Wu *et al. EURASIP Journal on Wireless Communications and Networking* (2018) 2018:13
DOI 10.1186/s13638-017-1018-9

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Energy-efficiency opportunistic spectrum allocation in cognitive wireless sensor network

Cheng Wu, Yiming Wang[*] and Zhijie Yin

Abstract

The developments in wireless sensor network (WSN) that enriches with the unique capabilities of cognitive radio technique are giving impetus to the evolution of Cognitive Wireless Sensor Network (CWSN). In a CWSN, wireless sensor nodes can opportunistically transmit on vacant licensed frequencies and operate under a strict interference avoidance policy with the other licensed users. However, typical constraints of energy conservation from battery-driven design, local spectrum availability, reachability with other sensor nodes, and large-scale network architecture with complex topology are factors that maintain an acceptable network performance in the design of CWSN. In addition, the distributed nature of sensor networks also forces each sensor node to act cooperatively for a goal of maximizing the performance of overall network. The desirable features of CWSN make Multi-agent Reinforcement Learning (RL) technique an attractive choice. In this paper, we propose a reinforcement learning-based transmission power and spectrum selection scheme that allows individual sensors to adapt and learn from their past choices and those of their neighbors. Our proposed scheme is multi-agent distributed and is adaptive to both the end-to-end source to sink data requirements and the level of residual energy contained within the sensors in the network. Results show significant improvement in network lifetime when compared with greedy-based resource allocation schemes.

Keywords: Wireless sensor network, Cognitive radio, Reinforcement learning, Multi-agent learning, Opportunistic spectrum allocation

1 Introduction

Wireless sensor network (WSN) is one of the most compelling technologies for performing various tasks such as disaster recovery [1], environmental monitoring [2], remote surveillance [3], non-destructive testing of structures [4], military communication [5], and industrial automation [6]. It is composed of small and resource-constrained wireless sensor nodes that are generally deployed on a large scale and suited for particular applications. Traditional sensor nodes in a WSN use a fixed spectrum assignment policy for data transmission, and the performance is limited due to restricted processing and communication power. In recent years, more and more applications such as multimedia wireless sensor networks [7] are posed to change the classical data-only, delay tolerant assumptions in the design of such networks. In addition, with the increasing number of commercial applications, sensors are being deployed in densely populated areas and may experience more significant levels of interference in the unlicensed frequency bands. Thus, the need to propose a frequency-agile design leads to rapid development of Cognitive Radio (CR) in WSN, which can flexibly choose their spectrum of transmission, regulate their transmit power, and thereby support high-bandwidth applications with enhanced network lifetime.

Cognitive Radio Networks (CRN) allow for opportunistic use of the licensed spectrum bands, under the constraint that the operation of the licensed users of the bands are not affected [8]. Specifically for CR-based WSN, the nodes in a sensor network operate independently of each other, owing to their correlated measurements. They may sense the same spectrum to be vacant [9]. Thus, how to share the available spectrum fairly with minimum messaging between the sensor nodes is a critical component of the CR-based framework of WSN. In addition, CWSN still impose great challenges due to the distributed multi-hop architecture, the dynamic network topology, and diverse quality-of-service (QoS) requirements [10]. The

*Correspondence: ymwang@suda.edu.cn
Soochow University, 8 Jixue Rd., 215123 Suzhou, China

major challenges necessitate novel design techniques that simultaneously integrate theoretical research of learning theory. Considering that interacting with network environment, learning from past experience and adapting its functioning in a CWSN, Reinforcement Learning with the ability of interaction with environment, as well as multiagent system in the level of network architecture, are adjacent to improve spectrum utilization.

Machine learning, a field of artificial intelligence, can be used to solve search problems using prior knowledge, known experience, and data. Many powerful computational and statistical paradigms have been developed, including supervised learning, unsupervised learning, trial- and-error learning, and reinforcement learning. These paradigm can struggle to solve large-scale problems with distributed state and action space. Various solutions to this problem under have been studied, such as dimensionality reduction, principle component analysis, support vector machines, and function approximation. Reinforcement Learning is a biologically inspired model using Machine Learning technique (ML), in which an intelligent agent can learn useful knowledge through continuous trial-and-error interactions with external environment [11]. Within a given environment of particular application domain, an agent does always attempt to take best (sometimes optimal) actions to maximize long-term rewards achieved from the environment. The long-term reward is actually the desired value of accumulated reward that the agent expects to receive in the future using the policy, which can be formulated by a value function. The value function is often represented by a look-up table that stores values of pairs of states and actions [12]. The dynamic interaction with the environment and the adaptivity of the learning process are two of the great causes that motivate RL technique to be used for CWSNs, mainly for routing and spectrum decision tasks [13, 14]. In some cases, various solutions based on RL techniques are proved to work better than traditional approaches [15]. However, the large-scale random deployment and distributed operation of the sensors makes the task of sharing the spectrum a non-trivial task.

Multi-agent systems [16] allow to build complex systems composed of multiple interacting intelligent agents. Each agent in the system can sense the environment and achieve its own local knowledge and experience. The agent can then select behaviors based on local information and attempt to maximize the global performance of the system [17]. A typical multi-agent system is decentralized, without a designated controlling agent [18]. The distributed network of wireless sensors and the multi-hop manner to satisfy the application-specific requirements are two of the key features which make the multi-agent framework appealing for CWSNs applications.

Developing a multi-agent learning-based CWSN has a benefit of opportunistic access of spectrum holes, there exists a number of technical issues that need to be addressed for industry practice. In our previous work [19], we have proposed some new spectrum management approaches based on multi-agent reinforcement learning and some function approximation techniques for CR ad hoc networks with decentralized control. However, the CWSN environment has its uniqueness. For example, energy consumption is inevitably limited by volume of batteries in a wireless sensor. Large-scale network topology is another important consideration as the nodes can not rapidly sense global spectrum utilization for opportunistic access. There exists a need to propose novel techniques to improve both computational efficiency and spectrum reuse.

In this work, we describe a reinforcement learning-based solution that allows each sensor sender-receiver pair to locally adjust its choice of spectrum, and transmit power, subject to connectivity and interference constraints. We model this as a multi-agent system, where each action, i.e., choice of power and spectrum, earns a reward based on the utility that is maximized, as shown in Fig. 1. We propose first, a throughput-only approach, where the reward is computed primarily based on successful transmissions, subject to a pre-decided threshold on the interference offered to the licensed or primary users (PUs). As an intuitive example, higher transmission power results in improved SNR at the receiver, but also increases the intra-network sensor-sensor and inter-network sensor-PU interference in the CWSN. Thus, our scheme demonstrates how power and spectrum can be jointly chosen in a distributed manner to reach an optimal assignment. Our second contribution incorporates energy costs in the reward assignment, where a higher rate of energy consumption is penalized. Thus, a sensor that depletes its energy beyond a permissible rate is forced to lower its spectrum switching instants and transmit at lower powers. Each spectrum change involves an added messaging overhead of coordinating the new frequency and transmission parameters on the link and, thus, must also be limited along with the transmit power. Our reward function appropriately weights the gain obtained by successful transmissions and the energy costs to ensure that the CWSN stays connected for long periods of time.

The rest of this paper is organized as follows. Section 2 describes the related work in the use of learning theories for sensor networks. Section 3 presents the network architecture and problem formulation. Section 4 describes the application of our multi-agent reinforcement learning scheme, and Section 5 presents our proposed approaches. We undertake a thorough performance evaluation in Section 6, and finally, Section 7 concludes our work.

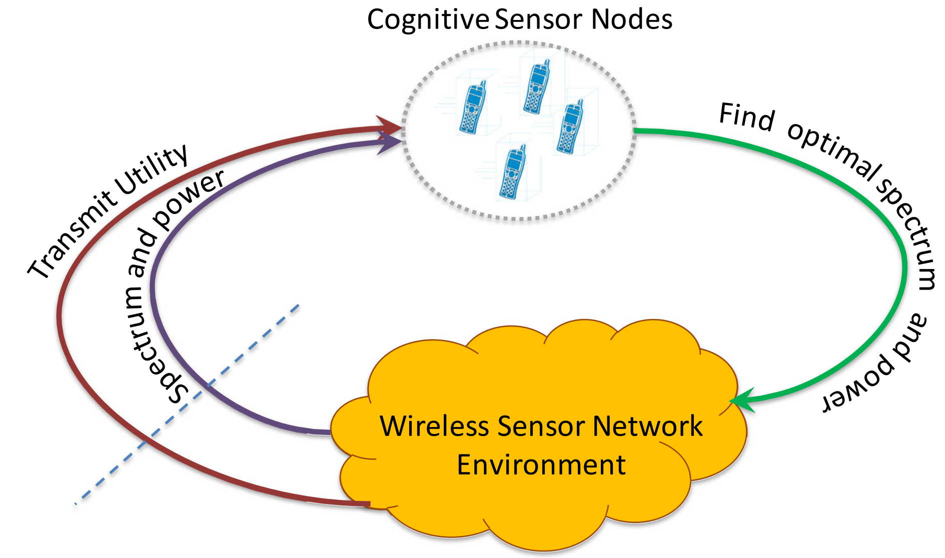

Cognitive Sensor Nodes
Find optimal spectrum and power
Transmit Utility
Spectrum and power
Wireless Sensor Network Environment

Fig. 1 Multi-agent learning framework for typical CWSNs

2 Related work

In recent years, a lot of researchers used cognitive radio to improve the performance of wireless communication. Letaief presented a cognitive space-time-frequency coding technique that can opportunistically adjust its coding structure by adapting itself to the dynamic spectrum environment [20]. Soyeon Kim proposed a CR operational algorithm for mobile cellular systems, which was applicable to the multiple secondary user environment [21]. These results proved CR technology can significantly reduce interference to licensed users, while maintaining a high probability of successful transmissions in a cognitive radio (CR) ad hoc network.

There have been several proposed solutions for spectrum sensing and sharing for distributed ad hoc networks, and these implementations are mainly developed for the link layer [22, 23]. For spectrum allocation, a graph-coloring scheme based on network structure is proposed in [24], where a topology-optimized allocation algorithm is used for the fixed topology which involves advance knowledge of the PU interference regions at a central network entity. Cao and Zheng [25] proposed a distributed scheme of local bargaining-based spectrum re-utilization, in which CR users continuously negotiate spectrum assignment with "local self-organized groups". The scheme uses a poverty threshold to ensure minimum usage of channel allocation to each user and further guarantees the fairness of each user. However, most of these spectrum management frameworks have been shown to be awkward to solve non-cooperative spectrum sharing.

The computer engineering machine learning community has begun to develop algorithms that allow collections of agents to learn to cooperate and compete with one another [26, 27]. There are performance guarantees on the quality of the resulting learned algorithms [28, 29] and these algorithms have been applied to a limited set of problems [30, 31]. In recently, [32] propose an energy-efficient game-based spectrum decision (EGSD) scheme for cognitive radio sensor networks to extend the network lifetime.

The artificial intelligence community has worked to develop game playing algorithms that allow agents to search for optimal moves and learn the biases and weaknesses of a human (or computer) opponent [33–35]. Environments where agents may behave randomly, or using some other suboptimal strategy such as a human strategy, are being developed using opponent modeling [36–38]. In [39], the principle from the game theory is employed to analyze the behavior of the CR user for adaptively assign frequency channels by keeping transmission power constant. For those of resource-constrained networks, such as WSNs and ad hoc networks, a natty rule-based device-centric spectrum assignment mechanism is proposed in [25]. It shows that most of cooperative cases can be modeled using a strategical rule, which can converge to a pure value of Nash-equilibrium.

Very little work has been published in the issue about comprehensive consideration of both spectrum and transmission power. Parekh et al. tried to assign a channel in a fixed power level to only one transmission every time in order to avoid intra-channel interference with other neighbors in [40]. It shows that using such an orthogonal scheme in spectrum and power allocation is optimal for achieve the maximum capacity of the entire network. In [41], not only single channel but also multiple channels using "asynchronous distributed pricing scheme" are further proposed, in which every CR user transfers its interference price, including the information about channel and power level, to other nodes. Although the approach considers both channel and power allocation at the same time, it does not address heterogeneous spectrum availability over time and space, or interference to the PUs which distinguishes CR networks from the classical wireless counterparts. Lunden et al. [10] gives a high-level survey of machine learning approaches to cognitive radio, but does not describe any experimental results. Yau [6] and Wang [42] evaluate the effectiveness of reinforcement learning in achieving context awareness in CR sensor networks, but do not consider energy effects.

In addition, Xiong et al. studied the group cooperation algorithm for optimal resource allocation in wireless powered communication network (WPCN) [43, 44]. These studies describe the optimization of energy cooperation and time allocation between different groups of sensor nodes, enabling the two groups to achieve the desired information transfer. By optimizing the time allocation, power allocation, and SWIPT beamforming vectors under the available power constraints and two sets of QoS requirement constraints, the system WSR is maximized and its overall power consumption is minimized. At the same time, he also studied the energy efficiency (EE) in multiple relay-assisted OFDM systems, which uses decoding-forward (DF) relay beamforming to aid in the transmission of information [45]. These studies give another feasible idea of the improvement of energy efficiency in WPCN, which has a positive inspiration for our research.

3 Network architecture and problem formulation

3.1 Network architecture

Now, we give a brief description about our network architecture. More detailed can be found in [19, 38, 46]. There are a few PUs that are distributing in a spatially overlapped region with the sensor nodes of WSN. These individual sensor nodes are working as CR users, all of which can communicate each other. Each sensor makes decisions on selecting its own channel frequency and transmission power independently of the others in the neighborhood.

In the wireless environment, we assume perfect sensing, that is, all sensors that are within the PU's transmission range can exactly infer whether the interference of the PUs is present or not. For the case of imperfect sensing, the wireless sensor network environment can be formulated as a partially observable Markov Decision Process (POMDP) problem. A typical solution is to use cooperative scheme to compensate for incomplete sensing information. Further, in the case of collision, the sensor can also detect whether the colliding node is a PU transmitter, or another sensor. In order to realize the scenario, we keep the power level of this PU transmitter an order of magnitude higher than the sensor, and this is readily verified in context of physical transmitters. Once energy detection is performed, the receiving sensor observes the received signal energy. If several multiples greater than the sensor-only case, it means that the collision occurs exactly with a PU transmitter. The information is returned back to the sender via a predefined control channel. The kind of simple energy-based PU detection is often called the fast sensing [6]. Another detection method is the fine sensing, which applies feature-based detection to categorize the PU signal according to its signature in order to understand the characteristics of the PU traffic. The fine sensing is more advanced and takes longer duration. Once the location of a PU receiver is unavailable or unknown, such the case is also considered as the occurrence of *PU interference*. It is because that there exists a potential collision in the range of the PU receiver due to concurrent transmission of any sensor. Thus, our network architecture is conservative, which does overestimate the effect of PU interference and guarantee strictly the performance of entire network.

Within the architecture, all of CR users have to correctly monitor spectrum utilization and continuously detect the presence of the PUs. According to the collected information and the changing demands of the higher layers, the system still need a more flexible capability to reorganize the front-end of radio environment. This capability can be realized by the cognitive cycle [47] including four functions about spectrum management:

- *Spectrum sensing*: perceiving the portions of the spectrum currently available
- *Spectrum decision*: selecting the best available spectrum
- *Spectrum sharing*: coordinating the priority to access this channel with other sensors
- *Spectrum mobility*: exactly vacating the channel when the access of a licensed agent is detected

Our design of network architecture attempts to use two critical techniques to enrich with these unique cognitive capabilities. Firstly, the tasks of spectrum sensing and sharing can be performed effectively by leveraging RL technique, wherein the patterns of transmit rules from particular interferers can be learned and recognized. Secondly, by constructing a multi-agent system of CR users and exploiting its unique mechanism of agent interaction, the network can converge to the solution to safeguard the fairness of spectrum sharing and fast recovery if the spectrum is reclaimed by a PU.

3.2 Problem formulation

We have formulated our application domain of CWSN in the preliminary work [19]. Now, we summarize and further supplement as follows.

3.2.1 Choice of power level and spectrum

We denote the transmission power of the ith sensor node as P_{tx}^{i}. The transmission range and interference range of the ith sensor node are given by ${R_{tx}}^{i}$ and ${R_{if}}^{i}$, respectively. The attenuated power incident as P_{rx}^{j} at the jth receiver can be calculated by the free-space path-loss equation, that is,

$$P_{rx}^{j} = \alpha \cdot P_{tx}^{i} \left\{ D^{i} \right\}^{-\beta},$$

where

D^{i} — the actual distance between the jth receiver and the ith sensor node
β — the exponent parameter of path-loss
α — the function of frequency f^{i} chosen by the transmission sensor i

$$\alpha = \frac{c^{2}}{\left(4\pi f^{i}\right)^{2}}$$

$\boldsymbol{c}$ — the speed of light

The values of power levels are discrete, and a jump from one given value to another is feasible in any consecutive time slot. A choice of spectrum by the sensor node, or CR user i is actually the choice of the frequency given by $f^{i} \in F$, where F represents the collection of licensed frequency bands.

3.2.2 Typical network scenarios

Typical network scenarios for different conditions are as follows, described in Fig. 2:

- *Link disconnection*: In wireless sensor network, power shortage at sensor nodes often leads to link disconnection. If the power P_{rx}^{i} of the ith receiver is less than the preset threshold P_{th}, then all the packets would be lost [48]. In such the scenario, the sender should fast jump its higher level of transmission power in order to reestablish a new link. We punish the scenario by applying a negative reward, which is exacly -15 in our frameworks.
- *PU-sensor interference*: It occurs when a PU and any sensor node concurrently select the same spectrum for transmission. Note that we can allow the occurrence of packet collision among the CR users, though it reduces the throughput of the entire network. But, we must strictly avoid the concurrent use of the spectrum shared with a PU. It violates the principle of protection of the licensed devices.
- *CWSN intra-network collision*: If a packet collision occurs due to the concurrent transmission of another sensor, then we say that the collision is intra-network. Collision among the sensors reduces network throughput, which should be avoided as much as possible by fair sharing of another available spectrum. By giving the CR users a reward to choose the unoccupied channel frequencies, if available, each pair of sensors can communicate without affecting its neighbors.
- *Channel induced errors:* Some frequency bands are certainly more robust due to lease channel errors obtained from lower attenuation rates. By selecting these frequency bands with the lowest packet-error-rate (PER) from the bit-error-rate (BER), the sensor nodes reduce the occurrence of retransmissions and eliminate the possible delays of network.
- *Bounded energy use:* The rate at which a sensor uses up its energy is a reflection on its approximate lifetime. Especially in the exploratory stage at the start of the network, the sensor may aggressively explore the search space, thereby consuming large amounts of energy is short durations. This rate of consumption must level out as the network progresses in its operation. The reward may be decided also by how much the current series of actions is depleting the energy of the sensor. Further details on calculating this metric are given in Section 6.3.
- *Successful transmission*: If none of the above scenarios are observed to be true in a given transmission slot, then a packet is considered to be successfully transmitted from a sender to a receiver.

4 Learning framework

A cognitive wireless sensor network can be formalized as a multi-agent system, in which the sensor agents can sense their environment, learn network scenarios, and further optimize their transmission parameters to maximize the performance of network communication. The formulation does fit quite well within the context of multi-agent reinforcement learning.

Actually, the overview of applying reinforcement learning to a cognitive wireless sensor network is illustrated in Fig. 1 in Section 1. Here, each cognitive sensor works as a reinforcement agent. All of the agents can sense current spectrum utilization and perceive their own states, i.e., spectrum bands and power levels for transmission. They then execute spectrum decisions and spectrum mobility to pick up optimal actions, i.e., channel switching or power level jumping. Finally, the agents perform spectrum sharing to signal transmission. After interacting with wireless environment, the agents further get their rewards which are used as the inputs for the coming sensing and cognitive cycle in next step.

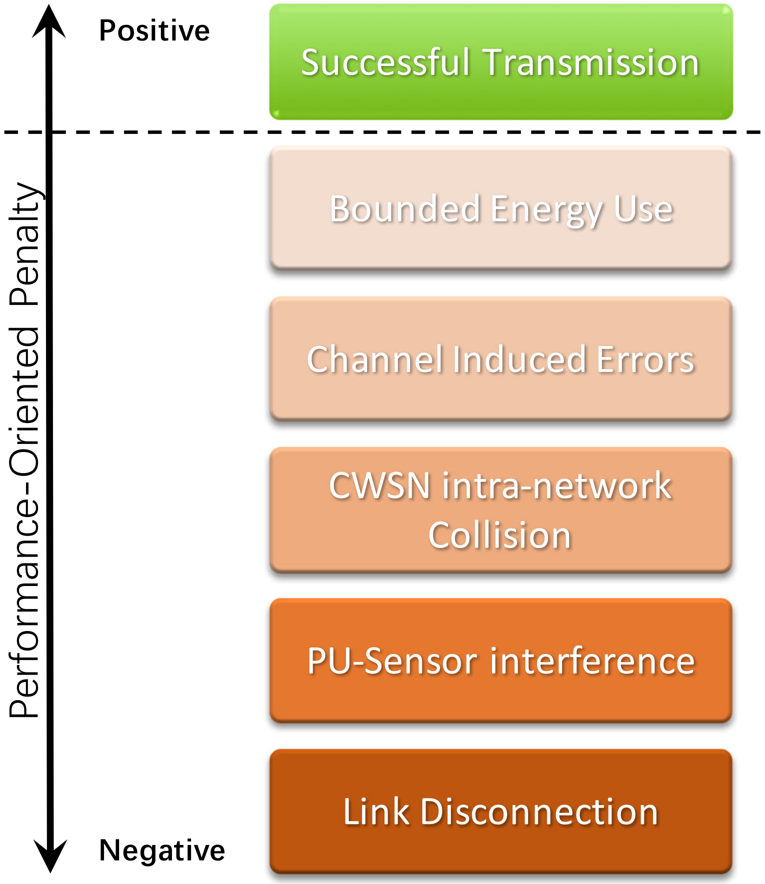


Fig. 2 Penalty levels for different observed scenarios

In a classical RL system, a *state* includes some information that an agent is perceiving from external environment. In RL-based CWSN, a *sensor agent's state* is its current spectrum and power value for signal transmission. In our multi-agent CWSN system, the state is defined as the set of every agent's state. We hence denote the *system state* at time t as $\vec{s_t}$, that is,

$$s_t = \left(\vec{sp}, \vec{pw}\right)_t,$$

where

$\vec{sp}$ — a vector of spectrum bands across all agents $[sp_1, sp_2, ..., sp_i, ...]$

$\vec{pw}$ — a vector of power levels across all agents $[pw_1, pw_2, ..., pw_i, ...]$

Here, sp_i and pw_i are its spectrum band and power level of the ith agent, and $sp_i \in SP$ and $pw_j \in PW$. Specially, assume M spectrum bands and N power levels in the system, we can index them in the way, $SP = \{1, 2, ..., M\}$ and $PW = \{1, 2, ..., N\}$.

An *action* in reinforcement learning is defined as the behavior of an agent's choice from one special state to another. In RL-based CWSN, the *action a of a sensor* allows to either switch from current spectrum to another new available one in the set *SP*, or jump from current power level to another new available one in the set *PW*. Here, we denote the *system action* at time t as $\vec{a_t}$, that is,

$$\vec{a_t} = (\vec{a})_t,$$

where

$\vec{a}$ — a vector of actions over all agents $[a_1, a_2, ..., a_i, ...]$.

Here, a_i is the action of the ith agent, and $a_i \in \{switch_{spectrum}, jump_{power}\}$.

A *reward* in reinforcement learning shows the desirability of an agent's action chosen under a specific state over the environment. In our multi-agent CWSN system, the *system reward*, denoted as r, is achieved according to network performance and/or network energy consumption.

4.1 Sarsa: online control for CWSN

Reinforcement learning enables learning from feedback achieved through interactions with an external environment. The typical algorithm of reinforcement learning is implemented as follows. At every time slot t, an agent senses its current state s_t and the set of possible actions A_{s_t}. The agent then chooses its optimal action $a \in A_{s_t}$, receives a reward r_{t+1} from the environment, and moves to the next state s_{t+1}. After a series of continuous interactions, the reinforcement learner would develop gradually an optimal policy $\pi : S \rightarrow A$ which maximizes the long-term reward

$$R = \sum_t \gamma r_t,$$

where γ is a factor of discounting for subsequent rewards, which satisfies $0 \leq \gamma \leq 1$.

In reinforcement learning, one of the most successful algorithm is Sarsa-learning [11, 49]. The Sarsa-learning algorithm is an on-policy temporal difference learning method. The temporal difference learning method combines the advantages of Monte Carlo methods and dynamic programming, can be applied to model-free, ongoing tasks, and has excellent performance. Compared to Q-learning, Sarsa-learning can select the action to follow the strategy and update the action value function to follow the same strategy. Q-learning way can simplify the execution of algorithm analysis and convergence difficulty, but Sarsa-learning has a higher learning efficiency and faster convergence rate [50]. The algorithm employs a simple update process for value iteration. At the given time t, for current state s_t and current action a_t, the algorithm observes the immediate reward r_t and the Q-value $Q(s_t, a_t)$. The value of $Q(s_t, a_t)$ shows the desirability of the agent under current state s_t and action a_t, which can be used to distinguish which action is optimal. δ is defined as an increment of the Q-value from $Q(s_t, a_t)$ to $Q(s_{t+1}, a_{t+1})$, which is calculated from current reward r_t and the maximal Q-value under the next state $maxQ(s_{t+1}, a_{t+1})$, that is, $r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$. The update of Q-value is calculated by $\alpha \times \delta$, where α is the learning rate such that $0 \leq \alpha_t(s, a) \leq 1$. In this way, the Sarsa-learning can calculate an update to its expected discounted reward, $Q(s_t, a_t)$ as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [\, r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

where γ is the discounting factor and $0 \leq \gamma < 1$. The Sarsa-learning often stores the state-action values in a value table with an exploration rate ε. Figure 3 describes our algorithm for implementing Sarsa-based CWSN scheme.

5 Application of reinforcement learning in CWSNs

We consider two applications in computing the rewards received by an agent given as (i) successful transmission-only and (ii) joint energy and successful transmission depending on the role of energy considerations in the reward assigning process.

5.1 Successful transmission-only

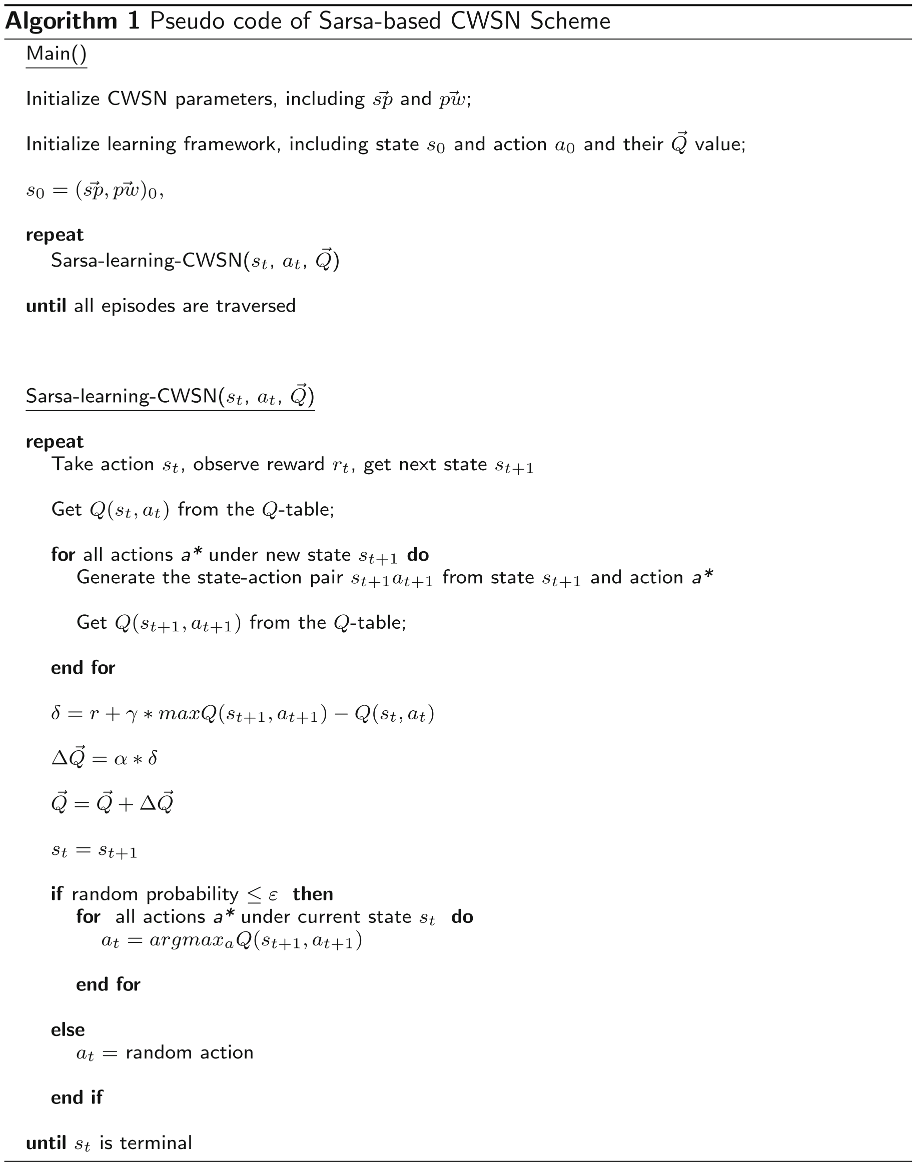
In the first application, we do not consider the energy cost and decide on the suitability of the current choice of spectrum and power based on the successful transmissions only. Such a model is suited for short-lived or replaceable sensors, where the delivery of data is most important for the user.

As shown in Fig. 2, successful transmissions are given a moderately positive reward. For the remaining cases, the rewards are negative. We assume a slow fading environment, where occasional bit-flips may be introduced by the channel. Thus, depending upon the currently observed BER, one or more bits of the packet may be in error. As this error may be present in only a few packets, if at all, and possibly recovered through error correction mechanisms, it is assigned a low negative reward. Collisions between nodes in a CWSN result in complete re-transmission of the packets, and this is a significant additional energy cost, which results in a comparatively higher penalty. When the collision occurs with a PU, the CWSN violates the critical premise of protecting the licensed user operation. In this case, the reward is comparatively greater on the negative scale, forcing the sensor to take a corrective action immediately. As the transmitting sensor is unaware of the power needed to reach the receiver on the other end of the link, it may try choices that result in an incoming signal power lower than the receiver threshold. As the link gets disconnected and the sensor loses energy without achieving any communication, this constitutes the most severe case of failure and earns the highest possible negative reward.

The reward assignment algorithm is shown in Fig. 4, where progressively, in the order of descending penalty, each node evaluates the reward that must be assigned to the state that it is in. The scoring mechanism is from the frequency of events and the impact of events on network performance. The detailed description can be found in [19]. For successful transmission-only, the *Bounded Energy Use* state is absent and, therefore, shown by a gray box. As long as the packet is received correctly by the destination without any CR-PU interference, a reward of +5 is assigned.

5.2 Joint energy and successful transmission

In the second application, each sensor node must decide if its current rate of energy consumption is within a predecided bound and, accordingly, adapt its exploration (i.e., curtail the use of higher power values and limit switching of channels) that is inherently a costly activity from the viewpoint of energy conservation. While the relationship between higher power values and energy cost is trivial, every time a channel switch occurs, the transmitting



```
Algorithm 1 Pseudo code of Sarsa-based CWSN Scheme
Main()
Initialize CWSN parameters, including $\vec{sp}$ and $\vec{pw}$;
Initialize learning framework, including state $s_0$ and action $a_0$ and their $\vec{Q}$ value;
$s_0 = (\vec{sp}, \vec{pw})_0$,
repeat
    Sarsa-learning-CWSN($s_t$, $a_t$, $\vec{Q}$)
until all episodes are traversed

Sarsa-learning-CWSN($s_t$, $a_t$, $\vec{Q}$)
repeat
    Take action $s_t$, observe reward $r_t$, get next state $s_{t+1}$
    Get $Q(s_t, a_t)$ from the $Q$-table;
    for all actions a* under new state $s_{t+1}$ do
        Generate the state-action pair $s_{t+1}a_{t+1}$ from state $s_{t+1}$ and action a*
        Get $Q(s_{t+1}, a_{t+1})$ from the $Q$-table;
    end for
    $\delta = r + \gamma * maxQ(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$
    $\Delta\vec{Q} = \alpha * \delta$
    $\vec{Q} = \vec{Q} + \Delta\vec{Q}$
    $s_t = s_{t+1}$
    if random probability $\leq \varepsilon$ then
        for all actions a* under current state $s_t$ do
            $a_t = argmax_a Q(s_{t+1}, a_{t+1})$
        end for
    else
        $a_t =$ random action
    end if
until $s_t$ is terminal
```

Fig. 3 Algorithm for our proposed learning

sensor must coordinate the new channel with the receiver, and exchange control packets to set up the environmental variables. Thus, by engaging in fewer channel switches, the control overhead is also reduced.

A general sensor node consists of four subsystems: sensing unit, microprocessor, communication unit, and power supply. Normally, the sensing unit is coupled to the microprocessor. The operation of the communication unit depends on whether the microprocessor is in active state. In a distributed architecture, wireless sensor nodes have three operation modes: the first is active mode, in which all of its devices are energized. The second is silent mode, in which its communication unit is closed and effectively temporarily isolate itself from the rest of the network. The third is sleep mode, in which all units shut down.

Once the topology of the sensor nodes is fixed, the only way to reduce the average power consumption of nodes is to periodically turn off some of these units. In most distributed sensor networks, nodes can communicate with the rest of the network only when necessary, in order to achieve significant energy-saving effects. Although censoring sensors is a direct way to save energy, the less intuitive way is to close the sensor nodes completely when the probability of the information content of the next observation is likely to be very small. Conceptually, the sensor node uses a priori knowledge about the process it

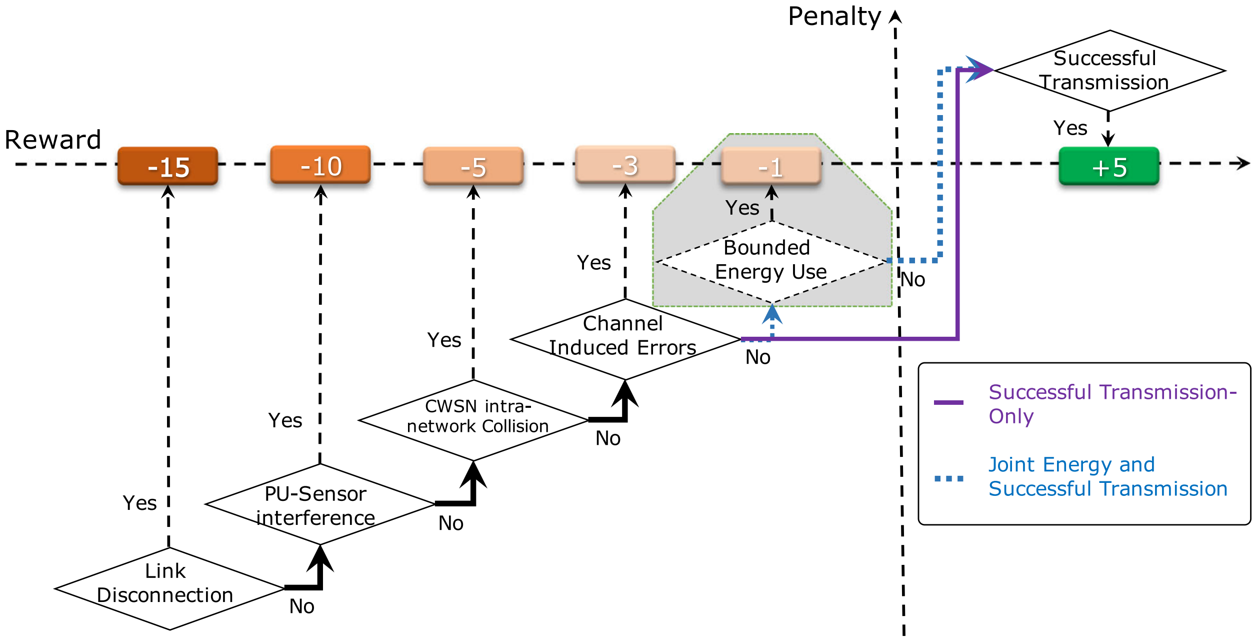


Fig. 4 Algorithm for assigning rewards

is monitoring together with its current and past observations to reduce energy consumption; a large number of research results show that the distributed detection system, and a slight decline in performance can save a lot of energy. For example, a minimal increase in expected control detection delay can more than double the expected lifetime of the sensor node, the results for the wireless sensor node for a long sleep interval support, whenever the information content of the value of the next few observations can be hours. When the event of interest becomes unlikely, the sensor node can sleep for a long time, thus saving energy. On the other hand, in critical situations, sensor nodes must be awake.

In our work, we assign a reward to the sensor based on the rate of energy consumption, R_E computed as:

$$R_E\left(t^e_{end}\right) = \frac{E_{end} - E_{start}}{t^e_{end} - t^e_{start}}, \tag{1}$$

where E_{end} and E_{start} are the residual energy levels within the sensors measured at the start and end of an epoch given by t^e_{end} and t^e_{start}, respectively. Each epoch consists of 50 time slots. We compare this observed rate R_E with the threshold value represented by the *energy curve*

$$R_{lim}(t) = -e^{(\tau/t)} + \phi,$$

at time $t = t^e_{end}$. Here, τ is a design constant chosen prior to the operation of the network, and ϕ is the correction factor that is used to bring down the exponentially falling rate down to zero when a sensor node can no longer reach its neighbor. The rewards based on the energy consumption rate try to guide the energy consumption of the sensor (hence, the power levels or propensity for switching) along the curve R_{lim}. Thus, if the rate is much lower than R_{lim} at a particular instant, the higher reward encourages aggressive exploration and, conversely, with a higher rate (i.e., greater than R_{lim}) of energy consumption, the reward are increasingly negative. The implication of scaling the exponential parameter τ with time is to force the network to get more conservative towards exploring channel and power choices towards the end of the network lifetime.

The reward assignment algorithm (Fig. 4) in this case introduces an additional decision for checking the energy. We check the energy used only if the packet is received correctly, without interference to the PU. Moreover, in this approach, the state may still earn a negative reward of −3 if the rate of energy consumption R_E is exceeds the limit imposed by R_{lim} at that instant of time. Otherwise, as before, a reward of +5 is assigned to the state. The intuition here is that energy conservation is important to a CWSN, but still secondary to the aim of delivering an uncorrupted packet without causing interference to the PU.

Thus, on the basis of rewards, the nodes in the CWSN apply multi-agent reinforcement learning to choose the best combination of spectrum and power, while keeping the energy usage bounded.

6 Performance evaluation

In the section, we evaluate the performance of our *successful transmission-only* (Section 6.2) and the *joint energy and successful transmission* (Section 6.3) applications.

6.1 Simulation platform and benchmarks

6.1.1 Simulation platform

Our novel CR network simulator is described in Fig. 5. We develop the framework using *NS*-2 network simulation tool. Some specific modifications to the physical, link, and network layers is done in the form of stand-alone *C*++ modules. The Multi-agent System Module provides the architecture of multi-agent system, which is composed of an agent object declaimer, agent cooperation mechanism, and agent communication protocol. The Learning Module describes several typical learning algorithms and some common functions. The PU Activity Block defines the activities of PUs using the on-off model, such as transmission and interference range, rule of spectrum occupancy, and location. The Channel Block Module is a channel table which contains the information about background noise and channel capacity. The Spectrum Sensing Block describes some functionalities about energy-aware spectrum sensing. One important function is to notify the Spectrum Management Block if a PU is detected. The Spectrum Management Block would trigger the sensor to switch to another available channel. The Spectrum Sharing Block coordinates the access of channels and calculates the interference of sensor nodes brought by any ongoing transmission. The Wireless Sensor Network Environment Repository facilitates the information about transmission power levels, spectrum bands, locations of sensors, and different network protocols.

We mainly investigate the topology with 100 sensor nodes placed over a square grid of side 1000 m. There are totally 25 PUs in our CWSN system. For each PU, it is randomly assigned its default channel, and the default channel can be kept with a probability of 0.4. Each PU can also switch, with the decreasing probabilities $\{0.3, 0.2, 0.1\}$, to three other pre-assigned successively placed channels, respectively. In the way, these PUs follow an underlying rule, with which they are active on their own given channels. But the information is unknown to the CR sensors.

There are a total of 100 licensed channels. The transmission in the CWSN occurs via these channels connecting multiple pairs of sensor nodes. We denote such a pair with a data link as $\{i, j\}$, which means a directional transmission from the ith sender to the jth receiver. The transmitting spectrum is chosen by the sender node and is notified to the receiver node via a common control channel, also called *CCC*. The information of possible collisions may also be returned back to the sender sensor using this CCC, which may be experienced by the receiver sensor. Here, all data are transmitted only using the chosen spectrum over the link between the pair of sensor nodes.

The permissible value of transmission power for cognitive sensors are uniformly distributed on the interval of $\{0.5\ \text{mW}, 4\ \text{mW}\}$, while the PUs always transmit at the level of 10 mW. We consider the time to be slotted, and the link layer at each sender node attempts to transmit with a probability of 0.2 in every slot. The time scale of the x axis on the following figures is represented by *epochs*, each of

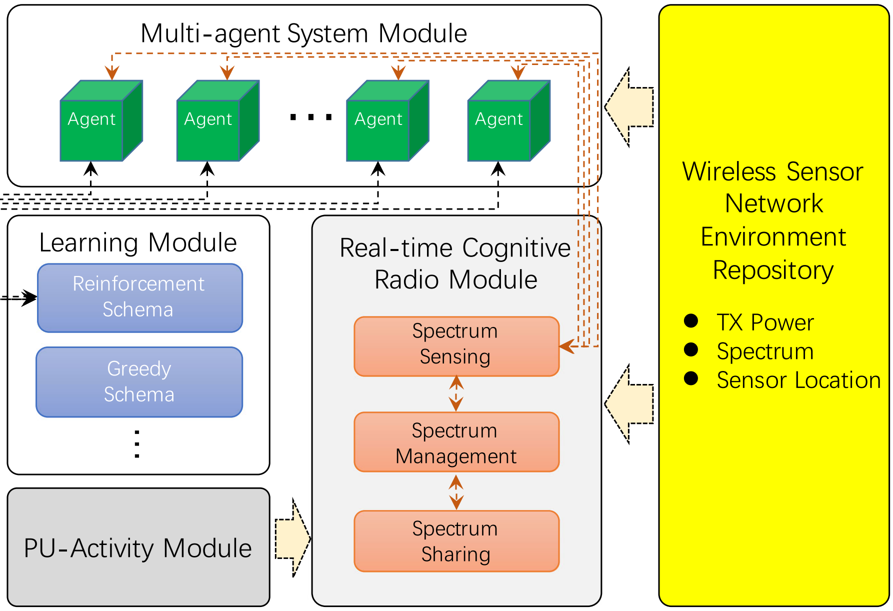


Fig. 5 Block diagram of the implemented simulator tool for RL-based cognitive radio

which is composed of 50 time slots, and we show results for over 600 epochs.

6.1.2 Simulation benchmarks

In our experiment evaluation, the proposed reinforcement-learning-based scheme, abbreviated as *RL Scheme*, is compared with the other three schemes, that is, (i) random assignment, (ii) greedy assignment with 1 memory slot, and (iii) greedy assignment with 20 memory slots.

Random assignment scheme, abbreviated as *RA Scheme*, uses a random combination of spectrum band and power level in each time slot.

Greedy assignment scheme with 1 memory slot, abbreviated as *GD-1 Scheme*, only stores the reward received the last time for every state, that is, for every combination of spectrum band and power level. In order to avoid local optimum, the scheme selects with a probability η the combination having the highest previous reward and explores with a probability $(1 - \eta)$ a random-chosen combination.

Greedy assignment scheme with 20 memory slots, abbreviated as *GD-20 Scheme*, performs a repository of rewards received in the 20 past time slots for every combination of spectrum band and power level and picks up the best one. In a similar way, it selects the best with the probability η and explores a random combination with the probability $(1 - \eta)$.

In *RL Scheme*, the probability of exploration ϵ is set to 0.2. The initial learning rate α is 0.8, which decreases gradually using a scaling factor of 0.995 in every time slot. Note that *GD-1 Scheme* occupies the same amount of memory as *RL Scheme*, but the *GD-20 Scheme* uses twenty times more.

6.2 Successful transmission-only

We now evaluate these four schemes, i.e., *RA Scheme*, *GD-1 Scheme*, *GD-20 Scheme*, and *RL Scheme*, in the small topology with 100 nodes and the large one with 500 nodes. We then observe the (i) average percentage of successful transmissions, (ii) average reward obtained of CR sensors, and (iii) average channel switches of CR sensors.

We apply the above schemes to the small network and give the average percentage of successful transmissions over all transmissions in Fig. 6a. The results show that, after 600 epoches, *RL Scheme* transmits successful packets up to approximately 94.8%, while *GD-20 Scheme*, *GD-1 Scheme*, and *RA Scheme* transmit successful packets with an average percentage of approximately 93.6, 45.7, and 46.1%, respectively. The results indicate that *RL Scheme* can clearly increase the portion of successful packages over all packages transmitted, and its learning performance is much better than the other schemes, even if *GD-20 Scheme* spends the order of magnitude of memory.

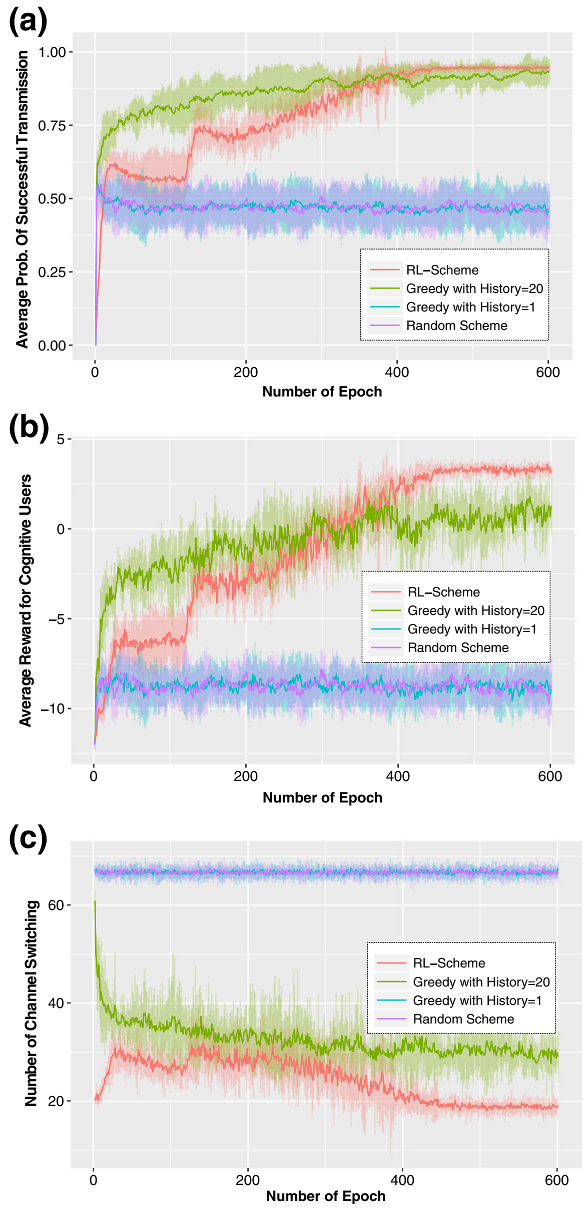


Fig. 6 Performance evaluation of RL-based scheme. **a** Average probability of successful transmission. **b** Average reward of CR users. **c** Average number of channel switches

We also evaluate the average rewards obtained by cognitive sensors with the four schemes. Figure 6b gives the results in the small network. The results show that *RL Scheme* gets the greatest reward about +3.5, and *GD-20 Scheme* has its reward of approximately +1.3, whereas *GD-1 Scheme* and *RA Scheme* have the negative rewards of approximately −8.9 and −8.8, respectively.

The results indicate that *RL Scheme* pushes CR sensors to gradually obtain the higher positive rewards and choose more suitable spectrum band and power level for package transmission. The results also indicate that the reward obtained tends to be proportional to the probability of successful transmission.

Figure 6c shows the average occurrences of channel switching by CR users, again for the small topology. We observe that after learning, *RL Scheme* tends to reduce the number of channel switches to 19.0, wherein *GD-20 Scheme* keeps the channel switches to approximately 29.0, *GD-1 Scheme* keeps the channel switches to approximately 67.4, and *RA Scheme* keeps the channel switches to approximately 66.8. The results indicate that our proposed approach can keep the occurrences of channel switching lower and converge to an optimal solution.

6.3 Joint energy and successful transmission

In the application, each sensor has a fixed amount of energy 1500 mW at the start of the network, which gets depleted with time. Unless specified, the parameter τ is assumed as 1. We also demonstrate the effect of varying τ on the lifetime in this section. Owing to space constraints, we show the measurements for the case of small topology of 100 nodes only, and a similar scenario is observed for the case of 500 nodes.

In Fig. 7, we observe that our energy-aware RL approach displays significant improvement over the basic RL scheme, which does not exploit rewards based on the rate of energy consumption. In the RL-aware approach, each node is allowed to consume energy during exploration phase, while it is forced to get more conservative towards exploring channel and power choices towards the end of the network lifetime. As a result, both the RL schemes show the same performance during the initial exploration phase, but the energy-aware scheme is still operational after the competing scheme show a completely dead network.

We investigate further the performance of our energy-aware approach in Fig. 8 by varying τ that decides the rate at which we allow the network to consume energy. For each of these experiments, the effect of the competing schemes remains the same (being independent of energy considerations) and, therefore, not displayed. We observe in Fig. 8, that at epoch 600, lower values of τ are able to sustain the network longer with a greater residual energy. As τ increases, we observe that the difference in residual energy is much greater in the range $1 - 100$, than in the subsequent range $100-200$. This is attributed to the exponentially increasing value of the R_{lim} function. In Fig. 8, we observe that the CWSN is still partially operational at $60 - 70\%$ nodes in the range $\tau \in [1, 10]$. For higher τ, while the optimal solution may be reached quicker, the network pays a strong penalty in terms of nodes that are alive towards the end of the simulation. Finally, Fig. 8 reveals that for moderate lengths of experiments, in which extreme lifetime of the sensors is not a factor, the value of τ can be freely chosen in the range $[1, 20]$. This also allows the network to converge faster, and the resulting loss of energy does not cripple the network entirely for moderate time scales.

7 Conclusions

We have proposed two approaches for realizing low-memory, low-power sensor networks that are capable of switching multiple spectrum and regulate their transmission power, leading to a novel CWSN paradigm. Our results reveal that RL-based techniques provide good convergence to the best choice of spectrum and power, while ensuring PU protection and energy conservation

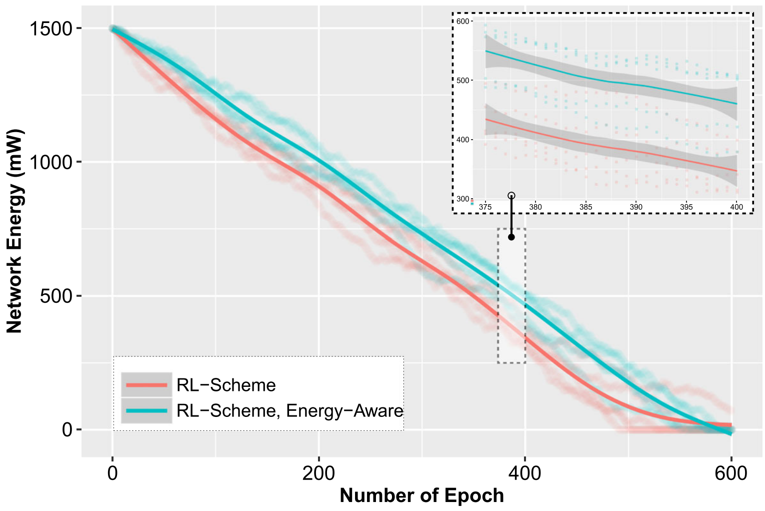


Fig. 7 Energy left in sensor network, for the RL scheme, and the RL energy-aware scheme

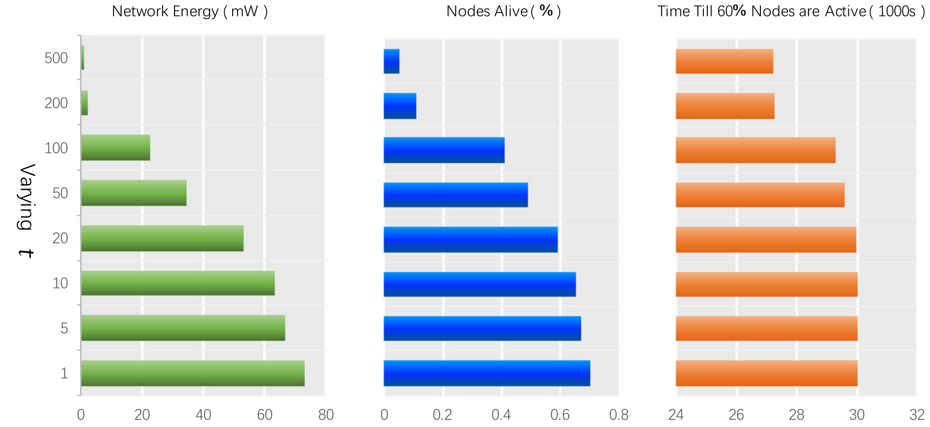


Fig. 8 The network energy after 600 epochs

in CWSNs. Each sensor arrives at its choice independently, by picking a state, jointly defined by spectrum and power based on its local observations. Our method is scalable and takes the first steps towards making a case for reinforcement learning techniques in sensor networks.

Acknowledgements
The work is supported by the National Natural Science Foundation of China (No. 61471252) and the Natural Science Foundation of Jiangsu Province (No. BK20130303).

Authors' contributions
The idea of this work was proposed by CW and YW. CW and ZY performed the experiments and analyzed the simulation results. CW wrote the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 3 June 2017 Accepted: 22 December 2017
Published online: 15 January 2018

References

1. J Horneber, A Hergenroder, A survey on testbeds and experimentation environments for wireless sensor networks. IEEE Commun. Surv. Tutorials. **16**(4), 1820–1838 (2014)
2. MA Alsheikh, Markov Decision processes with applications in wireless sensor networks: a survey. arXiv. **17**(3), 37 (2014)
3. GP Joshi, SY Nam, SW Kim, Cognitive radio wireless sensor networks: applications, challenges and research trends. Sensors (Basel Switzerland). **13**(9), 11196–11228 (2013)
4. IF Akyildiz, T Melodia, KR Chowdhury, A survey on wireless multimedia sensor networks. Comput. Netw. (Elsevier). **51**(4), 921–960 (2007)
5. IF Akyildiz, W-Y Lee, K Chowdhury, CRAHNs: cognitive radio ad hoc networks. Ad Hoc Netw. J. (Elsevier). **7**(5), 810–836 (2009)
6. K Yau, P Komisarczuk, P Teal, in *Proceedings of the 34th IEEE Conference on Local Computer Networks*. Cognitive radio-based wireless sensor networks: conceptual design and open issues (Zurich, 2009), pp. 955–962
7. L Khalid, A Anpalagan, Emerging cognitive radio technology: principles, challenges and opportunities. Comput. Electr. Eng. **36**(2), 358–366 (2010)
8. MA Alsheikh, S Lin, D Niyato, HP Tan, Machine learning in wireless sensor networks: algorithms, strategies, and applications. IEEE Commun. Surv. Tutorials. **16**(4), 1–23 (2014)
9. M Husain, R Guest, M Shadaram, S Zeadally, P Bellavista, Recent developments in cognitive radio sensor networks. Pervasive Mobile Comput. **22**, 1–2 (2015)
10. J Lunden, M Motani, HV Poor, Distributed algorithms for sharing spectrum sensing information in cognitive radio networks. IEEE Trans. Wirel. Commun. **14**(8), 4667–4678 (2014)
11. R Sutton, A Barto, *Reinforcement Learning: An Introduction*. (MIT Press, Cambridge, 1998)
12. C Wu, YM Wang, X Qiang, ZY Zhang, Adaptive spectrum management of cognitive radio in intelligent transportation system. Appl. Mech. Mater. **743**, 765–773 (2015)
13. K Chowdhury, R Doost Mohammady, W Meleis, MD Felice, L Bononi, in *Proceedings of the 5th IEEE Workshop on Wireless Mesh Networks*. To sense or to transmit: a learning-based spectrum management scheme for cognitive radio mesh networks (Boston, 2010), pp. 1–8
14. MD Felice, K Chowdhury, W Meleis, L Bononi, in *Proceedings of the IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM)*. Cooperation and Communication in Cognitive Radio Networks based on TV Spectrum Experiments, (Lucca, 2011), pp. 1–9
15. MD Felice, K Chowdhury, L Bononi, in *Proceedings of 12th ACM international conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. Modeling and performance evaluation of transmission control protocol over cognitive radio ad hoc networks, (Canary Islands, 2009), pp. 4–12
16. RO Saber, JA Fax, RM Murray, Consensus and Cooperation in Multi-Agent Networked Systems. Proc. IEEE. **95**(1), 215–233 (2007)
17. M Veloso, P Stone, Multiagent systems: a survey from a machine learning perspective. Autonomous Robots. **8**(3), 345–383 (2000)
18. R Babuska, BD Schutter, L Busoniu, A comprehensive survey of multiagent system. IEEE Trans. Syst. Man Cybernet. Part C Appl. Rev. **38**(2), 156–172 (2008)
19. C Wu, K Chowdhury, MD Felice, W Meleis, in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 10)*. Spectrum management of cognitive radio using multi-agent reinforcement learning, (Toronto, 2010), pp. 1705–1712
20. KB Letaief, W Zhang, Cooperative communications for cognitive radio networks. Proc. IEEE. **97**(5), 878–893 (2009)
21. S Kim, W Sung, Operational algorithm for wireless communication systems using cognitive radio. Communication Networks and Satellite (COMNETSAT), 2014 IEEE International Conference On, 29–33 (2014)
22. C Cormio, KR Chowdhury, A survey on mac protocols for cognitive radio networks. Ad Hoc Netw. (Elsevier) J. **7**(7), 1315–1329 (2009)
23. T Wysocki, Jamalipour A, Spectrum management in cognitive radio: applications of portfolio theory in wireless communications. IEEE Wirel. Commun. **18**(4), 52–60 (2011)
24. C Peng, H Zheng, BY Zhao, Utilization and fairness in spectrum assignment for opportunistic spectrum access. ACM Mobile Netw. Appl. (MONET). **11**(4), 555–576 (2006)

25. L Cao, H Zheng, Distributed rule-regulated spectrum sharing. IEEE J. Sel. Areas Commun. **26**(1), 130–145 (2008)
26. M Tan, in *Readings in Agents*, ed. by M Huhns, M Singh. Multi-agent reinforcement learning: independent vs. cooperative learning (Morgan Kaufmann, San Francisco, 1997), pp. 487–494
27. M Veloso, P Stone, Multiagent systems: a survey from a machine learning perspective. Auton. Robot. **8**(3), 345–383 (2000)
28. J Hu, M Wellman, in *Proceedings of the 15th International Conference on Machine Learning*. Multiagent reinforcement learning: theoretical framework and an algorithm, (Madison, 1998), pp. 242–250
29. ML Littman, in *Proceedings of the 11th International Conference on Machine Learning (ML-94)*. Markov games as a framework for multi-agent reinforcement learning (New Brunswick, 1994), pp. 157–163
30. P Stone, M Veloso, Towards collaborative and adversarial learning: a case study in robotic soccer. Int. J. Hum. Comput. Syst. (IJHCS). **97**(5), 878–893 (1997)
31. YH Chang, T Ho, in *Proceedings of the 1st International Conference on Autonomic Computing (ICAC'04)*. Mobilized ad-hoc networks: a reinforcement learning approach, (New York, 2004), pp. 240–247
32. S Salim, S Moh, An energy-efficient game-theory-based spectrum decision scheme for cognitive radio sensor networks. Sensors. **16**(7), 1009 (2016)
33. G Weiss, *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. (MIT Press, Cambridge, 2000)
34. J Schaeffer, N Burch, Y Bjvrnsson, A Kishimoto, M Miller, R Lake, PLS Sutphen, Checkers is solved. Science. **317**(5844), 1518–1522 (2007)
35. ML Benitez, F Casadevall, Spectrum usage in cognitive radio networks: from field measurements to empirical models. IEICE Trans. Commun. **E97-B**(2), 242–250 (2014)
36. D Billings, D Papp, J Schaeffer, D Szafron, in *Proceedings of the 15th National Conference on Artificial Intelligence*. Opponent modeling in Poker (Madison, 1998), pp. 493–498
37. C Wu, K Chowdhury, MD Felice, W Meleis, in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 10)*. Spectrum management of cognitive radio using multi-agent reinforcement learning, (Toronto, 2010), pp. 1705–1712
38. MD Felice, K Chowdhury, C Wu, L Bononi, in *Proceedings of the 8th International Conference on Wired/Wireless Internet Communications (WWIC)*. Learning-Based Spectrum Selection in Cognitive Radio Ad Hoc Networks (Lulea, 2010), pp. 133–145
39. N Nie, C Comaniciu, in *Proceedings of the IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*. Adaptive Channel Allocation Spectrum Etiquette for Cognitive Radio Networks, (Baltimore, 2005), pp. 269–278
40. REA Parekh, D Tse, Spectrum sharing for unlicensed bands. IEEE J. Sel. Areas Commun. **25**(3), 517–528 (2007)
41. J Huang, RA Berry, ML Honig, in *Proceedings of the IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*. Spectrum sharing with distributed interference compensation, (Baltimore, 2005), pp. 88–93
42. W Wang, in *Proceedings of the 3rd International Symposium on Intelligent Information Technology Application Workshops*. Spectrum Sensing for Cognitive Radio, (Nanchang, 2009), pp. 410–412
43. K Xiong, C Chen, G Qu, P Fan, KB Letaief, Group cooperation with optimal resource allocation in wireless powered communication networks. IEEE Trans. Wireless Commun. **16**(6), 3840–3853 (2017)
44. K Xiong, P Fan, C Zhang, KB Letaief, Wireless information and energy transfer MIMO-OFDM relay networks. IEEE J. Sel. Areas Commun. **33**(8), 1595–1611 (2015)
45. K Xiong, P Fan, S Member, Y Lu, KB Letaief, Energy Efficiency with proportional rate fairness in multirelay OFDM networks. IEEE J. Sel. Areas Commun. **34**(5), 1431–1447 (2016)
46. C Wu, Y Wang, Cognitive communication in rail transit: awareness, adaption and reasoning. IT Professional. **19**(4), 45–54 (2017)
47. IF Akyildiz, W Lee, KR Chowdhury, CRAHNs: cognitive radio ad hoc networks. Ad Hoc Netw. **7**(5), 810–836 (2009). https://doi.org/10.1016/j.adhoc.2009.01.001
48. P Santi, The critical transmitting range for connectivity in mobile ad hoc networks. IEEE Trans. Mobile Comput. **4**(3), 310–317 (2005)
49. Watkins CJCH, *Learning from delayed rewards*. Ph.D thesis. (Cambridge Univeristy, Cambridge, 1989)
50. C Wu, H Song, C Yan, Y Wang, A fuzzy-based function approximation technique for reinforcement learning. J. Intell. Fuzzy Syst. **32**(6), 3909–3920 (2017). https://doi.org/10.3233/IFS-.162212