Wireless Netw (2013) 19:1237–1250
DOI 10.1007/s11276-012-0530-4

Reinforcement learning for cooperative sensing gain in cognitive radio ad hoc networks

Brandon F. Lo · Ian F. Akyildiz

Published online: 25 December 2012

Abstract Spectrum sensing is a fundamental function in cognitive radio networks for detecting the presence of primary users in licensed bands. The detection performance may be considerably compromised due to multipath fading and shadowing. To resolve this issue, cooperative sensing is an effective approach to combat channel impairments by cooperation of secondary users. This approach, however, incurs overhead such as delay for reporting local decisions and the increase of control traffic. In this paper, a reinforcement learning-based cooperative sensing (RLCS) method is proposed to address the cooperation overhead problem and improve cooperative gain in cognitive radio ad hoc networks. The proposed algorithm is proven to converge and capable of (1) finding the optimal set of cooperating neighbors with minimum control traffic, (2) minimizings the overall cooperative sensing delay, (3) selecting independent users for cooperation under correlated shadowing, and (4) excluding unreliable users and data from cooperation. Simulation results show that the RLCS method reduces the overhead of cooperative sensing while effectively improving the detection performance to combat correlated shadowing. Moreover, it adapts to environmental change and maintains comparable performance under the impact of primary user activity, user movement, user reliability, and control channel fading.

Keywords Ad hoc networks · Cognitive radio · Control channel · Cooperative spectrum sensing · Cooperative gain · Reinforcement learning

1 Introduction

The primary goal of spectrum sensing in cognitive radio (CR) networks [2, 3] is to identify available licensed spectrum for secondary user (SU) transmission while reducing the interference with primary users (PUs) to a tolerable level to improve spectrum utilization. However, the spectrum sensing results obtained by an individual SU is more susceptible to detection errors due to shadowing effects and multipath fading in wireless channels. Imperfect sensing results in either wasting spectrum opportunity, known as false alarm when a PU is mistakenly considered present in an available spectrum, or interfering with PUs, known as miss detection when a present PU in the licensed band is wrongfully considered absent. Since SUs are located in different places and the probability of all SUs in deep fading or shadowing is small, combining local sensing results by an SU, known as fusion center (FC), to make cooperative detection decisions for all participating SUs reduces the possibility of making incorrect decisions at each individual user. The process of spectrum sensing with cooperation among SUs is called cooperative sensing [4]. Previous studies [5, 6] have demonstrated that cooperative sensing can effectively combat multipath and shadow

A preliminary version of this work was presented at IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, September 2010 [1].

B. F. Lo (✉) · I. F. Akyildiz
Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: blo3@gatech.edu

I. F. Akyildiz
e-mail: ian@ece.gatech.edu

I. F. Akyildiz
Faculty of Computing and Information Technology, King Abdul-Aziz University, P.O. Box 80221, Jeddah 21589, Saudi Arabia

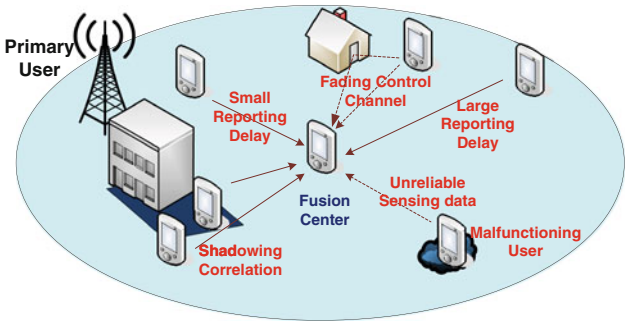

Fig. 1 Cooperative sensing and possible cooperation overhead that limits cooperative gain

fading to reduce the miss detection and false alarm probabilities by utilizing *spatial diversity* of cooperating SUs. Regardless of the benefits of cooperative sensing, cooperation incurs overhead that limits the cooperative gain [4]. Figure 1 illustrates an example of cooperative sensing and possibly incurred cooperative overhead in a CR ad hoc network (CRAHN).

The major types of cooperation overhead under consideration are (1) shadowing correlation, (2) control message overhead, (3) synchronization and reporting delay, and (4) user and data reliability. First, it is known that shadowing correlation degrades the performance of cooperative sensing [6]. This is because SUs, spatially located in proximity and blocked by the same obstacle, may experience correlated shadowing and have poor observations of PU signals. As a result, cooperative gain is limited by shadowing correlation. Second, cooperation requires extra control message exchange among SUs for reporting sensing data on a common control channel (CCC) [7, 8]. Such control transmission is also limited by the available CCC bandwidth. Third, synchronizing SUs in CRAHNs for sensing cooperation is not a trivial task. Since SUs have different transmission and sensing schedules, the local sensing results from cooperating SUs may not simultaneously arrive at the FC. Moreover, control packet collision and re-transmission in control channel result in extra reporting delay. Thus, asynchronous reporting and delay overhead should be considered in cooperative sensing. Finally, the reported sensing results may be unreliable due to the malfunctioning of SUs, or manipulation of malicious SUs, known as the Byzantine failure problem [9]. Furthermore, control channel fading incurs reporting errors, which may further complicate the reliability issue. Therefore, a mechanism to exclude unreliable cooperating users and their sensing results from cooperation must be included in cooperative sensing.

Existing cooperative sensing solutions are mainly based on the model of parallel fusion network in distributed detection [10], where all cooperating CR users generate local decisions and report them simultaneously to FC for making global decisions by data fusion. To mitigate correlated shadowing, [11] takes into account user correlation in the linear-quadratic fusion method to improve detection performance in correlated environment. In addition, [12] proposes user selection algorithms based on location information to find uncorrelated users for cooperative sensing. However, these solutions may not be adaptive to dynamic environmental changes in a timely fashion. To reduce control messages overhead, [13–16] report quantized and binary sensing data for soft and hard decision combining, respectively. Alternatively, [17] reduces the average number of reporting bits by restraining unreliable sensing results from being reported. For synchronization and delay issues, recent studies [16, 18, 19] consider the asynchronous case where cooperating SUs report local results at different time. However, conventional schemes based on the parallel fusion model [9, 11, 15] typically assume conditional independence on observations among SUs and perfect SU synchronization with instant reporting on an error-free CCC. Moreover, existing cooperative sensing methods seldom address all aforementioned cooperation overheads in response to dynamic environmental changes. Thus, it is clear that a new model for cooperative sensing with the capability of interacting with and learning from the environment is required to address all these problems in CRAHNs.

In this paper, we introduce a novel reinforcement learning-based cooperative sensing (RLCS) method to address incurred cooperation overheads and improve detection performance in multipath and correlated shadow fading. Reinforcement learning (RL) [20] is an adaptive method for a decision-making agent learning to choose optimal actions and maximize received rewards by interacting with its environment. In RLCS, the SU acting as the FC is the decision-making agent interacting with the environment that consists of its cooperating neighbors and their observations of PU activity. By requesting sensing results from its neighbors, the FC learns the behaviors of cooperating SUs and takes actions to select users for cooperation through periodic cooperative sensing. Among a variety of RL algorithms, we utilize temporal-difference (TD) learning [20] to address cooperation overhead issues in cooperative sensing because of its capability to evaluate the correlation between successive SU selections and adjust subsequent selection predictions based on the experiences accumulated over time to mitigate cooperation overheads by selecting an optimal set of cooperating SUs. More importantly, TD learning enables SUs to learn from and adapt to dynamic environmental change, such as the changes in PU activity, user movement, user reliability, and channel conditions, while maintaining satisfactory sensing performance without requiring a model or a priori knowledge about PU activity and SU's behavior. Obviously,

these benefits of TD learning cannot be obtained by pre-selection of SUs or cooperation of all SUs with no learning. Although RL algorithms have been applied to dynamic channel access [21, 22], user selections [23], and multi-band sensing policy [24, 25] in CR networks, to the best of our knowledge, RLCS is the first work applying the RL techniques to address both the cooperation overhead problems and the detection performance of cooperative sensing. Our contributions can be summarized as follows:

- We propose the novel RLCS model and algorithm for SUs to learn the optimal user selection policy for finding uncorrelated and reliable cooperating neighbors to improve cooperative gain and mitigate cooperation overhead in cooperative sensing.
- We show that the optimal solution obtained by RLCS approach greatly improves the detection performance under correlated shadowing while minimizing control channel bandwidth requirement by using binary local decisions and hard-combining strategy.
- We demonstrate that RLCS converges asymptotically with the option of optimal stopping for fast response in dynamic environment, mitigates the impact of control channel fading, improves the reliability of user and sensing data selection, and adapts to PU activity change and the movement of SUs.

The remainder of this paper is organized as follows: Sect. 2 presents our system model and assumptions. Section 3 describes the proposed RLCS scheme. Section 4 discusses the performance analysis of our proposed scheme. Section 5 evaluates the performance by test scenarios and numerical results, and Sect. 6 concludes the paper.

2 System model

We consider a group of SUs forming a CRAHN overlaid with a primary network to opportunistically share a set of N_C licensed channels. Each licensed channel is assumed to be occupied by one primary transmitter (the PU) and potential primary receivers in its transmission range. In order to protect these primary receivers from interference, the range of PU transmission R_P plus the range SU transmission R_S, $R_P \gg R_S$, forms the protected region [11]. The PU activity on channel m is modeled as a two-state birth-death process with the birth rate r_b^m and the death rate r_d^m [26]. In this PU model, the transitions follow a Poisson process with exponentially distributed inter-arrival time. Thus, the long-term average probabilities of PU active (P_{on}^m) and inactive (P_{off}^m) on channel m are $r_b^m/(r_b^m + r_d^m)$ and $r_d^m/(r_b^m + r_d^m)$, respectively. The PU activity is unknown to SUs a priori. To balance the traffic load and power consumption, the SUs in the CRAHN may either take turns to serve as the FC to cooperatively sense one licensed channel each time, or act as FCs simultaneously to sense multiple channels at the same time. However, there is only one SU acting as the the decision-making FC (learning agent) on each channel. Without loss of generality, we focus on RLCS with one FC and its cooperating neighbors on one channel and the channel index m will be omitted from the notation thereafter unless otherwise specified. How to determine which channel to sense is beyond the scope of this work.

Let $\mathcal{C}$ be the set of the neighbors of the FC where the FC is denoted by SU_0 and $|\mathcal{C}| = L$. Let y_i be the average SNR in dBm of the received PU signal observed at cooperating user SU_i. y_i are Gaussian distributed since the received signal power in shadowing is assumed to be log-normally distributed [11]. The observations $y_i, i = 1, \ldots, L$, may be correlated depending on the location of the SUs. The collection of these observations is the Gaussian distributed vector $\mathbf{Y} = \{y_i\}_1^L$ under the null hypothesis H_0, which indicates the absence of the PU transmit signal, and the alternative hypothesis H_1, which indicates the presence of the transmit signal, as follows [11]:

$$\mathbf{Y} \sim \begin{cases} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma_0^2 I}), & H_0 \\ \mathcal{N}(\boldsymbol{\mu_1}, \boldsymbol{\sigma_1^2 \Sigma}), & H_1 \end{cases}, \tag{1}$$

where $\mathbf{0}$ is the zero vector, $\boldsymbol{\mu_1}$ is the mean SNR that depends on the path loss from the location of the PU, σ_0^2 is the Gaussian noise variance under H_0, σ_1^2 is the variance of noise in correlated shadowing under H_1, $\mathbf{I}$ is the identity matrix, and $\boldsymbol{\Sigma}$ is the normalized covariance with elements ρ_{ij}. We assume that the correlation follows the exponential correlation model [27]. In this model, the correlation coefficients can be expressed as

$$\rho_{ij} = e^{-d_{ij}/D_c} = e^{-a \cdot d_{ij}}, \tag{2}$$

where d_{ij} is the distance between SUs i and j, D_c is the de-correlation distance, and $a = 1/D_c$ is the exponential decaying coefficient set to 0.1204 and 0.002 for urban and suburban settings, respectively [6]. Thus, two SUs are correlated if the distance between them is smaller than D_c, and uncorrelated otherwise.

Depending on the distance between the PU and the SU, and the degree of fading, the SNR observed at SUs may vary significantly. Due to these variations, SUs may take different number of observations to satisfy detection requirements and report their local decisions on the CCC *asynchronously* that causes different reporting delays. We assume that a narrowband CCC is shared by the FC for message broadcast and cooperating SUs for reporting local sensing data. The time-varying wireless channel between each SU and the FC, known as the reporting channel, is susceptible to independent Rayleigh fading, which is

modeled by the finite-state Markov channel (FSMC) model [28]. Since the received SNR γ_b varies with time and ranges from 0 to infinity, the entire SNR range is divided into J regions in which the jth region is defined as $\Gamma_j = [A_j, A_{j+1}) = \{\gamma_b : A_j \le \gamma_b < A_{j+1}\}$ where $\{A_j\}$ are region boundaries with $A_0 = 0$ and $A_J = \infty$. For transmitting binary local decisions, we assume BPSK modulation at reporting SU and the coherent demodulation at the FC. In this case, the error probability can be expressed in terms of the received SNR as $P_e(\gamma_b) = Q(\sqrt{2\gamma_b})$ where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du$ is the tail probability of standard normal distribution. The reported local decision in each channel state x_j follows a binary symmetric channel (BSC) where the local decisions are received in errors at the FC with crossover probability ε_j.

Based on the proposed RLCS model and algorithm discussed in Sect. 3, the FC selects and combines these local results, and makes a cooperative decision on the presence of the PU. In general, the data fusion of selected $K \le L$ local sensing results at FC is given by

$$\sum_{i=1}^{K} w_i f(y_i) = \sum_{i=1}^{K} w_i u_i \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda_0 \tag{3}$$

where $f(\cdot)$ is the local decision process, w_i is the weighting factor for local sensing data $u_i \in \{0, 1\}$ from cooperating user SU_i and λ_0 is the cooperative decision threshold at the FC. For hard combination with the majority rule, $w_i = 1$, $\forall i$ and $\lambda_0 = \lceil K/2 \rceil$. The majority rule is chosen over *AND* and *OR* rules for the balance of false alarms and miss detection. The cooperative decision $u_0 \in \{0, 1\}$ is then broadcast to all neighbors. This cooperative sensing process is periodically repeated for infinite iterations, called *episodes*.

3 RL-based cooperative sensing

In this section, we present the proposed RLCS model and algorithm for cooperative sensing. We formulate the problem as a cooperative sensing decision process (CSDP) and discuss the process of RLCS algorithm for improving cooperative gain in cooperative sensing.

3.1 Cooperative sensing decision process

In RLCS, the interactions between the FC and cooperative SUs are modeled as a CSDP. CSDP is a decision process with non-Markovian rewards for FC's sequential decisions on selecting cooperating neighbors. Figure 2 illustrates the RLCS model with the inherent CSDP and the environment with which the agent inside the FC interacts. In the figure,

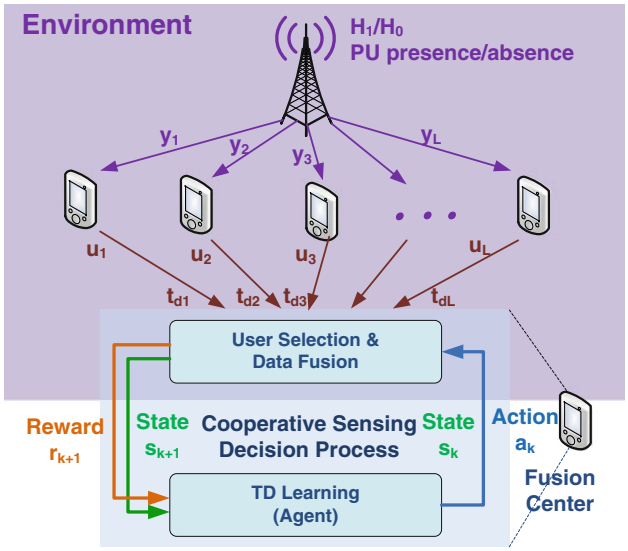

Fig. 2 The model of cooperative sensing with reinforcement learning

the FC interacts with L cooperating neighboring nodes that observe the PU activity in the environment and obtain Gaussian distributed and possibly correlated observations $\mathbf{Y} = \{y_i\}_1^L$ as in (1). In each state s_k, where k is the time step or stage index, the FC selects neighbor i by choosing action $a_k = i$ and receives local decision u_i determined from observed y_i with reporting delay t_{d_i} as reward r_{k+1} along with state change to s_{k+1}. By exploring the unknown states and accumulating the knowledge of receiving rewards from known states, the FC learns the sequence of optimal decision rules (optimal policy) that gives rise to the maximum reward.

The CSDP is represented by a quadruple $\langle \mathcal{S}, \mathcal{A}, f_p, f_R \rangle$ in which $\mathcal{S} = \{0, 1, \ldots, L\}$ is a finite set of all states, $\mathcal{A} = \cup_{j \in \mathcal{S}} \mathcal{A}_j = \{0, 1, \ldots, L\}$ is a finite set of actions, where $\mathcal{A}_j \subseteq \mathcal{A}$ is the set of actions available in state j, f_p is the state transition probability function, and f_R is the reward function. Each component is described as follows:

States: A state of the CSDP is the status of user selection and reporting in the environment that includes cooperating SUs and their observations of the PU signal. In each episode n, the states of the environment s_k, which take on values from $\mathcal{S} = \{0, 1, \ldots, L\}$, are defined as

$$s_k^n = i \cdot \mathcal{I}\{a_{k-1}^n = i \in \mathcal{A}_j, s_{k-1}^n = j \in \mathcal{S}, i \ne j, k \ne 0\}, \tag{4}$$

where $\mathcal{I}\{x\}$ is the indicator function and equals one if x is true and zero otherwise, and a_{k-1}^n is the action selected in s_{k-1}^n. The process starts with the FC state (SU_0 state) $s_0^n = 0$ when the FC initiates the process of cooperative sensing at time t_0^n. In each state $s_k^n = j \in \mathcal{S}$, the FC requests local decision u_i from SU_i and awaits a response from the environment by choosing action $a_k^n = i \in \mathcal{A}_j$, or terminates the cooperative sensing by choosing action $a_k^n = 0$ to return to

the FC state. The state changes from $s_k^n = j$ to $s_{k+1}^n = i$ when the FC obtains reported u_i and the corresponding reward r_{k+1}^n as the response at time t_{k+1}^n. The FC state is both the *start* state and the *terminal* state in each episode.

Actions: An action is the FC's decision on selecting an SU (including the FC itself) for reporting in a state. Let $h_k^n = (s_0^n, a_0^n, \ldots, s_{k-1}^n, a_{k-1}^n, s_k^n)$ be the history of state-action sequence from s_0^n to s_k^n in episode n. The decision rule μ_k^n is the function mapping h_k^n into a probability distribution $\Delta_{\mu_k^n}(\mathcal{A}_{s_k^n})$ on the set of actions $\mathcal{A}_{s_k^n}$ in state s_k^n of episode n. Let also $\mathcal{D}_{k-1}^n$ be the set of selected SUs from s_0^n to $s_{k\text{-}1}^n$ of episode n given by $\mathcal{D}_{k-1}^n = \{a_0^n, \ldots, a_{k-1}^n\}$ and $\mathcal{D}_{-1}^n = \emptyset$. Thus, the actions of the FC a_k^n in state s_k^n, which take on values from $\mathcal{A}_{s_k^n} = \mathcal{A} \setminus \{\{s_k^n\} \cup \mathcal{D}_{k-1}^n\}$, are defined as:

$$a_k^n = \mu_k^n(h_k^n) \in \mathcal{A}_{s_k^n}, \quad \text{w.p. } p(s_k^n, \mu_k^n(h_k^n)) \tag{5}$$

where $p(s_k^n, \mu_k^n(h_k^n))$, defined in (9), is the probability of selecting a_k^n in s_k^n according to $\Delta_{\mu_k^n}(\mathcal{A}_{s_k^n})$. In each state of episode n, the FC selects SU i with probability $p(s_k^n, a_k^n = i)$ for reporting in s_k^n. Specifically, the FC requests cooperating SU i to report local decision u_i by sending $a_k^n = i$, or informs all cooperating SUs the cooperative decision u_0 by sending $a_k^n = 0$ along with u_0. In the latter, action $a_k^n = 0$ also terminates one round of cooperative sensing. Nevertheless, how to choose the actions depends on the action selection strategy defined in Sect. 3.2.

Transition probability function: The transition probability function, $f_p : \mathcal{H} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ where $\mathcal{H} = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \cdots \times \mathcal{S}$, maps the state-action-state transitions to a probability of changing from the current state to the next state by choosing the action. The transition probability from state $s_k^n = j$ to $s_{k+1}^n = i$ by choosing action $a_k^n = i$ in s_k^n is denoted by $p_{ji} = P(i \mid j, \ \mu_k^n(h_k^n))$ and is generally not known a priori. Since a chosen action implies the transition to a particular state in our model, the probability of choosing an action $a_k^n = i$ in state $s_k^n = j$ can be considered as the state transition probability from $s_k^n = j$ to $s_{k+1}^n = i$. As a result, the FC gradually learns the state transitions from the action selection probabilities, even though the transition probabilities are not required by TD learning algorithms.

Reward function: The reward function, $f_R : \mathcal{H} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, maps the state-action-state transitions to a real-valued reward. The reward is used by the FC to evaluate action selections for choosing uncorrelated SUs with small reporting delay for cooperative sensing. The FC receives a reward r_{k+1}^n upon the arrival of the local sensing data u_i from SU_i with reporting delay $t_{d_i}^n$ as the result of action $a_k^n = i$ in state $s_k^n = j$. The reward r_{k+1}^n corresponding to the action a_k^n in state s_k^n of episode n is given by:

$$r_{k+1}^n(s_k^n, \mu_k^n(h_k^n)) = r_{\rho_{k+1}}^n \mathcal{I}\{C_{\rho_{k+1}}^n \neq 0\} + r_{d_{k+1}}^n \mathcal{I}\{C_{\rho_{k+1}}^n = 0\}, \quad k = 0, \ldots, K^n - 1, \tag{6}$$

where $r_{\rho_{k+1}}^n = -C_{\rho_{k+1}}^n$ and $r_{d_{k+1}}^n = 1 - C_{d_{k+1}}^n$ are the rewards attributed to correlation cost $C_{\rho_{k+1}}^n$ in (7) and delay cost $C_{d_{k+1}}^n$ in (8), respectively, and $K^n \leq L$ is the number of selected cooperating SUs in episode n. Note that $r_{k+1}^n = 0$ for $K^n \leq k \leq L$. (6) states that r_{k+1}^n is determined by the delay cost if the selected SU is uncorrelated with previously selected SUs, and by the correlation cost if the selected SU incurs correlation. r_{k+1}^n is positive only when selected SUs are uncorrelated and their cumulative reporting delay is within the delay constraint.

The correlation between SUs' observations in correlated shadowing is captured by the covariance matrix $\boldsymbol{\Sigma}$ in (1). The elements of $\boldsymbol{\Sigma}$ correlation coefficients ρ_{ij}, are estimated by using location information and (2). These correlation coefficients affect the correlation cost (7) and the reward (6) obtained in each state. Given different SU_j, $j \neq i$ selected in state $s_l, l = 0, \ldots, k-1$ and $\Sigma = \{\rho_{ij}\}$, correlation cost $C_{\rho_{k+1}}$ is given by:

$$C_{\rho_{k+1}}^n = \left[\frac{1}{k}\sum_{\ell=0}^{k-1} |\rho_{ij}(s_\ell, a_\ell = j)|\right] \mathcal{I}\{k > 0\}, \quad j \neq i. \tag{7}$$

Thus, the correlation cost is simply the average of correlation coefficients between the newly selected SU_i and each selected SUs in previous k states.

The delay cost $C_{d_{k+1}}^n$, on the other hand, is attributed to reporting delays. The reporting delay of SU_i in $s_k^n, t_{d_i}^n = t_d(s_k^n, a_k^n = i)$, is the interval between the time of the FC requesting SU_i's cooperation with the action $a_k^n = i$ and the arrival time of the local sensing data u_i at the FC. Thus, the delay cost $C_{d_{k+1}}^n$ is given by:

$$C_{d_{k+1}}^n = \frac{\sum_{\ell=0}^{k-1} t_d(s_\ell^n, a_\ell^n = j) + t_d(s_k^n, a_k^n = i)}{T_{lim}}, \quad i \in A_{s_k}, j \in A_{s_\ell}, i \neq j \tag{8}$$

where $T_{lim} = \min\{T_{c_{max}}, T_{d_{avg}}\}$ is the total reporting delay constraint, $T_{c_{max}}$ is the maximum allowed cooperative sensing time, and $T_{d_{avg}} = \sum_{j=1}^{L} \bar{t}_{d,j}$ is the total average reporting delay of all SUs. It is simply the cumulative reporting time up to the start of next state s_{k+1}^n normalized by the factor of maximum cooperative sensing time or total average reporting delay whichever is smaller. This means that the reward attributed to the delay cost $r_{d_{k+1}}^n$ is lower for SUs to be selected in the later stage than the earlier stage, which enforces the SU with large average reporting delay to be less attractive for participation, especially in the later stage when the cooperative decision needs to be determined within the limit $T_{c_{max}}$.

From (7), (8) and the definition of r_{k+1}^n in (6), we know that negative rewards are obtained when the selected SUs are correlated or their cumulative reporting delays exceed the maximum tolerable cooperative sensing time. Such selections are learned and will be depreciated from future selections. Positive rewards are possible only when all selected SUs are uncorrelated. Thus, large positive rewards are more likely attributed to selecting more uncorrelated SUs with small reporting delays within the time constraint $T_{c_{max}}$.

3.2 RL-based cooperative sensing algorithm

Based on the CSDP model, the RLCS algorithm learns the environment by iteratively choosing actions, receiving rewards, and evaluating action selections with the objectives of maximizing received rewards and improving cooperative gain. In the following, we discuss action selection strategy, expected cumulated reward for optimal policy, state-action value updates for action evaluation, user selection for reliable cooperation, and the impact of control channel fading on reporting errors and detection performance.

Action Selection Strategy: The action selection strategy affects how the FC interacts with the environment. In RLCS, the softmax approach based on Boltzmann distribution is utilized for action selections. In this action strategy, the probability of selecting action $a_k^n = i$ in state s_k^n is given by:

$$p(s_k^n, a_k^n = i) = \frac{e^{Q(s_k^n, a_k^n = i)/\tau^n}}{\sum_{j=1}^{|\mathcal{A}_{s_k^n}|} e^{Q(s_k^n, a_k^n = j)/\tau^n}}, \quad i \in \mathcal{A}_{s_k^n} \tag{9}$$

where $Q(s_k^n, a_k^n)$ is the state-action value (*Q-value*) function that evaluates the quality of choosing action a_k^n in state s_k^n, and τ^n is an episode-varying parameter called *temperature* that controls the degree of exploration versus exploitation. For large values of τ^n, all actions are equally probable. In this case, the FC explores the opportunities of more uncorrelated cooperating SUs to achieve potentially higher detection probability in the future with large τ^n. For small τ^n, on the other hand, the action with maximum $Q(s, a)$ is favored. Hence, the agent exploits the current knowledge of best selections of cooperating SUs to achieve the potentially highest detection probability with small τ^n. As a result, τ^n remains a large value for exploration in highly dynamic environment while τ^n is decreased to a small value for exploitation in static environment where the convergence can be assured [29]. To achieve the convergence in a certain number of episodes, we use a linear function to decrease the value of τ^n over episodes as follows:

$$\tau^n = -(\tau^0 - \tau^N) \cdot n/N + \tau^0, \quad n \leq N \tag{10}$$

where N is the number of episodes to reach the convergence, τ^0 and τ^N are the initial and the last value, respectively, of the temperature in N episodes. Note that $\tau^n \neq 0$, $\forall n$ and $\tau^n = \tau^N \approx 0$ for $n \geq N$ until any environmental change.

Expected Cumulative Rewards: The expected cumulative reward R^n of episode n is defined as

$$R^n = E[Y^n] = E\left[\sum_{k=0}^{L} r_{k+1}^n(s_k^n, \mu_k^n(h_k^n))\right], \tag{11}$$

where $Y^n = \sum_{k=0}^{L} r_{k+1}^n$ is the cumulative reward of episode n. If there are K^n SUs selected in episode n, $Y^n = \sum_{k=0}^{K^n} r_{k+1}^n$ and $r_{k+1}^n = 0$ for $k = K^n, \ldots, L$. The objective of TD learning in RLCS is to find the optimal policy $\pi^* = \{\mu_0^*, \ldots, \mu_{K^*}^*\}$, where $K^* \leq L$ is the optimal number of selected cooperating SUs, to achieve the maximum cumulative reward R_{π^*}, which leads to higher detection performance in cooperative sensing.

State-Action Value Updates: To evaluate the quality of action selections, a table known as *Q-table* of size $|\mathcal{S}| \times |\mathcal{A}|$ is used to store the Q-values for all state-action pairs. In each state s_k^n with the selection of action a_k^n, the Q-value $Q(s_k^n, a_k^n)$ for the state-action pair (s_k^n, a_k^n) needs to be updated according to the received reward r_{k+1}^n and future state-action value estimates $Q(s_{k+1}^n, a_{k+1}^n)$. The general form of the Q-value update in TD learning can be expressed as

$$Q_k^n \leftarrow (1 - \alpha_k^n)Q_k^n + \alpha_k^n\left[r_{k+1}^n + \gamma f(Q_{k+1}^n)\right], \tag{12}$$

where $Q_k^n = Q(s_k^n, a_k^n)$, α_k^n is the learning rate, γ is the discount factor for future state-action value estimates Q_{k+1}^n, and $f(Q_{k+1}^n)$ is the function of future estimates Q_{k+1}^n that depends on the TD learning algorithms used. For example, $f(Q_{k+1}^n) = \max_{a_{k+1}^n} Q_{k+1}^n$ and $f(Q_{k+1}^n) = Q_{k+1}^n$ are the future estimate functions for Q-learning [30] and Sarsa [20], respectively. The discount factor γ determines the weight of the future Q-value estimates compared to the current Q-value for the (s_k^n, a_k^n) pair. For faster convergence, the learning rate α_k^n is generally decreased as the (s_k^n, a_k^n) pair is explored more often. However, α_k^n should remain sufficiently large and constant to take into account the latest changes in the highly dynamic environment that may be caused by, for example, the movement of SUs.

User Selection for Reliable Cooperation: The selection of cooperating users ensures that reliable users can be constantly selected to participate in cooperative sensing and contribute correct local decisions to improve detection performance while the unreliable ones are excluded. Let $p(\mathbf{u}_i)$ be the distribution of local decisions reported from user SU_i and $p(\mathbf{u}_0)$ be the distribution of cooperative decisions at the FC. These are SUs' and FC's estimates of

PU activity P_{on} and P_{off}. After receiving SU_i's report u_i and having the cooperative decision u_0, the FC will update $p(\mathbf{u}_i)$ and $p(\mathbf{u}_0)$, respectively. Based on the assumption that cooperative decisions are more accurate than local decisions statistically, we use the *Kullback-Leibler (KL) distance*, $D(p(\mathbf{u}_0)\|p(\mathbf{u}_i))$, to measure how far the distribution of local decisions $p(\mathbf{u}_i)$ diverge from the distribution of cooperative decisions $p(\mathbf{u}_0)$ at the FC. Since $D(p(\mathbf{u}_0)\|p(\mathbf{u}_i)) \geq 0$ and is zero when $p(\mathbf{u}_0) = p(\mathbf{u}_i)$), larger DL distance indicates that the degree of the divergence is higher and implies that SU_i is more unreliable. To determine the user reliability, we compare the KL distance with a threshold δ_{DL} and claim that a user SU_i is considered unreliable for cooperative sensing if

$$D(p(\mathbf{u}_0)\|p(\mathbf{u}_i)) = \sum_{u_i \in \{0,1\}} p(u_i) \log \frac{p(u_i)}{p(u_f)} > \delta_{DL} \tag{13}$$

and reliable otherwise. The threshold δ_{DL} is set to $\mu_\delta + c\sigma_\delta$, where μ_δ and σ_δ are the mean and the standard deviation, respectively, of the DL distances of all reliable cooperating SUs, and c is a constant. Let $\mathcal{U} \subseteq \mathcal{C}$ be the set of uncorrelated SUs selected from the first user selection step. The set of SUs selected for data fusion $\mathcal{D}$ is the subset of $\mathcal{U}$ given by

$$\mathcal{D} = \mathcal{U} \cap \{\cup_i \{SU_i \mid D(p(\mathbf{u}_i)\|p(\mathbf{u}_f)) \leq \delta_{DL}, \forall i\}\}. \tag{14}$$

Thus, when a previously unreliable SU becomes reliable, for example, owing to better CCC conditions, and satisfies the condition in (14), it will be included in $\mathcal{D}$ for cooperation.

Control Channel Fading: As indicated in Sect. 2, FSMC [28] is used to model Rayleigh fading in the control channel for reporting local sensing decisions. In Rayleigh fading, the received SNR γ is exponentially distributed with distribution $f_\Gamma(\gamma) = \frac{1}{\bar{\gamma}} e^{-\gamma/\bar{\gamma}}$, where $\bar{\gamma}$ is the average received SNR. The probability of received SNR γ stays in the SNR region $[A_j, A_{j+1})$ is the probability of staying in channel state x_j, which is given by $p_j = \int_{A_j}^{A_{j+1}} f_\Gamma(\gamma) \mathrm{d}\gamma = e^{-\frac{A_j}{\bar{\gamma}}} - e^{-\frac{A_{j+1}}{\bar{\gamma}}}$. Since the bit error rate of BPSK modulation in additive white Gaussian noise is $Q(\sqrt{2\gamma})$, the crossover probability of the BSC channel for state x_j is given by [28]

$$\varepsilon_j = \frac{\int_{A_j}^{A_{j+1}} f_\Gamma(\gamma) Q(\sqrt{2\gamma}) \mathrm{d}\gamma}{\int_{A_j}^{A_{j+1}} f_\Gamma(\gamma) \mathrm{d}\gamma} = \frac{\gamma_j - \gamma_{j+1}}{p_j} \tag{15}$$

where $\gamma_j = e^{-\frac{A_j}{\bar{\gamma}}} Q(\sqrt{2A_j}) + \bar{\gamma}_c \left(1 - Q\left(\frac{\sqrt{2A_j}}{\bar{\gamma}_c}\right)\right)$ and $\bar{\gamma}_c = \sqrt{\frac{\bar{\gamma}}{\bar{\gamma}+1}}$. Hence, the average error probability is

$$P_e = \sum_{j=0}^{J} p_j \varepsilon_j = \gamma_0 - \gamma_\infty = \frac{1}{2}(1 - \bar{\gamma}_c), \tag{16}$$

where the second equality is obtained by canceling intermediate γ_j terms and the last equality is obtained by using $A_0 = 0$ and $A_\infty = \infty$.

Let $\mathcal{D}^n$ be the set of selected SUs for data fusion in episode n and $K^n = |\mathcal{D}^n|$. Let also the false alarm probability of SU_i be P_f^i, the detection probability P_d^i, and the average CCC reporting error probability P_e^i in (16). For SU_i to report a false alarm in CCC fading, there are two possibilities: 1) a false alarm ($u_i = 1$) is reported and received at FC with no error, and 2) correct local decision $u_i = 0$ is reported and received at FC in error ($u_i = 1$) due to CCC fading. The false alarm probability for local decisions reported by SU_i via fading CCC and perceived by FC is then $P_f^i(1 - P_e^i) + (1 - P_f^i) P_e^i$. Similarly, we can find the probability of SU_i with correct local decisions under H_0 received at FC as $(1 - P_f^i)(1 - P_e^i) + P_f^i P_e^i$. As a result, the false alarm probability for the cooperative decision in episode n is given by

$$Q_f^n = \sum_{\ell=\lceil K^n/2 \rceil}^{K^n} \sum_{\substack{i=1 \\ \mathcal{L}_i \cup \mathcal{M}_i = \mathcal{D}}}^{\binom{K^n}{\ell}} \left(\prod_{\substack{l \in \mathcal{L}_i \\ |\mathcal{L}_i| = \ell}} P_{fe}^l \right) \cdot \left(\prod_{\substack{m \in \mathcal{M}_i \\ |\mathcal{M}_i| = K - \ell}} P_{fe}^m \right) \tag{17}$$

where $P_{fe}^l = P_f^l (1 - P_e^l) + (1 - P_f^l) P_e^l$ and $P_{fe}^m = (1 - P_f^m)(1 - P_e^m) + P_f^m P_e^m$ are the probability of SU_l's error reporting and SU_m's correct reporting, respectively, perceived at the FC, $\mathcal{L}_i$ is the set of the ℓ selected SUs with received false alarms in the "ℓ out of K^n" data fusion rule from the ith combination of all $\binom{K^n}{\ell}$ combinations, and $\mathcal{M}_i = \{\mathcal{D} \backslash \mathcal{L}_i\}$ is the set of the rest of $K^n - \ell$ SUs with correctly received local decisions under H_0 from the ith combination. For the majority decision rule, ℓ ranges from $\lceil K^n/2 \rceil$ to K^n. Similarly, the detection probability for the cooperative decision in episode n, Q_d^n, can be obtained by replacing P_f^l with P_d^l in (17).

RLCS Algorithm: The RLCS algorithm is listed in Algorithm 1. The algorithm takes the array of all state-action values ($Q(|\mathcal{S}|, |\mathcal{A}|)$) as the input and initializes the array entries to zero. The output is the optimal solution: optimal number of cooperating SUs ($K^* = |\mathcal{P}|^*$), optimal reporting sequence (π^*), and total reporting delay (T_d^*). In the algorithm, RLCS is performed repeatedly (lines 3 to 19) unless the optimal stopping criterion is met. In each episode, there are $K^n \leq L$ SUs selected for sensing and reporting based on the action strategy. In each state of an episode (lines 6–14), the FC sends the request to the selected SU, receives the local decision, calculates the reward, and updates the Q-value using (12). At the end of the episode, the FC terminates the episode with $a_k^n = 0$, determines the reliable set of SUs and the cooperative sensing decision, and broadcasts it to all neighbors (lines 16–18).

Algorithm 1 : RL-based Cooperative Sensing (RLCS)

Input: $Q(|\mathcal{S}|,|\mathcal{A}|), L, N$
Output: $\pi^* \leftarrow \pi, \mathcal{P}^* \leftarrow \mathcal{D}^n, K^* \leftarrow |\mathcal{P}^*|, T_d^* \leftarrow T_d$
repeat
  Initialize: $\mathcal{D}^n = \mathcal{U}^n = \emptyset, K^n = 0, T_d = 0$
  for $k \leftarrow 0$ to $(L+1)$ **do**
    $a_k^n \leftarrow ActionStrategy(\mu_k^n, s_k^n, Q, \tau^n)$
    if $a_k^n \neq 0$ **then**
      $SendReqToNeighbor(SU_i; a_k^n = i)$
      $r_{k+1}^n \leftarrow WaitForReward(u_i; t_d^n(s_k, a_k = i))$
      $Q_k^n \leftarrow (1-\alpha_k^n)Q_k^n + \alpha_k^n\left[r_{k+1}^n + \gamma f(Q_{k+1}^n)\right]$
      $Update(\pi, \alpha_k^n, \bar{t}_{d_i}, T_d, K^n, Y^n, \mathcal{U}^n)$
    else
      break
    end if
  end for
  $\mathcal{D}^n = \mathcal{U}^n \cap \{\cup_i \{SU_i \mid D(p(\mathbf{u}_i)\|p(\mathbf{u}_f)) \leq \delta_{DL}, \forall i\}\}$
  $u_0 \leftarrow CoopSensingDecision(u_i, k, L, \forall i \in \mathcal{D}^n)$
  $BroadcastCoopDecision(u_0; a_k = 0)$
until $Y^n \geq \tilde{Y}^*$

4 Performance analysis

In this section, we analyze the performance of the proposed RLCS scheme by first showing the optimal solution of RLCS, proving its convergence, and evaluating the rate of convergence. We then provide the optimal stopping alternative for performance improvement in expected cumulative rewards.

4.1 Optimal solution of RLCS algorithm

The RLCS algorithm is capable of learning the changes in dynamic environment to reach the optimal solution. The optimal solution of RLCS is the optimal set of spatially uncorrelated SUs, $\mathcal{P}^* \subseteq \mathcal{C}$, selected for cooperation in sequence by optimal policy $\pi^* = (\mu_0, \ldots, \mu_{K^*})$ that achieves maximum cumulative reward R_{π^*}, where $K^* = |\mathcal{P}^*|$ is the optimal number of selected SUs. In the following, we present the necessary conditions of achieving the optimal solution in static environment, where SU locations and their reporting delays are known, as two lemmas followed by the corresponding theorem.

Lemma 1 *Given the set of L cooperating SUs, $\mathcal{C}$, with their locations, the optimal number of selected SUs, $K \leq L$, is the maximum number of spatially uncorrelated SUs that maximize $R_{\rho_L} = \sum_{k=0}^{L-1} r_{\rho_{k+1}}$ with maximum value $R_{\rho_K} = 0$.*

Proof Let $\mathcal{P}_k$ be the set of selected SUs from s_0 to s_k and $SU_j, j \in \{\mathcal{C} \setminus \mathcal{P}_{k-1}\}$, be the SU selected in s_k. From (6) and (7), $r_{\rho_{k+1}} = -C_{\rho_{k+1}}$ and $C_{\rho_{k+1}} \geq 0$. The maximum value of $r_{\rho_{k+1}}$ is 0 and can be obtained if and only if $C_{\rho_{k+1}} = 0$. By (7), $C_{\rho_{k+1}} = \frac{1}{k}\sum_{\ell=0}^{k-1} |\rho_{ij}(s_\ell, a_\ell = i)| = 0, \forall i \in \mathcal{P}_{k-1}$. All selected SUs in $\mathcal{P}_k$ must be spatially uncorrelated to maximize $r_{\rho_{k+1}}$. If SU_j is spatially correlated to any SU in $\mathcal{P}_{k-1}, r_{\rho_{k+1}} < 0$. Thus, $r_{\rho_{k+1}} = 0$ for $k = 0, \ldots, K-1$ for selecting up to maximum $K \leq L$ spatially uncorrelated SUs in $\mathcal{C}$ to maximize R_{ρ_L} with maximum value $R_{\rho_K} = 0$ as $r_{\rho_{k+1}} < 0, k = K, \ldots, L-1$, such that $R_{\rho_k} < R_{\rho_K}, k = K+1, \ldots, L$. □

Lemma 2 *Given the set of K selected SUs, $\mathcal{P}$, with their reporting delays $t_{d_i}, i \in \mathcal{P}$, there exists an optimal user selection sequence $\pi_{K^*}^* = (a_0^*, \ldots, a_{K^*-1}^*)$ that maximizes $R_{d_K} = \sum_{k=0}^{K-1} r_{d_{k+1}}$, where $a_k^* = \arg\min_{a_k \in \mathcal{A}_{s_k}} t_d(s_k, a_k)$ and $K^* \leq K$ is the maximum number of selected SUs that satisfy total reporting delay constraint $T_{d_{K^*}} = \sum_{k=0}^{K^*-1} t_d(s_k, a_k^*) \leq T_{lim}$.*

Proof Let $d_{k,i} = t_{d_i}(s_k, a_k), i \in \mathcal{A}_{s_k}$, be the reporting delay of SU_i selected in s_k. From (6) and (8),

$$r_{d_{k+1}} = 1 - C_{d_{k+1}} = 1 - \frac{\sum_{\ell=0}^{k} d_{\ell,i}}{T_{lim}} = \left(1 - \frac{\sum_{\ell=0}^{k-1} d_{\ell,i}}{T_{lim}}\right) - \frac{d_{k,i}}{T_{lim}} = r_{d_k} - \frac{d_{k,i}}{T_{lim}}. \quad (18)$$

If $T_{d_{k+1}} = \sum_{\ell=0}^{k} d_{\ell,i} \leq T_{lim}$, we have $0 < C_{d_{k+1}} \leq 1$ and $r_{d_{k+1}} \geq 0$. Since $d_{k,i}$, $T_{lim} > 0$ and T_{lim} is constant, maximizing $r_{d_{k+1}}$ is equivalent to minimizing $d_{k,i}$ given r_{d_k} obtained in the previous state. As a result, the optimal user selection that maximizes $r_{d_{k+1}}$ in state s_k is $a_k^* = \arg\min_{i \in \mathcal{A}_{s_k}} d_{k,i}$, $k = 0, \ldots, K^* - 1$. Since K^* is the maximum number of SUs that satisfy $T_{d_{K^*}} \leq T_{lim}$, we have $T_{d_{k+1}} > T_{lim}$ and $r_{d_{k+1}} < 0$ for $k = K^*, \ldots, K-1$, due to $C_{d_{k+1}} > 1$ from (8). Thus, there exists an optimal user selection sequence $\pi_{K^*}^* = (a_0^*, \ldots, a_{K^*-1}^*)$, where $a_k^* = \arg\min_{i \in \mathcal{A}_{s_k}} d_{k,i}$, that maximizes R_{d_K} and achieves maximum value $R_{d_{K^*}} = \sum_{k=0}^{K^*-1} r_{d_{k+1}}$. □

Theorem 1 (Optimal Solution of RLCS Algorithm) *The optimal solution of RLCS is a 3-tuple* $\langle K^*, \mathcal{P}^*, T_d^* \rangle$ *obtained by the optimal policy* $\pi^* = \{\mu_0^*, \ldots, \mu_{K^*-1}^*\}$ *that achieves the maximum cumulative reward given by*

$$R_{\pi^*} = K^* - \frac{\sum_{k=0}^{K^*-1} (K^* - k) \cdot t_d(s_k, \mu_k^*)}{T_{lim}} \tag{19}$$

and total reporting delay $T_d^* = \sum_{k=0}^{K^*-1} t_d(s_k, \mu_k^*) < T_{lim}$, *where* $\mu_k^* = \arg\min_{a_k \in \mathcal{A}_{s_k}} t_d(s_k, a_k)$.

Proof In (11), r_{k+1} received in s_k can be negative if $C_{\rho_{k+1}} > 0$ in (7) or $C_{d_{k+1}} > 1$ in (8). That is, the SU selected in s_k either causes spatial correlation with any previously selected SUs or incurs reporting delay that does not satisfy the total reporting delay constraint T_{lim}. To avoid negative rewards, $C_{\rho_{k+1}} = 0$ and $0 < C_{d_{k+1}} \leq 1$ are required. Using Lemma 1, we obtain maximum K spatially uncorrelated SUs to ensure $r_{\rho_{k+1}} = C_{\rho_{k+1}} = 0$. By plugging (6) into (11) and setting $C_{\rho_{k+1}} = 0, R_\pi$ is reduced to $R_\pi = \sum_{k=0}^{K-1} r_{d_{k+1}} = R_{d_K}$. Using Lemma 2, we obtain optimal user selection sequence $\pi_{K^*}^*$ to maximize R_π and ensure $0 < C_{d_{k+1}} \leq 1$ with maximum K^* selected SUs that satisfy $T_{d_{K^*}} < T_{lim}$. Since the decision rules μ_k are deterministic, $\mu_k^* = a_k^* = \arg\min_{a_k \in \mathcal{A}_{s_k}} t_d(s_k, a_k)$ in $\pi^* = \pi_{K^*}^*$ with $T_d^* = T_{d_{K^*}}$ to achieve

$$\begin{aligned} R_{\pi^*} = R_{d_{K^*}}^* &= \sum_{k=0}^{K^*-1} r_{d_{k+1}}^* = K^* - \sum_{k=0}^{K^*-1} C_{d_{k+1}}^* \\ &= K^* - \frac{1}{T_{lim}} \sum_{k=0}^{K^*-1} \sum_{\ell=0}^{k} t_d(s_\ell, \mu_\ell^*). \end{aligned} \tag{20}$$

After some algebraic manipulation, (19) follows. □

4.2 Convergence of RLCS algorithm

The optimal solution is achieved when the RLCS algorithm converges with sufficient exploration of state-action pairs. In the case of insufficient exploration, a suboptimal solution may be obtained upon convergence. To prove the convergence, we first show that the sequence of expected cumulative rewards $\{R^n\}$ is a submartingale in Lemma 3, and the result follows the Martingale convergence theorem given in Lemma 4 whose proof can be found in [31] and is omitted here due to limited space.

Lemma 3 *The sequence of expected cumulative rewards* $R^n, n = 1, 2, \ldots$ *is a submartingale that satisfies* $E|R^n| < \infty$ *and* $E[R^{n+1} \mid R^i] \geq R^n, \forall\, i \leq n$.

Proof From (11), we can easily show that $E|R^n| < \infty, n = 1, 2, \ldots$ since L and r_{k+1}^n are finite. Next, we show $E[R^{n+1} \mid R^i] \geq R^n, \forall\, i \leq n$. Let Π be the set of all policies.

$$E[R^{n+1}|R^i] = \sum_{\pi \in \Pi} R_\pi^{n+1} p(R_\pi^{n+1}|R^i) \tag{21}$$

$$= \sum_{\pi \in \Pi} \sum_{k=0}^{L} E\left[r_{k+1}^{n+1}(s_k^{n+1}, \mu_k^{n+1})\right] p(R_\pi^{n+1}|R^i) \tag{22}$$

$$= \sum_{\pi \in \Pi} \sum_{k=0}^{L} \sum_{s_k^{n+1}, \mu_k^{n+1}} r_{k+1}^{n+1}(s_k^{n+1}, \mu_k^{n+1}) \cdot p_r(s_k^{n+1}, \mu_k^{n+1}) p(R_\pi^{n+1}|R^i) \tag{23}$$

$$= \sum_{\pi \in \Pi} \sum_{k=0}^{L} \sum_{s_k^{n+1}, a_k^{n+1}} r_{k+1}^{n+1}(s_k^{n+1}, a_k^{n+1}) \cdot p_\pi(s_{k+1}^{n+1}|s_k^{n+1}, a_k^{n+1}) \tag{24}$$

where R_π^{n+1} in (21) is the expected cumulative reward obtained by π, p_r in (23) is reward distribution for π, and p_π in (24) is the transition probability for π in episode $n+1$ given R_i, equivalently, all Q-value updates with $r_{k+1}^i, k = 0, \ldots, L, i \leq n$, in previous n episodes. Since, as indicated in Sect. 3, the state transition probability is the action selection probability and let $s_k^{n+1} = x$ and $a_k^{n+1} = s_{k+1}^{n+1} = y$, from (9), we have

$$p_{xy}^{n+1} = \frac{e^{Q^{n+1}(s_k = x, a_k = y)/\tau^{n+1}}}{\sum_{j=1}^{|\mathcal{A}_x|} e^{Q^{n+1}(s_k = x, a_k = j)/\tau^{n+1}}}, \quad y \in \mathcal{A}_x. \tag{25}$$

We now compare $r_{k+1}^{n+1}(x,y)\ p_{xy}^{n+1}$ with $r_{k+1}^n(x,y)\ p_{xy}^n$ for a policy π. For simplicity, we first assume no future estimates ($\gamma = 0$). Since the Q-value for each (x, y) is updated no more than once in one episode, Q^{n+1} before the update in episode $n+1$ equals Q^n after the update (12) in episode n. Hence, (12) simplifies to

$$Q_k^{n+1} = (1 - \alpha) Q_k^n + \alpha r_{k+1}^n = Q_k^n + \alpha(r_{k+1}^n - Q_k^n). \tag{26}$$

If the delays and locations of SUs are fixed, the reward of the same state-action pair (x, y) in a policy π is the same for different episodes: $r_{k+1}^{n+1}(x, y) = r_{k+1}^n(x, y)$. Moreover, p_{xy}^{n+1} is the function of $\frac{Q^{n+1}}{\tau^{n+1}}$ and we know from (10) that $\tau^{n+1} < \tau^n$. As a result, depending on how Q-values change from episode n to $n+1$, we have the following six cases:

(i) $Q^{n+1} \geq Q^n \geq 0$: In this case, $p_{xy}^{n+1} > p_{xy}^n$ because $\frac{Q^{n+1}}{\tau^{n+1}} > \frac{Q^n}{\tau^n}$. From (26), $Q_k^{n+1} - Q_k^n = \alpha(r_{k+1}^n - Q_k^n) \geq 0$. We have $r_{k+1}^{n+1} = r_{k+1}^n \geq Q_k^n \geq 0$. Thus, $r_{k+1}^{n+1} p_{xy}^{n+1} \geq r_{k+1}^n p_{xy}^n$.

(ii) $Q^{n+1} \geq 0 \geq Q^n$: As in case (i), $p_{xy}^{n+1} > p_{xy}^n$. Using (26) and $Q_k^n \leq 0$, we obtain $r_{k+1}^{n+1} = r_{k+1}^n \geq (1 - \frac{1}{\alpha}) Q_k^n \geq 0$. Thus, $r_{k+1}^{n+1} p_{xy}^{n+1} \geq r_{k+1}^n\ p_{xy}^n$.

(iii) $0 \geq Q^{n+1} \geq Q^n$: Similarly, $p_{xy}^{n+1} > p_{xy}^n$. In this case, $r_{k+1}^n \in [Q_k^n, (1 - \frac{1}{\alpha}) Q_k^n]$. Thus, $\Pr(r_{k+1}^{n+1} p_{xy}^{n+1} \geq r_{k+1}^n\ p_{xy}^n) < \Pr(r_{k+1}^{n+1}\ p_{xy}^{n+1} < r_{k+1}^n\ p_{xy}^n)$.

(iv) $Q^{n+1} \le Q^n \le 0$: Since $\tau^{n+1} < \tau^n$ and $\frac{Q^{n+1}}{\tau^{n+1}} < \frac{Q^n}{\tau^n}$, we have $p_{xy}^{n+1} < p_{xy}^n$. From (26), $Q_k^{n+1} - Q_k^n = \alpha(r_{k+1}^n - Q_k^n) \le 0$. We obtain $r_{k+1}^{n+1} = r_{k+1}^n \le Q_k^n \le 0$. Thus, $r_{k+1}^{n+1} p_{xy}^{n+1} \ge r_{k+1}^n p_{xy}^n$.

(v) $Q^{n+1} \le 0 \le Q^n$: Similar to (iv), $p_{xy}^{n+1} < p_{xy}^n$. Using (12) and $Q_k^n \ge 0$, we obtain $r_{k+1}^{n+1} = r_{k+1}^n \le (1 - \frac{1}{\alpha}) Q_k^n \le 0$. Thus, $r_{k+1}^{n+1} p_{xy}^{n+1} \ge r_{k+1}^n p_{xy}^n$.

(vi) $0 \le Q^{n+1} \le Q^n$: Similarly, we obtain $p_{xy}^{n+1} < p_{xy}^n$ and $r_{k+1}^n \in [(1 - \frac{1}{\alpha}) Q_k^n, Q_k^n]$. Since $Q^{n+1}, Q^n \ge 0$, $\Pr(r_{k+1} < 0) < \Pr(r_{k+1} \ge 0)$. Thus, $\Pr(r_{k+1}^{n+1} p_{xy}^{n+1} \ge r_{k+1}^n p_{xy}^n) < \Pr(r_{k+1}^{n+1} p_{xy}^{n+1} < r_{k+1}^n p_{xy}^n)$.

If $\gamma > 0$, one may replace r_{k+1}^n in (26) with $r_{k+1}^n + \gamma f(Q_{k+1}^n)$ and the analysis above still applies with variations of future estimates. Since cases (i)–(vi) are applicable to all (x, y) pairs in any policy π and equally likely, we conclude that

$$E[R^{n+1}|R^i] = \sum_{\pi \in \Pi} \sum_{k=0}^{L} \sum_{x,y} r_{k+1}^{n+1}(x,y) p_{xy}^{n+1} \tag{27}$$

$$\ge \sum_{\pi \in \Pi} \sum_{k=0}^{L} \sum_{x,y} r_{k+1}^{n}(x,y) p_{xy}^{n} \tag{28}$$

$$= E[R^n|R^i] = R^n, \quad \forall i \le n, \tag{29}$$

and $\{R^n\}$ is a submartingale. □

Lemma 4 (Martingale Convergence Theorem) *(Theorem 5.14 in [31]) Let $R^1, R^2, \ldots$ be a submartingale such that $\sup E|R^n| < \infty$, then there exists a random variable (r.v.) R such that $R^n \to R$ almost surely (a.s.) and $E|R| < \infty$.*

Based on Lemma 3 and Lemma 4, we present the convergence theorem of RLCS.

Theorem 2 (Convergence of RLCS Algorithm) *The sequence of expected cumulative rewards $R^n, n = 1, 2, \ldots$ converges to a value R almost surely (a.s.).*

Proof Since $\{R^n\}$ is a submartingale (Lemma 3), by following Lemma 4, there exists an r.v. R such that $R^n \to R$ a.s. and $E|R| < \infty$ due to $E|R^n| < \infty$. The convergence of RLCS follows.

From (24), R^n increases with p_{xy}^n according to (25). If we set $a = Q^n/\tau^n$ such that $p_i = \frac{e^{a_i}}{\sum_j e^{a_j}}$, we obtain that $\frac{\partial p_i}{\partial a_i} = p_i - p_i^2$ and $\frac{\partial^2 p_i}{\partial a_i^2} = (p_i - p_i^2)(1 - 2p_i)$. p_i is convex if $p_i \in [0, 0.5]$ and concave if $p_i \in [0.5, 1]$. As a result, when τ^n is large at the beginning of learning where exploration takes place, all possible actions i in that state are equally likely and $p_i \in [0, 0.5]$. R^n is convex in this region. On the opposite, when τ^n is close to zero at the end of learning where exploitation takes effect, $p_i = 1$ for the best action i in all states. R^n in this region is concave. The region in between where the transition from exploration to exploitation occurs is, thus, linear. Based on this observation, we show the rate of convergence in the following theorem. □

Theorem 3 (Rate of Convergence) *The sequence of expected cumulative rewards R^n in RLCS converges sublinearly.*

Proof Let R^1 be the initial R^n and $R^n = R^*$ at episode $n = N$. The increasing rate of R^n in the linear region between exploration and exploitation can be approximated as $\frac{R^* - R^1}{N-1}$. From $R^n = \frac{R^* - R^1}{N-1}(n-1) + R^1$, we obtain $R^{n+1} - R^* = \frac{n-N+1}{N-1}(R^* - R^1)$ and $R^n - R^* = \frac{n-N}{N-1}(R^* - R^1)$. Using the rate of convergence $\hat{K}$ defined in [32], we have

$$\hat{K} = \limsup_{n \to \infty} \frac{\|R^{n+1} - R^*\|}{\|R^n - R^*\|^{\zeta}} = \limsup_{n \to \infty} \frac{\|n - N + 1\|}{\|n - N\|} = 1 \tag{30}$$

where ζ is the order of convergence and $\zeta = 1$ indicates the convergence of the first order. The same result can be obtained by using (19). Since R^* is upper bounded by K^* in (19) and $N \gg K^*$ for sufficiently large N, $\Delta R = R^{n+1} - R^n = \frac{R^* - R^1}{N-1} \le \frac{K^* - R^1}{N-1} \approx 0$. Thus, $\hat{K} = 1$ is obtained in (30) with $R^{n+1} \approx R^n$. $\{R^n\}$ is said to converge sublinearly. □

4.3 Optimal stopping time

In Sect. 3, the RLCS algorithm is introduced to find the optimal solution. However, the number of episodes needed to reach the optimal solution may be large due to the exploration of all state-action pairs in state and action spaces. In this section, we aim to find the optimal stopping time T^* to reduce the total learning time by formulating the problem as Markov optimal stopping with finite horizon N.

Let $\mathcal{F}^n$, $n = 1, \ldots, N$ be a nondecreasing sequence of sub-σ-algebras of event class $\mathcal{F}$ called a filtration. Consider a set of stopping times $\mathcal{T}_n^N = \{T \in \mathcal{T} | n \le T \le N\}, 1 \le n \le N$, where $\mathcal{T}$ is the set of all stopping time. Thus, with cumulative reward sequence $\{Y^n\}$, the optimal stopping problem is to find the stopping time T such that the expected cumulative reward is maximized:

$$\sup_{T \in \mathcal{T}} E[Y^T] \tag{31}$$

According to the optimal stopping theory [33, 34], the solution to (31) can be obtained by backward induction as defined by a sequence of random variables:

$$S_n^N = \max\{Y^n, E[S_{n+1}^N | \mathcal{F}^n]\}, \quad n = N-1, \ldots, 1 \tag{32}$$

with $S_N^N = Y^N$. Thus, we stop at $n = T$ if $Y^T \ge E[S_{T+1}^N | \mathcal{F}^T]$ and continue otherwise. The stopping time is given by

$$T_n^N = \inf\{n \le T \le N | S_T^N = Y^T\}, \quad 1 \le n \le N. \tag{33}$$

However, unlike $Y^n, E[S_{n+1}^N|\mathcal{F}^n]$ is difficult to obtain in each episode n. This is because the probability distributions of the r.v.'s in the sequence S_{n+1}^N is unknown in episode n and may be significantly changed owing to the action selections in future episodes $n+1, \ldots, N$. Thus, we use the optimal reward estimate $\tilde{Y}^*$ as the alternative to $E[S_{n+1}^N|\mathcal{F}^n]$. $\tilde{Y}^*$ is the best-known cumulative reward estimate in episode n based on the optimal set of uncorrelated SUs, $\tilde{\mathcal{K}}^*$, obtained from the sequence of sorted average reporting delay, $\mathcal{T}_d^n = \{\bar{t}_{d_i}^n\}$ in ascending order. By using (19), we obtain the following:

$$\tilde{Y}^* = \tilde{K}^* - \frac{\sum_{k=0}^{\tilde{K}^*-1}(\tilde{K}^* - k)\bar{t}_{d_i}^n(k)}{T_{lim}}, i \in \tilde{\mathcal{K}}^* \tag{34}$$

where $\tilde{K}^* = |\mathcal{K}^*|$ and $\bar{t}_{d_i}^n(k) \in \mathcal{T}_d^n$ is indexed for the kth uncorrelated SU in $\tilde{\mathcal{K}}^*$. Thus, (32) reduces to $S_n^N = \max\{Y^n, \tilde{Y}^*\}$ and the optimal stopping time T^* in the RLCS algorithm is the smallest episode number $n = T^*$ whose cumulative reward Y^n is greater than or equal to the optimal reward estimate $\tilde{Y}^*$:

$$T^* = \inf\{n \le N \mid Y_n \ge \tilde{Y}^*\}. \tag{35}$$

In other words, the optimal stopping occurs if the cumulative reward of episode n, Y^n, is greater than or equal to the current best known reward estimate, $\tilde{Y}^*$, or the RLCS algorithm continues to find the optimal solution.

5 Performance evaluation

In this section, we evaluate the performance of our proposed RLCS scheme by showing the convergence the RLCS algorithm, the improvement of detection probability, and the adaptability to environmental changes.

We consider an SU (the FC) and its 9 neighbors deployed in a 600 m × 600 m square area placed in the first quadrant of the Cartesian coordinate system. A PU with $r_b = 0.3$ and $r_d = 0.2$ at 900 MHz is located at (0, 0). The FC is located at (500, 500) and its neighbors are located within 40 m of FC's location. For channel model and sensing parameters, we set $\gamma_{pl} = 3.1$ for path loss, noise uncertainty $\sigma_0 = 6$ dB and lognormal dB-spread $\sigma_1 = 6$ dB in (1), decaying coefficient $a = 0.1204$ in (2) for urban settings, and local detection threshold $\lambda_{th} = 0.2$ dB for all SUs. The FSMC model for independent Rayleigh fading CCC consists of 1,024 SNR regions in which the range of each SNR region is of 0.1 dB. We set the SU transmit power to 20 mW with path loss exponent $\gamma_{pl} = 4.1$. For data fusion at the FC, hard combinations of local decisions with the majority rule are used. With these settings, the SUs are approximately located at the boundary of the protected region of the PU. At this border location, the received power is close to the noise floor set to −101 dBm. Thus, cooperative sensing is essential for SUs to improve their detection performance. A test scenario can be found in [1].

5.1 Convergence of RLCS algorithm

Figure 3 shows the expected cumulative rewards with $N = 1{,}000$, $\tau_0 = 1$, and $\tau_N = 0.01$, for Q-learning, Sarsa, and Action-Critic TD learning strategy over 1,000 runs. The sublinear convergence of RLCS (both with and without optimal stopping) is evident. The maximum value R^* is obtained by (19). For both cases, Q-learning converges to the optimal value while the other two settle for a suboptimal value due to more exploitation than exploration in early stages ($n < 500$). All three methods show significant improvement in R_n with optimal stopping.

5.2 Detection performance

Figure 4 shows the improvement of detection performance (both Q_d and Q_f) under Poisson and bursty PU traffic during the RLCS process. Q_d and Q_f are averaged over the most recent 500 cooperative decisions at the FC. The detection performance of full cooperation (FCS) case is approximately the same for all episodes. It is evident that Q_d of RLCS is gradually improved and reaches above 0.9 after 3,000 episodes, significant improvement over FCS. The initial large Q_f of RLCS is attributed to non-optimal SU selections at the beginning of learning during the exploration phase. However, Q_f is constantly decreasing

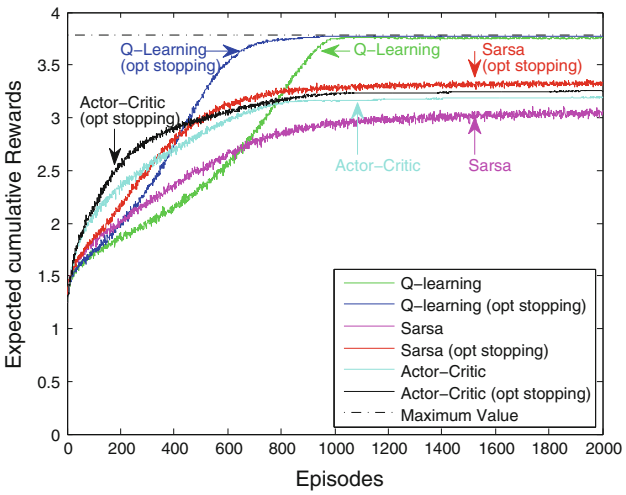

Fig. 3 Expected cumulative rewards of RL-based cooperative sensing

and considerably reduced to 0.025 to be comparable to Q_f of FCS at the end. Thus, with RLCS, the detection performance improves as soon as the learning from the environment takes effect. Figure 5 shows the receiver operating characteristic (ROC) curves of FCS and RLCS in correlated shadowing and possible user movement and CCC fading. We see that the cooperative gain achieved by FCS in independent shadowing is compromised by correlated shadowing with Q_d dropping from 0.97 to 0.85 at $Q_f = 0.1$. The detection performance of full cooperation in independent shadowing is attainable only when all cooperating SUs are uncorrelated. However, with RLCS, Q_d is increased to 0.91. Hence, RLCS scheme is effective in combating correlated shadowing.

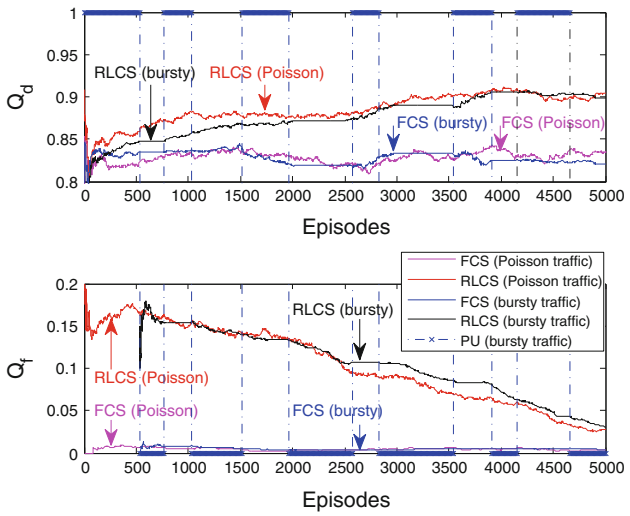

Fig. 4 Improvement of Q_d/Q_f during RLCS and adaptability to random and bursty PU traffic

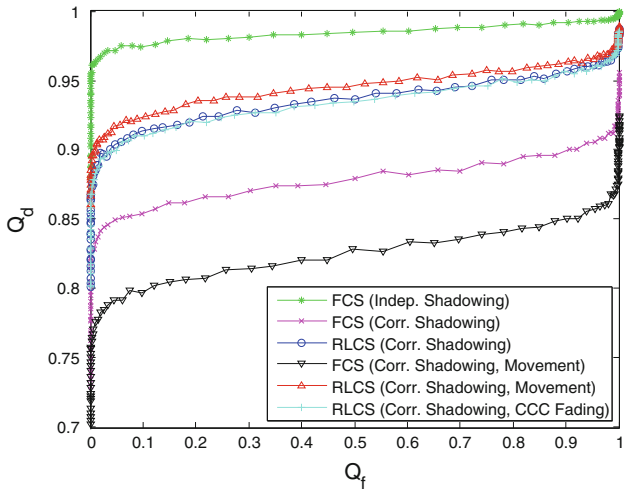

Fig. 5 ROC of FCS and RLCS in correlated shadowing with possible user movement and fading control channel

5.3 Adaptability to environmental change

With the learning capability, the RLCS algorithm is able to adapt to changes in the environment. In this subsection, we evaluate the adaptability of RLCS based on the changes of PU activity, user location, user reliability, and fading control channel.

PU Activity Changes: Figure 4 shows adaptability to PU activity changes for different PU traffic types in addition to constant improvement in detection performance. To manifest the effect of PU activity changes on detection performance, we generate bursty PU traffic by staying in either PU state with high probability for a period that spans a random number of episodes, and toggling the ON/OFF states with low probability. As seen in the figure, Q_d is improved mostly during the period of active PU while the Q_f is primarily improved during the period of no PU activity. For this reason, $Q_f = 0$ during the first 500 episodes. Thus, RLCS is adaptable to PU activity and consistently improves detection performance for arbitrary PU traffic patterns.

User Movement: In this test scenario, an SU with independent observations at original location moves to a new location (distance 33.12 m) at the pedestrian speed 1.25 m/s. Since it takes 26.5s to reach the destination, the movement spans 133 RLCS episodes for cooperative sensing period of 200 ms. Since this movement incurs correlation with other SUs, the optimal solution is changed accordingly when the algorithm converges. Figure 5 shows the ROC curve before and after the movement for both FCS and RLCS. For $Q_f = 0.1$, Q_d of FCS drops from 0.85 to below 0.8 after the movement while Q_d of RLCS is even slightly improved after the movement due to the selection of all uncorrelated SUs. This shows the capability of RLCS adapting to user movement while maintaining detection performance.

User Reliability: Figure 6 shows the KL distance curve of an unreliable SU, the average KL distance values of all reliable users, and the detection threshold $\delta_{KL} = \mu_\delta + 2\sigma_\delta$ over 5,000 episodes. The low KL values before episode 1,000 indicate that the SU is normally a reliable user. Its KL value is dramatically increased after the user becomes unreliable at episode 1,000. It is detected and removed from the set of cooperation at episode 1,391 when its KL value exceeds the threshold δ_{KL}. When the SU becomes reliable again, its KL distance is gradually reduced. After its KL distance meets the threshold in episode 4,016, the user may be selected by the FC again for cooperation.

Fading Control Channel: We first compare the detection performance obtained by simulations with that obtained by using (17) and its counterpart for Q_d. Local P_d and P_f are set to 0.7633 and 0.1466, respectively, for all SUs. This corresponds to $Q_d = 0.91$ and $Q_f = 0.025$, respectively, with $P_e = 0$. The results are obtained at episode 5000 when the optimal solution is reached, and are

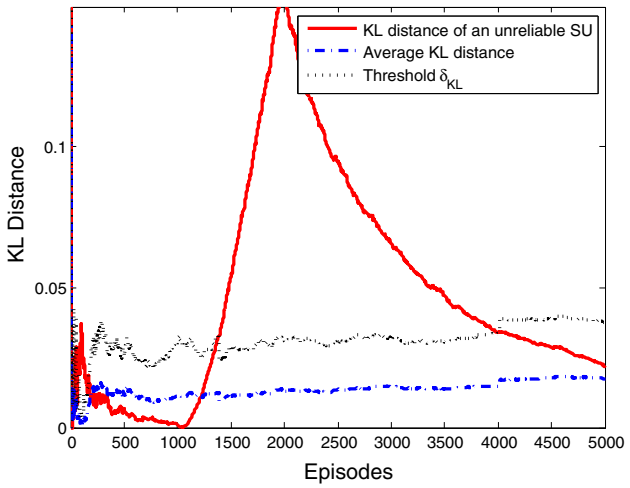

Fig. 6 Average and user KL distance values for detection of unreliable users

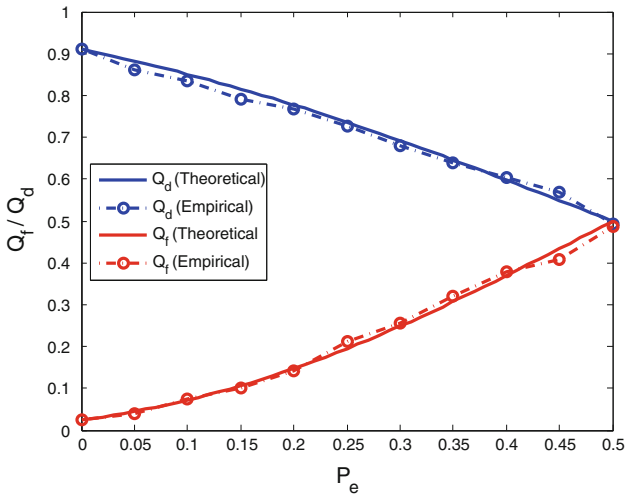

Fig. 7 Theoretical and empirical detection performance (Q_d/Q_f) versus average error probability (P_e) on fading control channel

averaged over 10 runs. Figure 7 shows the theoretical and empirical detection performance versus P_e ranging from 0 to 0.5 on fading control channel. The simulation results follow the theoretical curves closely for both Q_d and Q_f. Figure 5 also shows the ROC curve of RLCS with fading CCC versus perfect CCC. The detection performance of the fading CCC is similar that of the perfect CCC case with slight degradation ($Q_f = 0.0255$, $Q_d = 0.9027$). Thus, RLCS can effectively maintain the detection performance with reporting in fading control channel.

6 Conclusions

In this paper, we propose a novel cooperative sensing model via RL to improve the cooperative gain and mitigate the cooperation overhead in correlated shadowing and dynamic environment. We show that the proposed RLCS method is capable of converging to an optimal solution asymptotically and enhancing rewards by using optimal stopping. The optimal solution achieved by an optimal user selection policy includes finding the optimal set of cooperating neighbors with minimum control traffic, reducing the overall reporting delay, selecting independent users for cooperation under correlated shadowing, and excluding unreliable users and data from cooperation. The results show that RLCS improves or maintains the comparable detection performance while adapting to environmental change, such as the changes in PU traffic patterns, user locations, user reliability, and fading control channel conditions, that may compromise the cooperative gain in cooperative sensing.

Acknowledgments This work was supported by the U.S. National Science Foundation under Award ECCS-0900930.

References

1. Lo, B. F., & Akyildiz, I. F. (2010). Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks. In *Proceedings of IEEE PIMRC*, pp. 2244–2249.
2. Akyildiz, I. F., Lee, W. Y., & Chowdhury, K. R. (2009). CRAHNs: Cognitive radio ad hoc networks. *Ad Hoc Networks Journal (Elsevier), 7*, 810.
3. Akyildiz, I. F., Lee, W. Y., Vuran, M. C., & Mohanty, S. (2006). NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer Networks Journal (Elsevier), 50*, 2127.
4. Akyildiz, I. F., Lo, B. F., & Balakrishnan, R. (2011). Cooperative spectrum sensing in cognitive radio networks: A survey. *Physical Communication (Elseviar) Journal, 4*(1), 40.
5. Cabric, D., Mishra, S. M., & Brodersen, R. W. (2004). Implementation issues in spectrum sensing for cognitive radios. In *Proceedings of 38th Asilomar Conference on Signals, Systems, and Computers*, pp. 772–776.
6. Ghasemi, A., & Sousa, E. S. (2005). Collaborative spectrum sensing for opportunistic access in fading environments. In *Proceedings of IEEE DySPAN*, pp. 131–136.
7. Lo, B. F., Akyildiz, I. F., & Al-Dhelaan, A. M. (2010). Efficient recovery control channel design in cognitive radio ad hoc networks. *IEEE Transactions on Vehicular Technology, 59*(9), 4513.
8. Lo, B. F. (2011). A survey on common control channel design for cognitive radio networks. *Physical Communication (Elseviar) Journal, 4*(1), 26.
9. Chen, R., Park, J. M., & Bian, K. (2008). Robust distributed spectrum sensing in cognitive radio networks. In *Proceedings of IEEE INFOCOM*, pp. 1876 –1884.
10. Varshney, P. K. (1997). *Distributed detection and data fusion*. New York: Springer.
11. Unnikrishnan, J., & Veeravalli, V. V. (2008). Cooperative sensing for primary detection in cognitive radio. *IEEE Journal of Selected Topics in Signal Processing, 2*(1), 18.
12. Selen, Y., Tullberg, H., & Kronander, J. (2008). Sensor selection for cooperative spectrum sensing. In *Proceedings of IEEE DySPAN*, pp. 1–11.

13. Visotsky, E., Kuffner, S., & Peterson, R. (2005). On collaborative detection of TV transmissions in support of dynamic spectrum sharing. In *Proceedings of IEEE DySPAN*, pp. 338–345.
14. Ma, J., Zhao, G., & Li, Y. (2008). Soft combination and detection for cooperative spectrum sensing in cognitive radio networks. *IEEE Transactions on Wireless Communications, 7*(11), 4502.
15. Visser, F. E., Janssen, G. J., & Pawelczak, P. (2008). Multinode spectrum sensing based on energy detection for dynamic spectrum access. In *Proceedings of IEEE VTC-Spring 2008*, pp. 1394–1398.
16. Zhou, X., Ma, J., Li, G., Kwon, Y., & Soong, A. (2010). Probability-based combination for cooperative spectrum sensing. *IEEE Transactions on Communications, 58*(2), 463.
17. Sun, C., Zhang, W., & Letaief, K. (2007). Cooperative spectrum sensing for cognitive radios under bandwidth constraints. In *Proceedings of IEEE WCNC 2007* (pp. 1–5).
18. Zhang, X., Qiu, Z., & Mu, D. (2008). Asynchronous cooperative spectrum sensing in cognitive radio. In *Proceedings of IEEE ICSP*, pp. 2020–2023.
19. Song, C., & Zhang, Q. (2008). Sliding-window algorithm for asynchronous cooperative sensing in wireless cognitive networks. In *Proceedings of IEEE ICC*, pp. 3432–3436.
20. Sutton, R., & Barto, A. (1988). *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press.
21. Berthold, U., Fu, F., van der Schaar, M., & Jondral, F. K. (2008). Detection of spectral resources in cognitive radios using reinforcement learning. In *Proceedings of IEEE DySPAN*.
22. Di Felice, M., Chowdhury, K. R., Meleis, W. & Bononi, L. (2010). To sense or to transmit: A learning-based spectrum management scheme for cognitive radiomesh networks. In *Proceedings of IEEE Workshop on Wireless Mesh Networks (WIMESH)*, pp. 1–6.
23. Vucevic, N., Akyildiz, I. F., & Perez-Romero, J. (2011). Dynamic cooperator selection in cognitive radio networks. *Ad Hoc Networks (Elsevier) Journal* (to appear).
24. Oksanen, J., Lundén, J., & Koivunen, V. (2010). Reinforcement learning-based multiband sensing policy for cognitive radios. In *Proceedings of 2nd Int'l Workshop on Cognitive Information Processing (CIP)*, pp. 316–321.
25. Oksanen, J., Lundén, J., & Koivunen, V. (2010). Reinforcement learning method for energy efficient cooperative multiband spectrum sensing. In *Proceedings of IEEE Int'l Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 59–64.
26. Lee, W. Y., & Akyildiz, I. F. (2008). Optimal spectrum sensing framework for cognitive radio networks. *IEEE Transactions on Wireless Communications, 7*(10), 3845.
27. Gudmundson, M. (1991). Correlation model for shadow fading in mobile radio systems. *Electronics Letters, 27*(23), 2145.
28. Wang, H. S., & Moayeri, N. (1995). Finite-state markov channel–A useful model for radio communication channels. *IEEE Transactions on Vehicular Technology, 44*(1), 163.
29. Singh, S., Jaakkola, T., Littman, M. L., & Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement learning algorithms. *Machine Learning, 38*(3), 287.
30. Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning, 8*, 279.
31. Breiman, L. (1968). *Probability*. Boston, MA: Addison-Wesley.
32. Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York, NY: Wiley.
33. Shiryayev, A. N. (1978). *Optimal stopping rules*. New York, NY: Springer.
34. Poor, H. V., & Hadjiliadis, O. (2009). *Quickest Detection*. Cambridge, UK: Cambridge University Press.

Author Biographies

Brandon F. Lo received the B.S. degree in Computer Science, with honors, from Tunghai University, Taichung, Taiwan in 1992 and the M.S. degree in Electrical Engineering from the University of Michigan, Ann Arbor, MI in 1995. He is pursuing the Ph.D. degree in Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta, GA. Before his doctoral study, Mr. Lo designed processors and ASIC chips for computers and broadband communications in semiconductor industry. His research interests include cognitive radio networks, wireless sensor networks, and next generation cellular networks.

Ian F. Akyildiz received the B.S., M.S., and Ph.D. degrees in Computer Engineering from the University of Erlangen-Nürnberg, Germany, in 1978, 1981 and 1984, respectively. Currently, he is the Ken Byers Chair Professor in Telecommunications with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, the Director of the Broadband Wireless Networking Laboratory and Chair of the Telecommunication Group at Georgia Tech. Dr. Akyildiz is an honorary professor with the School of Electrical Engineering at Universitat Politècnica de Catalunya (UPC) in Barcelona, Catalunya, Spain and the founder of N3Cat (NaNoNetworking Center in Catalunya). Since 2011, he is a Consulting Chair Professor at the Department of Information Technology, King Abdulaziz University (KAU) in Jeddah, Saudi Arabia. Since January 2013, Dr. Akyildiz is also a FiDiPro Professor (Finland Distinguished Professor Program (FiDiPro) supported by the Academy of Finland) at Tampere University of Technology, Department of Communications Engineering, Finland. He is the Editor-in-Chief of Computer Networks (Elsevier) Journal, and the founding Editor-in-Chief of the Ad Hoc Networks (Elsevier) Journal, the Physical Communication (Elsevier) Journal and the NanoCommunication Networks (Elsevier) Journal. He is an IEEE Fellow (1996) and an ACM Fellow (1997). He received numerous awards from IEEE and ACM. His current research interests are in nanonetworks, Long Term Evolution (LTE) advanced networks, cognitive radio networks and wireless sensor networks.