiently tackles CSI imperfections without forming an explicit network model to learn is proposed.

A decentralized Cloud Assisted cooperative Multi-Agent Reinforcement Learning scheme is proposed with CSI imperfections modeled as an ellipsoidal approximation. This scheme allows cooperation among multi-agents in a decentralized manner by sharing their past strategies as historical information in the cloud. The cooperation expedites the learning process without increasing cooperation overhead. Numerical results demonstrate the merits of the proposed scheme in terms of convergence speed, average network capacity, and outage probability. In general, presented results authenticate that the proposed scheme provides an outstanding approach in a highly dynamic environment especially when it is difficult to model complex network dynamics.

Furthermore, it is interesting to extend this work with consideration of multiple application requirements of CRs in cooperative CR networks.

Acknowledgments

This work was supported by the Department of Science and Technology, New Delhi, Government of India under Women Scientist Scheme-A (No. SR/WOS-A/ET-52/2017), and Science and Engineering Research Board (No. EEQ/2017/000592).

References

[1] I. Al Qerm and B. Shihada, "Energy-efficient power allocation in multitier 5G networks using enhanced online learning," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11086–11097, Dec. 2017.
[2] White paper: Cisco Visual Networking Index: Forecast and Trends, 2017-2022, Cisco, Tech. Rep., San Jose, USA, 2018. [Online]. Available: https://networking.report/whitePapers/cisco-visual-networking-index-forecast-and-trends-20172022/6552
[3] FCC, Docket no. 03-322 Notice of proposed rule making and order, 2003. [Online]. Available: http://www.cs.ucdavis.edu/liu/2891/Material/FCC-03-322A1.pdf
[4] K. Kumar, A. Prakash, and R. Tripathi, "Context aware spectrum handoff scheme in cognitive radio vehicular networks," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 24, no. 1/2, pp. 101–116, 2017.
[5] K. Kumar, A. Prakash, and R. Tripathi, "Spectrum handoff in cognitive radio networks: A classification and comprehensive survey," *J. Netw. Comput. Appl.*, vol. 61, pp. 161–188, 2016.
[6] D. Wang, B. Song, D. Chen, and X. Du, "Intelligent cognitive radio in 5G: AI-based hierarchical cognitive cellular networks," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 54–61, Jun. 2019.
[7] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1136–1159, Third quarter 2013.
[8] T. Jiang, D. Grace, and Y. Liu, "Two-stage reinforcement-learning-based cognitive radio with exploration control," *IET Commun.*, vol. 5, no. 5, pp. 644–651, Mar. 2011.
[9] I. Mustapha, B. M. Ali, A. Sali, M. F. Rasid, and H. Mohamad, "An energy efficient reinforcement learning based cooperative channel sensing for cognitive radio sensor networks," *Pervasive Mob. Comput.*, vol. 35, pp. 165–184, Feb. 2017.
[10] R. S. Sutton and A. G. Barto, *An introduction to Reinforcement Learning*, Cambridge, UK: MIT Press, 1998.
[11] N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated Q-Learning for robust cognitive spectrum management in LTE cellular systems," *IEEE Trans. Mobile Comput.*, vol. 15, no. 4, pp. 817–825, Apr. 2016.
[12] A. M. Koushik, F. Hu, and S. Kumar, "Intelligent spectrum management based on transfer actor-critic learning for rateless transmissions in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1204–1215, May 2018.
[13] K. K. Nguyen, T. Q. Duong, N. A. Vien, N. A. Le-Khac, and M. N. Nguyen, "Non-cooperative energy-efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100480–100490, Jul. 2019.
[14] I. Alqerm and B. Shihada, "A cooperative online learning scheme for resource allocation in 5G systems," in *Proc. IEEE Int. Conf. Commun.*, 2016, pp. 1–7.
[15] J. B. Wang *et al.*, "A machine learning framework for resource allocation assisted by cloud computing," *IEEE Netw.*, vol. 32, no. 2, pp. 144–151, Mar./Apr. 2018.
[16] J. B. Wang, N. Li, J. Y. Wang, Y. P. Wu, M. Cheng, and M. Chen, "Online learning based transmission scheduling for delay-sensitive data over a fading channel with imperfect channel state information," *IEEE Access*, vol. 5, pp. 13225–13235, Jul. 2017.
[17] J. B. Wang, M. Feng, X. Song, and M. Chen, "Imperfect CSI based joint bit loading and power allocation for deadline constrained transmission," *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 826–829, May 2013.
[18] R. Masmoudi, E. V. Belmega, and I. Fijalkow, "Impact of imperfect CSI on resource allocation in cognitive radio channels," in *Proc. IEEE 13th Int. Conf. Wireless Mob. Comput. Netw. Commun.*, 2017, pp. 293–299.
[19] J. B. Wang *et al.*, "Imperfect CSI-based joint resource allocation in multi-relay OFDMA networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3806–3817, Oct. 2014.

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.

[20] T. Al-Khasib, M. B. Shenouda, and L. Lampe, "Dynamic spectrum management for multiple-antenna cognitive radio systems: Designs with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 2850–2859, Sep. 2011.
[21] W. Wang, A. Kwasinski, D. Niyato, and Z. Han, "A Survey on applications of model-free strategy learning in cognitive wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1717–1757, Third quarter 2016.
[22] C. J. C. H. Watkins and P. Dayan, "Q-Learning," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.
[23] A. H. Ko, R. Sabourin and F. Gagnon, "Performance of distributed multi-agent multi-state reinforcement spectrum management using different exploration schemes," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4115–4126, Aug. 2013.
[24] A. Szepesvári and M. L. Littman, "A unified analysis of value function based reinforcement learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 2017–2060, Nov. 1999.
[25] An introduction to software defined radio with NI LabVIEW and NI USRP, National Instruments, Texas, Austin, USA, 2013. [Online]. Available: http://www.ni.com/tutorial/1451/en/

Amandeep Kaur (Student Member, IEEE) received the BTech degree from the Department of Electronics and Communication Engineering, Guru Nanak Dev University, Regional Campus, Gurdaspur, Punjab, India, in 2011, and MTech degree in electronics and communication engineering (communication systems) from Guru Nanak Dev University, Amritsar, Punjab, India, in 2013. Currently, she is working toward the PhD degree in electronics and communication engineering from the National Institute of Technology, Hamirpur, India. From 2014 to 2018, she was an assistant professor at the National Institute of Technology, Hamirpur, India (Institute of National Importance). She is also handling one project under Women Scientist Scheme sponsored by Department of Science and Technology, Government of India. Her research interests include wireless communication, and machine learning especially the application of artificial intelligence in cognitive radio networks.

Krishan Kumar (Member, IEEE) received the BE degree from the Department of Electronics and Communication Engineering, CR State College of Engineering, Murthal, Haryana, India, in 2002, and the MTech degree in electronics and communication engineering from the National Institute of Technology, Kurukshetra, India (Institute of National Importance), in 2005, and the PhD degree from the Department of Electronics and Communication Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India, under Quality Improvement Program scheme. Presently, he has been working as an assistant professor at the National Institute of Technology, Hamirpur, India (Institute of National Importance), since 2006. He is the reviewer of various SCI-indexed journals. He is handling various projects sponsored by Department of Science and Technology, Government of India. His research interest includes wireless communication especially cognitive radio networks, vehicular networks with mobility management issues, and application of artificial intelligence.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: Central South University. Downloaded on March 22,2024 at 08:00:32 UTC from IEEE Xplore. Restrictions apply.