Journal of Network and Computer Applications 35 (2012) 253–267

Contents lists available at SciVerse ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues

Kok-Lim Alvin Yau [a], Peter Komisarczuk [a,b], Paul D. Teal [a,*]

[a] Network Engineering Research Group, School of Engineering and Computer Science, Victoria University of Wellington, P.O. Box 600, Wellington 6140, New Zealand
[b] Networks Research Group, School of Computing and Technology, University of West London, St Mary's Road, Ealing, London W5 5RF, United Kingdom

ARTICLE INFO

Article history:
Received 27 January 2011
Received in revised form
17 August 2011
Accepted 23 August 2011
Available online 8 September 2011

Keywords:
Wireless networks
Context awareness
Intelligence
Reinforcement learning
Mobile ad hoc networks
Wireless sensor networks
Cognitive radio networks

ABSTRACT

In wireless networks, context awareness and intelligence are capabilities that enable each host to observe, learn, and respond to its complex and dynamic operating environment in an efficient manner. These capabilities contrast with traditional approaches where each host adheres to a predefined set of rules, and responds accordingly. In recent years, context awareness and intelligence have gained tremendous popularity due to the substantial network-wide performance enhancement they have to offer. In this article, we advocate the use of reinforcement learning (RL) to achieve context awareness and intelligence. The RL approach has been applied in a variety of schemes such as routing, resource management and dynamic channel selection in wireless networks. Examples of wireless networks are mobile ad hoc networks, wireless sensor networks, cellular networks and cognitive radio networks. This article presents an overview of classical RL and three extensions, including *events*, *rules* and *agent interaction and coordination*, to wireless networks. We discuss how several wireless network schemes have been approached using RL to provide network performance enhancement, and also open issues associated with this approach. Throughout the paper, discussions are presented in a tutorial manner, and are related to existing work in order to establish a foundation for further research in this field, specifically, for the improvement of the RL approach in the context of wireless networking, for the improvement of the RL approach through the use of the extensions in existing schemes, as well as for the design and implementation of RL in new schemes.

Contents

* Corresponding author. Tel.: +64 4 463 5966; fax: +64 4 463 5045.
E-mail addresses: kok-lim.yau@ecs.vuw.ac.nz (K.-L. Yau), peter.komisarczuk@tvu.ac.uk (P. Komisarczuk), paul.teal@vuw.ac.nz (P.D. Teal).

1084-8045/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2011.08.007

1. Introduction

In wireless networks, context awareness enables each host to be aware of its operating environment; while intelligence enables each host to make the right decision at the right time to achieve optimum performance. In this article, we advocate the use of reinforcement learning (RL) (Sutton and Barto, 1998) to achieve context awareness and intelligence. RL is an unsupervised machine-learning technique that improves system performance. The phrase 'unsupervised' means the machine-learning technique enables a host to learn knowledge about its operating environment by itself without being overseen by an external teacher or critic.

The rest of this article is organized as follows. Sections 1.1 and 1.2 provides background and motivation for the application of the RL approach in wireless networks through the introduction of *policy-based* and *intelligent-based* approaches, respectively. Section 2 discusses traditional RL including *state*, *action* and *reward*, and open issues in the context of wireless networking. In addition, this section discusses features that are not used in the traditional RL approach, and are not widely applied in the wireless networking literature and yet have great potential for performance enhancement. These are *events*, *rules*, and *agent interaction and coordination*. Section 3 provides tutorial-based discussions on how problems in wireless networks, including routing, resource management and dynamic channel selection (DCS) are solved using RL. Specifically, Section 2 provides discussion on the RL approach; while Section 3 provides discussion on the application of the RL approach. Section 4 provides insight into real applications by presenting our recent work on the implementation of RL for cognitive radio networks. Section 5 discusses open issues in the application of RL approach in wireless networks. All discussions are presented in a tutorial manner in order to establish a foundation for further research in this field; specifically, the improvement of RL in the context of wireless networking, the improvement of RL in existing schemes, and the design and implementation of RL in new schemes.

1.1. Policy-based approach

Traditionally, without the application of intelligence, each host adheres to a strict and static predefined set of rules that is hardcoded. A widely used policy is to define rules through *if–then–else* conditional statements as shown in Fig. 1 or express them in a *state–event* table. When a host encounters a particular condition (or state) and an event in the operating environment, it performs a corresponding action. A *state*, such as queue size, is monitored at all times; while an *event*, such as a call handoff, happens occasionally and it is detected whenever it occurs. An example of the policy-based approach is the backoff mechanism in various medium access control protocols (Bianchi, 2000). The average backoff period is typically doubled on each successive transmission attempt due to failed transmission for a particular packet. A host determines its backoff period without considering its operating environment such as the number of neighbor nodes and the channel quality.

```
if      (state S1,event E1)  then  (action A1);
else if (state S2,event E2)  then  (action A2);
else if (state S3,event E3)  then  (action A3);
...
else    (state Sn,event En)  then  (action An);
end if;
```

Fig. 1. The *if–then–else* predefined policy.

A major drawback of the policy-based system is that since actions are hardcoded, they cannot be changed "on the fly" with respect to the continually changing operating environment. Specifically, the relationships between the states, events and actions are static. Wireless communication is a complex and dynamic system. For instance, the channels, spectrum usage, topology and nodal availability are uncertain factors that affect performance in a complex manner. Hence, a policy-based system may not be able to cater for all possible states and events encountered throughout its operation, resulting in suboptimal performance.

1.2. Intelligent-based approach

An alternative to policy-based approach is to incorporate intelligence into the system, and it is called the intelligent-based approach. Intelligence enables each host to learn new states, events and actions, as well as matching them so that optimal or near-optimal actions can be taken. The adage "practice makes perfect" describes the concept of intelligence. While making a decision on an optimal action is a difficult endeavor, intelligence approximates ans optimal action, in other words, achieves an optimal or near-optimal action as time goes by. A host learns about its actions by evaluating feedback, i.e., the consequences of executing its actions. Since the traditional policy-based system is not receptive to feedback, it does not achieve intelligence. Through learning on the fly, the policy in Fig. 1 evolves with time in order to approximate an optimal policy. The rest of this section discusses the necessity of continuous learning in the intelligent-based approach, as well as introduces RL and its applications in wireless networks.

1.2.1. Necessity of continuous learning

Continuous learning is necessary so that the policy remains optimal or near-optimal with respect to the dynamic environment.

Specifically, there are two main reasons for continuous learning. Firstly, the operating environment evolves with time such that new state–event pairs may be encountered, and new actions may be discovered, hence the policy must be constantly updated to match the state and event pairs with the optimal or near-optimal actions. Secondly, network performance brought about by an action for a particular state–event pair may deteriorate as time goes by, and so rematching may be necessary. Additionally, it should be noted that most operating environment in wireless networks exhibit statistical properties, e.g. traffic load may be a Poisson process; hence, it may take many trials to learn a policy, so continuous learning may be necessary.

1.2.2. Reinforcement learning and its applications in wireless networks

This article applies RL to achieve context awareness and intelligence in wireless networks. RL is a simple and model-free approach, and this implies two characteristics. Firstly, RL represents the performance metric(s) of interest and improves it as a whole, rather than modeling the complex operating environment. For instance, instead of tackling every single factor that affects network performance such as the wireless channel condition and nodal mobility, RL monitors the *reward* resulting from its actions. This reward may be throughput, which covers a wide range of factors that can affect the performance. Secondly, RL does not build explicit models of the other agents' strategies or preferences on action selection. Further discussion on advantages of the application of RL in wireless networks is provided in Giupponi et al. (2010).

Reinforcement learning has seen increasing applications in wireless networks including mobile ad hoc networks (MANETs), wireless sensor networks (WSNs), cellular networks, and recently the next generation wireless networks, such as cognitive radio networks (CRNs) (Mitola and Maguire, 1999). Context awareness and intelligence, and hence the RL approach, are imperative to the successful implementation of CRNs.

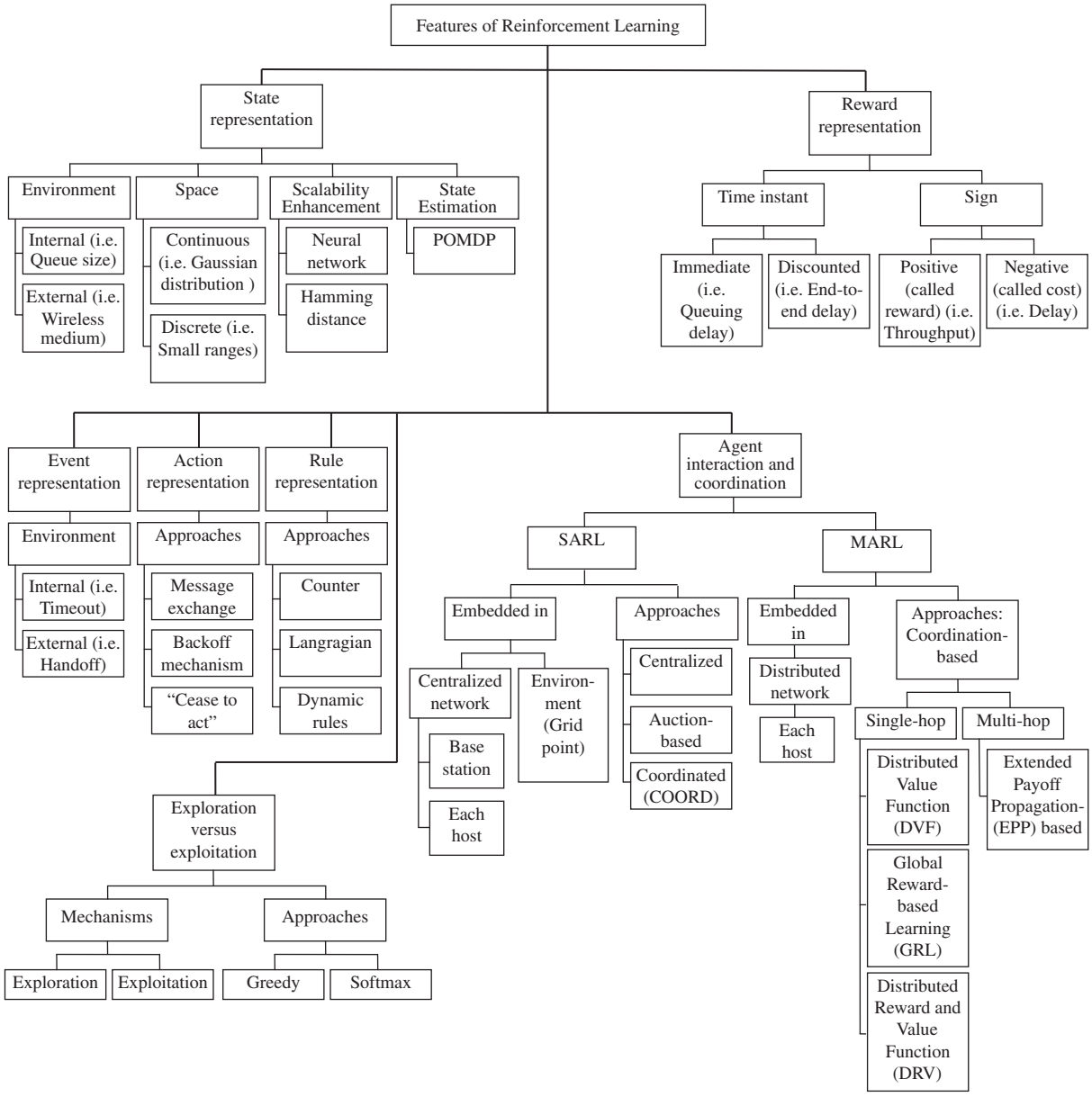

Fig. 2. Taxonomy of the features of RL for its application in wireless networks.

2. Reinforcement learning in the context of wireless networking

Q-learning (Sutton and Barto, 1998) is the most widely used RL approach in wireless networks. This section presents Q-learning, and its overview is presented in Section 2.1, its mathematical representation in Section 2.2, its flowchart in Section 2.3, as well as its components and features in Section 2.4. Figure 2 shows the taxonomy of the features of RL for its application in wireless networks, and it will be explained in Sections 2.1–2.4.

2.1. Q-learning: an overview

In Q-learning, we represent each host in the network as a learning agent as shown in Fig. 3. At a particular time instant, the agent observes a state and an event, as well as a reward, from its operating environment, performs learning, decides and carries out an action. The state and event may describe internal phenomena, which are within the agent, such as instantaneous queue size, or external to the agent, such as the usage of the wireless medium. The state and event are differentiated in that the *state* is monitored at all times, whereas *events* happen occasionally and in general are detected whenever they occur. At any time instant t, the agent carries out a proper action so that the reward is the maximum possible in the next time instant $t+1$. The most important component in Fig. 3 is the learning engine that provides knowledge of the operating environment through observing the consequences of its prior action including the state, event and reward. As an example, the learning engine is used to learn to take the best possible action under complex channel conditions including channel quality and channel utilization level. Various kinds of actions can be carried out by the agent including a message exchange, a backoff mechanism, and even "cease to act". As time progresses, the agent learns to carry out a proper action given a particular state–event pair.

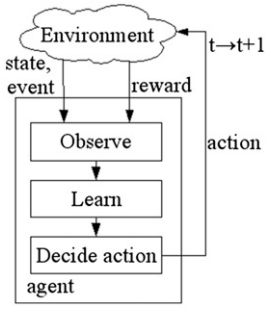

Fig. 3. Abstract view of a RL agent in its environment.

In Q-learning, the *learnt action value or Q-value, Q(state, event, action)* is updated using *immediate reward* and *discounted reward*, and maintained in a two-dimensional lookup Q-table with size $|(state,\ event)| \times |action|$, with $|arg|$ being the cardinality of *arg*. The immediate reward is the reward received at time $t+1$ for an action taken at time t. For each state–event pair, an appropriate action is rewarded and its Q-value is increased. In contrast, an inappropriate action is punished and the Q-value is decreased. Hence, the Q-value indicates the appropriateness of an action selection in a state–event pair. At any time instant, the agent chooses an action that maximizes its Q-value. The reward corresponds to performance metric such as throughput. The expected future return is the cumulative reward that an agent receives in the long run. Since the Q-value provides an estimate of the present value of the rewards, the expected future return is discounted to its present value, hence the term discounted reward.

2.2. Q-learning: mathematical representation

Denote state by s, event by e, action by a, reward by r, learning rate by α and discount factor by γ. A negative reward represents a cost; thus if reward is maximized, cost is reduced. We refer to cost as negative reward henceforth. At time $t+1$, the Q-value of a chosen action in a state–event pair at time t is updated as follows:

$$Q_{t+1}(s_t,e_t,a_t) \leftarrow (1-\alpha)Q_t(s_t,e_t,a_t) + \alpha(r_{t+1}(s_{t+1},e_{t+1}) + \gamma \max_{a \in A} Q_t(s_{t+1},e_{t+1},a)) \quad (1)$$

where $0 \le \alpha \le 1$ and $0 \le \gamma \le 1$. If $\alpha=1$, the agent will forget its previous learnt Q-value, and replace it with the most recent estimated reward. The higher the value of γ, the greater the agent relies on the discounted reward, which is the maximum Q-value in the next state–event pair. Unless $\gamma=1$ where the discounted and immediate rewards share the same weight, the discounted reward always has lower weight than the immediate reward. RL approximates an optimal policy π that maximizes its accumulated reward or value function by choosing the action with maximum Q-value as follows:

$$V^{\pi}(s,e) = \max_{a \in A} Q_t(s_t,e_t,a) \quad (2)$$

An example of the use of discounted reward (or negative reward) is found in multi-hop routing (Arroyo-Valles et al., 2007). A negative immediate reward represents the time delay introduced by a

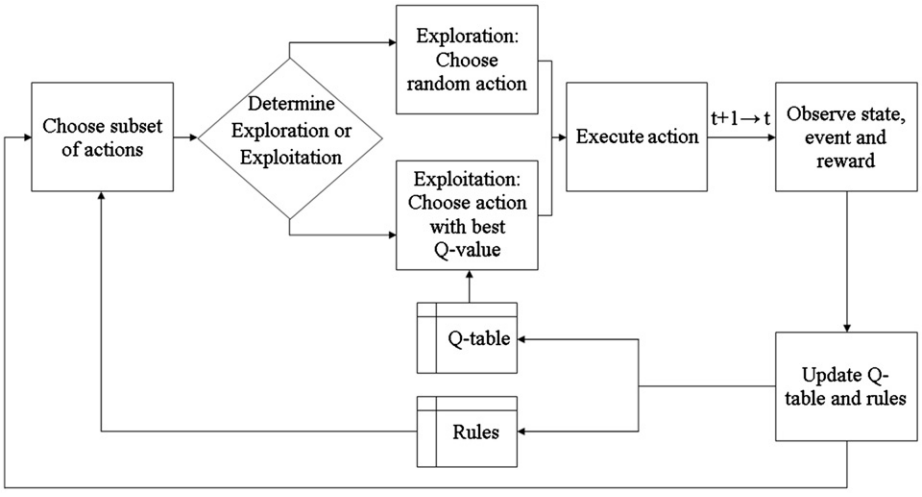

Fig. 4. Flowchart of the RL approach.

particular upstream node (action), while a negative discounted reward represents the amount of end-to-end delay from an upstream node choice (action) to a destination (state).

2.3. Q-learning: flowchart

Figure 4 shows the flowchart of the RL approach. At time t, an agent chooses a subset of actions in adherence to a set of *rules* that exclude actions that violate the network requirements. Next, it chooses either an *exploitation* action, which is the best known action derived from its Q-table, or an *exploration* action, which is a random action chosen to increase knowledge of the environment. At the next time instant $t+1$, it observes the consequences of its previous action including the state, event, and reward; and updates its Q-table and rules accordingly. Further explanation is given in the next few subsections. In general, to apply RL, these representations are necessary: 1) state, event, action and reward and 2) rules. Based on a scheme, some of the representations can be ignored while still achieving its goals. For instance, the state representation can be omitted in some RL approaches and these are called stateless models. Clearly, if an agent's state never changes throughout its operation, or the state is comprised of one entity only, the stateless model can be applied.

2.4. Q-learning: components and features

This section presents the components and features of Q-learning, namely space representation, exploration and exploitation, rules, as well as agent interaction and coordination. Table 1 summarizes the approaches and references for each component and feature.

2.4.1. Space representation

Not all the elements in the operating environment within which a wireless host resides may be important unless network performance can be improved by addressing them. The state, event, action and reward spaces incorporate the important decision-making factors of a scheme. The state characterizes the environmental factors that require constant monitoring; while the event represents the occurrence of events of particular interest that may happen occasionally in the environment. For instance, there are many factors in the operating environment in which a host resides, such as channel condition (signal-to-noise ratio), number of neighbor nodes, and queue size; suppose only queue size is an important factor in a routing algorithm, so the host represents the queue size as state. Additionally, if call handoff, which happens occasionally, can affect the rewards, then it can be represented as event.

The variables for the state, event, action and reward can be discrete or continuous. For discrete space, the variable may represent an interval of values segregated into smaller ranges representing different stages or levels as applied in Fu and Schaar (2009), or may be counter to keep track of the number of occurrences, or a Boolean representing an occurrence. In a complex scenario, the space can be too large to be stored in memory or lack of scalability. To reduce the number of states and events, two states or events that are close to each other can be merged if the difference between them is less than a threshold value (Sutton and Barto, 1998). This difference may take the form of a Hamming distance (the number of bits at which two state representations differ). In Galindo-Serrano and Giupponi (2010a), a neural network approach is applied to represent the Q-values for all combination pairs of state and action in order to improve scalability; however, a disadvantage is the requirement of larger number of computational operations than the lookup Q-table approach. For continuous space, it provides an added advantage of scalability as the agent does not keep track of each entry of state-event-action in its Q-table. An example is called REINFORCE that uses the Gaussian distribution to generate real-valued actions using the mean and variance of the state, which is updated using the reward (Vucevic et al., 2007). However, in Q-learning, it is not possible to represent continuous space in a tabular format. Future research could be pursued for effective approximation-based techniques to achieve continuous space representation. In some scenarios, such as a noisy environment, an agent may not be able to observe its state clearly, and a state estimator based on a partially observable Markov decision process (POMDP) (Littman et al., 1995; Murphy, 2000) may be applied as in Galindo-Serrano and Giupponi (2010a) to compute a belief state, which is a probability distribution over the known states of the environment.

Table 1
A summary of components and features of the RL approach.

RL representation/feature	Approaches	Reference
State	Segregate discrete space into small ranges	Fu and Schaar (2009)
	Improve scalability using Hamming distance	Sutton and Barto (1998)
	Improve scalability using neural network	Galindo-Serrano and Giupponi (2010a)
	Represent continuous space using Gaussian distribution	Vucevic et al. (2007)
	Estimate state using POMDP	Galindo-Serrano and Giupponi (2010a)
Exploration and exploitation	ε-greedy and the softmax approach	Sutton and Barto (1998)
Rules	Count the number of times an entry is visited and violated using counters	Yu et al. (2008)
	Embed constraints in Q-value to reward an entry negatively if it is violated using the Lagrangian approach	Salodkar et al. (2010)
	Reduce the number of entries using dynamic rules in order to decrease the number of explorations	Shiang and Schaar (2010)
Agent interaction and coordination	Embed SARL in base station of centralized networks	Yu et al. (2008)
	The auction-based approach: embed SARL in wireless host of centralized networks	Fu and Schaar (2009) and Salodkar et al. (2010)
	The COORD approach: embed SARL in entities in the operating environment	Seah et al. (2007)
	Eliminate stale or outdated entries using decay model	Dowling et al. (2005)
	Single-hop coordination-based MARL: apply the Distributed Value Function (DVF) approach	Renaud and Tham (2006) and Seah et al. (2007)
	Single-hop coordination-based MARL: apply the distributed Global Reward-based Learning (GRL) and Distributed Reward and Value (DRV) function approaches	Naddafzadeh-Shirazi et al. (2010)
	Multi-hop coordination-based MARL: extend the payoff propagation approach	Yau et al. (2010c) and Yau et al. (2010d)

2.4.2. Exploration versus exploitation

The update of the Q-value in Eq. (1) does not cater for actions that are never chosen (Sutton and Barto, 1998). *Exploitation* chooses the best known action, or the greedy action at all times. *Exploration* chooses other actions once in a while in order to improve the estimates of all the Q-values in the Q-table so that better actions may be discovered. The balance between exploitation and exploration depends on the accuracy of the Q-value estimation and the level of dynamic behavior in the environment. Examples of tradeoff methodologies are ε-greedy and the softmax approach (Sutton and Barto, 1998). In the ε-greedy approach, an agent chooses the greedy action as its next action with probability $1-\varepsilon$, and a random action with a small probability ε.

2.4.3. Rules

Q-learning must achieve a high level of reward without violating the constraints or rules, which can be imposed by the system requirements (i.e. quality of service (QoS) parameters such as end-to-end delay and packet dropping probability). Thus, (*state*, *event*, *action*) entries that violate the rules are marked. Whenever a state and event pair is encountered, the actions that violate the rules are prohibited. Following is an example of a method by which statistical information is collected so that the rules can be applied to mark (*state*, *event*, *action*) entries as violations (Yu et al., 2008). The agent keeps track of two counters, $C^M_{(s,e,a)}$ and $C^V_{(s,e,a)}$. The $C^M_{(s,e,a)}$ counts the number of times a particular (*state*, *event*, *action*) is found to violate the rules; while $C^V_{(s,e,a)}$ counts the number of times the (*state*, *event*, *action*) is visited. An action becomes illegitimate when the ratio of $C^M_{(s,e,a)}$ to $C^V_{(s,e,a)}$ is greater than a threshold value. Another approach is to embed the contraints in the Q-value so that actions that violate the constraints are negatively rewarded (Salodkar et al., 2010). This is accomplished through converting a constrained problem into an unconstrained problem using the Lagrangian approach.

The definition of the rules is dependent on the schemes, and hence the rules can be static or dynamic. For instance, $T_{(s,e,a)}$ may be static in order to conform to QoS requirements; or alternatively it may be dynamically adjusted so that the number of entries is reduced in order to decrease the number of explorations necessary to approximate the optimal action as applied in Shiang and Schaar (2010). The illegitimate entries may need to be explored in the future as they may become legitimate and optimal actions in a dynamic environment. Future research could be pursued to investigate types of rules, their objectives and necessities; and the timing and the conditions under which the rules could be applied.

2.4.4. Agent interaction and coordination

In a wireless network, which uses a shared medium, the actions of an agent may change the environment experienced by another agent, and multi-agent RL (MARL) may be necessary. For instance, if two neighbor hosts access a similar channel in a multi-channel environment, they share the rewards or transmission opportunities among themselves and this may result in a lower level of overall reward. Conflict does not always arise as actions such as channel sensing may not change the environment, and so the single-agent RL (SARL) approach may be sufficient. The SARL approach has been called RL in most of the literature. In this paper, we use SARL and RL to refer to the single-agent approach, and MARL to refer to the multi-agent approach. The two types of RL approaches are:

- Single-agent reinforcement learning (SARL) (Sutton and Barto, 1998) is suitable to be applied in centralized networks. The SARL approach, as shown in Fig. 3, can be embedded in the base station, the wireless host or even entities in the operating environment. For scenarios with SARL embeded in the base station (Yu et al., 2008), the base station is the only agent to learn and take actions that maximize its *individual network performance*.
  As the size of the state space increases exponentially with the number of wireless hosts, the SARL approach can be embedded in each wireless host to make their own action selection, which is subsequently sent to the base station, in order to improve scalability and computational efficiency. The base station makes the final decision on action selection and broadcasts the outcome to the hosts. Embedding SARL in each wireless host in a centralized network is called an auction-based approach and it has been applied in Fu and Schaar (2009) and Salodkar et al. (2010) with low communication overheads.
  In Seah et al. (2007), an SARL approach called coordinated (COORD) is applied to the sensing coverage scheme in WSNs to reduce the power consumption of sensor nodes. The COORD approach is embedded in some grid points that an agent, which is the sensor node, covers within a network-wide region. An agent considers the state of its grid points, consolidates the Q-values of the possible actions available, and subsequently makes a decision on action selection (Seah et al., 2007).
  In a multi-agent scenario (Kok and Vlassis, 2006), SARL may not achieve stability, specifically, the agents may change their respective actions frequently, or oscillate between actions, and fail to achieve an optimal joint action as shown in Gelenbe and Gellman (2007) and Yau et al. (2010d). Despite its limitation, the SARL approach has been applied to a number of multi-agent scenarios and it has been shown to achieve stability in Niyato and Hossain (2009).
- By contrast, multi-agent reinforcement learning (MARL) (Kok and Vlassis, 2006) is suitable to be applied in distributed networks where there are multiple agents. The MARL approach shown in Fig. 5 decomposes a network-wide problem into components, each of which is solved by a self-organized agent. In Fig. 5, each host is represented as an agent, and the hosts share information related to the local rewards among themselves so that each of them can evaluate their own action as part of the joint action in a shared environment (Kok and Vlassis, 2006). The messages exchanged among the agents can be piggybacked in the control packets, and they are called payoff messages. The information exchanged may become stale as time goes by, and a *decay* model is applied in Dowling et al. (2005): the stored values in the absence of new message exchange are decayed so that outdated actions are gradually degraded, and not chosen. The joint action is the combination of actions taken by all the agents throughout

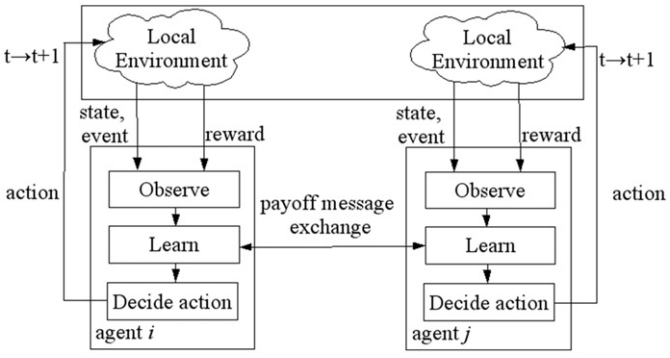

Fig. 5. Abstract view of MARL agents in their environment.

the entire network, and it converges to optimal or near-optimal *network-wide performance*. In other words, the agents share *local rewards* among themselves with the objective of maximizing the *global reward* in order to approximate an optimal joint action. The global reward is the sum of local rewards at each agent. There are two categories of MARL approaches in terms of the number of hops involved in payoff message propagation. *Single-hop coordination-based MARL approach* requires coordination among agents within one hop from each other; and *Multiple-hop coordination-based MARL approach* requires coordination among agents within multiple hops from each other. The two categories of the MARL approaches are shown in the following subsections.

2.4.4.1. Single-hop coordination-based MARL. Consider a scenario where an agent chooses an action, whether to sense or not to sense its environment, based on the action selections of its one-hop neighbor agents. This approach has been applied in sensing coverage scheme in WSNs (Seah et al., 2007; Renaud and Tham, 2006) to reduce power consumption of wireless sensor nodes, which are the agents, in order to increase network lifetime. The optimal or near-optimal action is achieved when there is the smallest possible number of sensor nodes sensing several landmark locations in a sensing field. In Seah et al. (2007) and Renaud and Tham (2006), a MARL approach called distributed value function (DVF) is applied and it is described here (Seah et al., 2007):

Each agent i constantly broadcasts payoff messages to the set $\Gamma(i)$ of its neighbor agents. Each payoff message contains agent i's value function or Eq. (2). In other words, using payoff message, agent i informs agent $j \in \Gamma(i)$ about the Q-values of itself so that agent j can consider agent i when agent j is deciding on its own action. The following equation is applied to update the local Q-values at each agent i:

$$Q^i_{t+1}(s^i_t,a^i_t) \leftarrow (1-\alpha)Q^i_t(s^i_t,a^i_t)+\alpha(r^i_{t+1}(s^i_{t+1})+\gamma \sum_{j \in \Gamma(i)} f^i(j)V^j_t(s^j_t)) \quad (3)$$

where $f^i(j)$ is the weigh factor of agent j's value function at agent i. The payoff messages are exchanged among the agents until a fixed optimal point is reached. During exploitation, the agent selects an action that maximizes the Q-value following Eq. (2). The action is part of the joint action that maximizes the accumulated global reward. In Naddafzadeh-Shirazi et al. (2010), other single-hop coordination-based MARL approaches, namely global reward-based learning (GRL) and distributed reward and value function (DRV), are applied to maximize throughput per transmitted energy in a cooperative retransmission scheme. The GRL approach updates Q-value using an approximate value of global reward, which is the average immediate reward received from an agent's neighbor agents; hence each payoff message contains an agent's immediate reward or $r^i_{t+1}(s^i_{t+1})$. The DRV approach, which is a hybrid of DVF and GRL, updates Q-value using neighbor agents' value function and approximate value of global reward; hence each payoff message contains an agent's immediate reward and value function. The DRV is shown to outperform DVF and GRL since it maximizes both immediate and discounted rewards. In other words, DRV enhances short-term and long-term performance.

2.4.4.2. Multiple-hop coordination-based MARL. Consider a multi-channel scenario where an agent selects its actions, which are the operating channels, based on the channel selection of its two-hop neighbor agents. Two hops are considered because these neighbor nodes are within the interference region of a node. This approach has been applied to dynamic channel selection in distributed

Table 2
Performance enhancements using RL in routing schemes.

Routing scheme																				
Performance enhancement compared to traditional approach	**Reference**																			
	Ad hoc								CR	Wireless sensor										
	Chang et al. (2004)	Curran and Dowling (2005)	Fu et al. (2005)	Henkel and Brown (2008)	Tao et al. (2005)	Usaha (2004, 2007)	Zaine and Mellouk (2010)	Bing et al. (2009)	Arroyo-Valles et al. (2007)	Baruah and Urgaonkar (2004)	Dong et al. (2007)	Egorova-Forster and Murphy (2007)	Forster and Murphy (2007, 2009)	Hu and Fei (2010)	Liang et al. (2008a, 2008b)	Ouferhat and Mellouk (2009)	Zhang and Fromherz (2006)
Lower end-to-end delay			×	×	×		×	×	×						×	×	
Higher throughput	×	×	×		×		×		×				×		×	×	
Lower network overhead						×						×	×				
Higher probability of searched path						×											
Higher network lifetime										×	×		×	×		×	×

CRNs (Yau et al., 2010c, 2010d) to enhance throughput performance. In Yau et al. (2010c), a stateless model is applied. The coordination is achieved using an extended payoff propagation (EPP) mechanism so that the joint action converges to optimal or near-optimal network-wide performance including in distributed CRNs with cyclic topology; and fast convergence is possible. Optimal or near-optimal action is achieved when the channel selection of all the agents provides the best possible throughput in the presence of channel contention and channels with different levels of channel utilization and channel quality in the licensed users' spectrum. In Yau et al. (2010c, 2010d) multiple-hop coordination-based MARL encompasses SARL and EPP, and it is briefly described next.

A major different between single-hop and multiple-hop coordination-based MARL is that, the single-hop approach uses Q-values, which are included in the payoff messages from agents within one hop from each other; and the multiple-hop approach uses Q-values from agents within multiple hops from each other. So, each payoff message from agent i contains agent i's Q-values and those of its two-hop neighbor agents $\Gamma(i)$. The Q-values are updated using Eq. (1), while the agent selects an action that maximizes the local payoff $g_t^i(a_t^i)$:

$$g_t^i(a_t^i) = \max_{a^i \in A} \left[Q_t^i(a^i) + \sum_{j \in \Gamma(i)} Q_{t,a^i}^j(a_t^j) + \sum_{k \in \Gamma(j)} Q_{t,a^i}^k(a_t^k) \right] \quad (4)$$

where $Q_{t,a^i}^j(a_t^j)$ and $Q_{t,a^i}^k(a_t^k)$ are the Q-values at one-hop and two-hop neighbor agents, respectively, when agent i takes action a_t^i. The payoff messages are exchanged among the agents until a fixed optimal point is reached. Future research could be pursued to further investigate interaction and coordination among the agents.

3. Application of reinforcement learning in wireless networks

Reinforcement learning has been applied in a variety of schemes, particularly routing, resource management and dynamic channel selection in wireless networks especially in ad hoc networks, cognitive radio networks cellular networks and wireless sensor networks. This section provides a survey on the performance enhancements brought about by the RL approach in various schemes and various kinds of wireless networks.

Using the RL approach, Table 2 presents performance enhancements compared to traditional approaches (or schemes) in routing. From the table, performance enhancement of higher throughput also indicates higher packet delivery rate and lower packet loss rate; while higher network lifetime also indicates lower energy consumption. As an example, the RL approach has been applied to provide lower end-to-end delay in ad hoc networks (Fu et al., 2005; Henkel and Brown, 2008; Tao et al., 2005; Zaine and Mellouk, 2010), CRNs (Bing et al., 2009) and wireless sensor networks (Arroyo-Valles et al., 2007; Liang et al., 2008a, 2008b; Ouferhat and Mellouk, 2009). An example of traditional scheme is ad hoc on-demand distance vector (AODV).

Using the RL approach, Table 3 presents performance enhancements compared to traditional approaches (or schemes) in resource management. From the table, performance enhancement of lower blocking probability also indicates lower dropping probability; while higher throughput also indicates higher resource utilization (or spectrum efficiency) and hence higher total revenue. Note that, the main objectives of achieving higher throughput are different in routing and resource management.

Using the RL approach, Table 4 presents performance enhancements compared to traditional approaches (or schemes) in dynamic channel selection. From the table, performance enhancement of lower blocking probability also indicates lower dropping probability; higher throughput also indicates higher packet delivery rate and lower packet loss rate; and lower number of channel switches also indicates lower number of channel reassignments.

The RL approach has also been applied in other schemes, and the performance enhancements brought about by RL in those schemes are summarized as shown in Table 5. From the table, *single link* indicates the application of RL at a particular transmission pair comprised of a transmitter and a receiver, hence it may be applicable to any kinds of wireless networks.

The rest of this section discusses the applications of RL using three examples. In general, we describe the objectives of the scheme and define the state, action, and reward, as well as the achievements of the RL approach. The first scheme on *routing in MANETs* adopts a direct application of the traditional RL approach and applies rules; the second scheme on *resource management in centralized networks with mobile hosts* shows how the events and rules are applied in addition to the traditional RL approach; and the third scheme on *dynamic channel selection in distributed CRNs* shows how the feature of agent interaction and coordination is applied in addition to the traditional RL approach.

3.1. Example 1: RL-based routing in mobile ad hoc networks

A MANET is a dynamic and distributed multi-hop wireless network established by a number of mobile hosts in the absence

Table 3
Performance enhancements using RL in resource management schemes.

Resource management scheme											
Performance enhancement compared to traditional approach	**Reference**										
	Ad hoc		Cellular								Wireless sensor
	Liu et al. (2004)	Yagan and Tham (2005)	Alexandri et al. (2002a, 2002b, 2002c)	Chanloha and Usaga (2007)	Giupponi et al. (2008)	Lilith and Dogancay (2007)	Martinez-Bauset et al. (2008)	Vucevic et al. (2009)	Xue et al. (2008))	Yu et al. (2004a, 2004b, 2008)	Saoseng and Tham (2006)
Lower end-to-end delay	×										
Higher bandwidth availability		×									
Higher new call arrival rate										×	
Lower blocking probability			×		×	×	×		×		
Higher throughput			×			×					
Higher user satisfaction probability								×	×		
Lower complexity				×							
Higher reliability											×

Table 4
Performance enhancements using RL in dynamic channel selection schemes.

Dynamic channel selection scheme									
Performance enhancement compared to traditional approach	**Reference**								
	Cellular				Cognitive radio				
	Bernardo et al. (2009a, 2009b)	El-Alfy et al. (2001, 2006)	Lilith and Dogancay (2004)	Tao et al. (2008, 2009)	Felice et al. (2010)	Galindo-Serrano and Giupponi (2010b)	Venkatraman et al. (2010)	Yang and Grace (2009)	Yau et al. (2010d)
Lower blocking probability		×	×	×				×	
Higher throughput	×				×		×		×
Higher user satisfaction probability	×								
Lower end-to-end delay					×				
Lower outage probability of licensed users						×			
Lower number of channel switches								×	×

Table 5
A summary of the application of RL in wireless networks.

Scheme	Type of wireless network	Performance enhancement compared to traditional schemes	Reference
Cooperative communications	Ad hoc	Higher number of successful packet transmissions per transmitted power	Naddafzadeh-Shirazi et al. (2009, 2010)
	Wireless sensor	Higher throughput (or higher packet delivery rate or lower packet loss rate) Lower end-to-end delay	Liang et al. (2009)
Rate adaptation	Single link	Higher throughput	Joshi et al. (2008)
Scheduling	Cellular	Higher throughput Lower end-to-end delay	Yu et al. (2007)
	Wireless personal area	Lower decoding failure rate	Moradi et al. (2007)
Power management	Ad hoc	Without third-party involvement in power control	Long et al. (2007)
	Single link	Lower number of backlogged packets	Vengerov et al. (2005)
	Wireless sensor	Higher throughput (or lower packet loss rate) per total consumed energy	Pandana and Liu (2004, 2005) Udenze and McDonald-Maier (2009)
		Lower energy consumption	Sridhar et al. (2007)
Security management	Ad hoc	Higher throughput	Lee et al. (2009) Usaha and Karnkamon (2005) Usaha and Maneenil (2006)
Queue management	Ad hoc	Lower packet dropping rate	Zhou et al. (2009)
Medium access control	Cognitive radio	Higher throughput (or lower packet loss rate)	Li et al. (2010)
	Wireless sensor	Lower energy consumption	Gummeson et al. (2010)
		Higher throughput and lower energy consumption	Liu et al. (2006)
Network coding	Cognitive radio	Higher throughput	Chen et al. (2009)
	Ad hoc	Faster code construction rate	Jabbarihagh and Lahouti (2007)
Service discovery	Ad hoc	Higher average bandwidth savings	Gonzalez-Valenzuela et al. (2008)

of fixed network infrastructure. Each host sends and forwards packets from its neighbors to upstream neighbors until the destination is reached. The objective is to maximize the proportion of packets successfully transmitted to the destination from the source in order to mitigate end-to-end errors caused by packet drop and buffer overflow. The work of Arroyo-Valles et al. (2007) is an example of RL-based routing, and it is discussed in this section. Similar RL-based routings are Shiang and Schaar (2010), Chang et al. (2004) and Dong et al. (2007).

In general, there are two main challenges faced by a node (or agent) to choose its upstream neighbor node (or action). Firstly, routing overhead, which conveys the information about a path, must be minimized to alleviate congestion. Secondly, with reduced routing information, there is partial observability. This means that when making an end-to-end routing decision, a node can only rely on messages received from its neighbor nodes.

3.1.1. Reinforcement learning model

A node chooses its upstream node based on the Q-values that are accumulated along a path. The Q-value is the cumulative link costs of a path, such as end-to-end delay. Traditional routing protocols that use minimum hop counts as a metric would have each link cost equal to one. The RL approach conveys the quality of a path, such as queuing delay and link reliability, using a Q-value. The SARL approach is applied. The state, action and reward representations are described below.

State: a node represents each of its destination nodes as a state. This is normally an entry in the routing table.
Action: a node represents all of its possible next hops for relay of a received packet as a set of actions. Thus, the cardinality of the set of actions equals the number of the node's adjacent neighbors.

Reward: a node represents the expected end-to-end delivery time as the *Q*-value. Higher negative *Q*-value indicates shorter estimated end-to-end delay; higher negative immediate reward indicates shorter estimated queuing delay at a node itself; higher negative discounted reward indicates shorter end-to-end delay from upstream neighbor to destination.

Eqs. (1) and (2) are applied in this approach. The *Q*-value *Q*(*state*, *action*) represents the estimated amount of end-to-end delay in order to transmit a packet to the destination *state* through a next hop *action*. The *event* in Eq. (1) is not represented. A node chooses its next hop with the highest *Q*-value that provides the minimum amount of end-to-end delay to the destination. This maximizes the proportion of packets successfully transmitted to the destination through avoiding unreliable links, dropped packets and buffer overflows or congestion throughout the path.

An important requirement is to select a route that fulfills a flow's QoS requirement, which is achieved using rules. For instance, only routes with *Q*-value greater than a threshold that provides the maximum allowable end-to-end delay are chosen as the candidate paths. To ensure link reliability, two nodes become 'practical' neighbors only if the delivery probability between them is above a predefined threshold. The delivery probability can be approximated based on the signal strengths or statistics, which can be estimated using the number of messages sent and received between the nodes.

3.1.2. Achievements of the RL model

Applying RL in routing has been shown in Arroyo-Valles et al. (2007), Chang et al. (2004) and Dong et al. (2007) to improve network performance such as providing a lower end-to-end delay for packet and a lower packet loss rate compared to traditional routing schemes such as ad hoc on-demand distance vector (AODV).

3.2. Example 2: RL-based resource management in centralized mobile networks

A centralized mobile network is a centralized single-hop wireless network established by fixed network infrastructure, such as a base station. A host moves from one cell to another, resulting in bandwidth fluctuation and scarcity within a base station. Yu et al. (2008) use RL applied to the provision of adaptive multimedia services in centralized networks with mobile hosts, and this approach is discussed in this section. The objective is to maximize network utilization, and hence revenue, by admitting new calls without jeopardizing the QoS of existing ongoing calls. There are two components in the scheme, namely, bandwidth adaptation, which adjusts the bandwidth allocation of individual ongoing calls, and call admission control, which decides whether to admit or reject new and handoff calls.

There are a number of service classes associated with each call. Within each class, there are a number of discretized bandwidth levels that bandwidth adaptation can choose from for a particular call. A call is served with un-degraded service if it is allocated with the highest bandwidth level within its service class. Different service classes and bandwidth levels generate different amount of revenue. During congestion, call admission control rejects calls and/or bandwidth adaptation degrades the bandwidth level of the calls. When a call releases its allocated bandwidth due to call completion or handoff to another cell, bandwidth adaptation increases the bandwidth level of its ongoing calls. To maintain the QoS of ongoing calls, handoff calls are given higher priority than new calls.

3.2.1. Reinforcement learning model

The main task is to determine whether a call from a particular service class is accepted and which call(s) should have their bandwidth changed. The RL approach is embedded in each base station. The SARL approach is applied. The state, event, action and reward representations are described below:

State–event pair: the state is the number of each type of ongoing call in a cell. For example, x_{ij} represents the number of ongoing calls of service class i using bandwidth b_{ij} in a cell. The possible events are a new call arrival, a handoff call arrival, a call termination, and a call handoff to a neighboring cell.
Action: there are three types of actions. Firstly, the admission action accepts or rejects a new or handoff call. Secondly, the set of actions of bandwidth degradation when a call is accepted is represented as $\left\{d^n_{(i,j)}\right\}$. The $d^n_{(i,j)}$ indicates the number of ongoing calls of service class i using bandwidth b_{ij} being degraded to b_{in}. Thirdly, the set of actions of bandwidth upgrade when a call is terminated or there is a handoff to a neighboring cell is represented as $\{u^n_{(i,j)}\}$. The $u^n_{(i,j)}$ indicates the number of ongoing calls of service class i using bandwidth b_{ij} being upgraded to b_{in}. The state variable x_{ij} must be updated after performing the action.
Reward: the reward is the total revenue or reward rate generated by the ongoing calls from all service classes using different levels of bandwidth minus the cost of signaling overhead, which consumes bandwidth and energy, during a sojourn time. The sojourn time is the duration between decision epochs when changes of state or event occur.

The base station applies Eq. (1) to update the *Q*-value, where *Q*(*state*, *event*, *action*) represents the estimated revenue of ongoing calls when an action is taken given a state–event pair. It chooses the action with the maximum *Q*-value to maximize its revenue following Eq. (2).

Several rules are imposed to achieve a certain level of QoS. Firstly, the total amount of bandwidth consumption within a cell must be less than the channel capacity. Secondly, the handoff dropping probability must be less than a certain threshold. Thirdly, the proportion of ongoing calls that receive un-degraded service must be greater than a particular threshold.

3.2.2. Achievements of the RL model

Applying RL has been shown in Yu et al. (2008) and Alexandri et al. (2002a) to improve several network performance metrics such as the amount of reward for different levels of new call arrival rate at each cell compared to existing schemes such as the Guard Channel scheme (Yu et al., 2008), and hence increases the revenue of the network operator, potentially optimizing return on investment.

3.3. Example 3: RL-based dynamic channel selection in distributed cognitive radio networks

Traditional static spectrum allocation policies grant each wireless service exclusive usage of certain frequency bands, leaving several bands unlicensed, such as the industrial, scientific, and medical (ISM), bands for general purposes. The tremendous growth in wireless applications that utilize the unlicensed frequency bands has caused spectrum scarcity in those bands. Cognitive radio (CR) (Mitola and Maguire, 1999) enables unlicensed spectrum users or secondary users (SUs) to exploit underutilized licensed spectrum (or white space) to optimize the utilization of the overall radio spectrum conditional on the interference to the licensed spectrum users or primary users (PUs) being below an acceptable level.

The white space is defined by time, frequency and maximum transmission power at a particular location. To alleviate collision with PU transmissions, using CR, data packets are allocated opportunistically to white space at different channels by changing transmission and reception operating channels.

The objective of the DCS scheme (Yau et al., 2010d) is to maximize the probability of successful packet transmission, and hence throughput, for each SU transmission pair in distributed CRNs. Some RL-based DCS schemes are described in Tao et al. (2008) and Yang and Grace (2009). A distributed CRN is a distributed single-hop wireless network established by a number of SU transmission pairs in the absence of fixed network infrastructure. The probability of successful packet transmission is dependent on many factors including the PU channel utilization level and packet error rate in the channel.

3.3.1. Reinforcement learning model

The main task for a sender SU is to choose a channel that maximizes the probability of successful packet transmission for data transmission. The multiple-hop coordination-based MARL approach is applied. The state, action and reward representations are described below:

State: a sender SU represents each of its neighbor nodes, such as their respective identification number, by a state. Hence, a SU sender can choose a different action for a different neighbor SU.
Action: the action is to choose an available channel out of a set of operating channels.
Reward: the reward is throughput; specifically, it is the number of successful data packet transmission achieved by a SU transmission pair within an epoch. Data packet transmission is successful when a link layer acknowledgment is received for a data packet sent. Throughput level is chosen as the reward because it is a good measure of contention level in addition to PU channel utilization level and channel quality. For instance, high levels of reward indicate low levels of contention and vice versa. To further explain the concept, consider a situation where all the SU pairs choose a similar data channel with low PU channel utilization level but high channel quality for data transmission. The Q-value of the chosen channel for all the SU pairs would be low due to high contention level.

Each SU transmission pair applies Eq. (1) to update the Q-value, where $Q(state, action)$ represents the estimated throughput of a chosen channel. It chooses the action with the maximum local reward using Eq. (4) to maximize network-wide throughput. As each agent does not include its own historical Q-value in its local reward computation, the local reward does not increase without bound in a cyclic topology (Yau et al., 2010c).

3.3.2. Achievements of the RL model

Applying multiple-hop coordination-based MARL has been shown in Yau et al. (2010d) to increase network-wide throughput and decrease the number of channel switches compared to the traditional SARL approach. The number of channel switches is an important criterion for reducing transmission delay and energy consumption.

4. Implementation of reinforcement learning in wireless platforms

This section provides a summary of RL implementations in wireless platforms, as well as a case study on an implementation in CR platform.

4.1. A summary of implementations

There is a limited research in the literature regarding the implementation of RL in wireless platform. Table 6 provides a list of references for various types of schemes in wireless networks, their respective type of network as well as the wireless platform that the scheme is applied, and the important performance enhancement brought about by the RL approach.

4.2. Implementation of reinforcement learning in cognitive radio platform: a case study

This section shows a case study on our implementation of a RL-based DCS scheme in the GNU radio platform (GNU Radio), and it is presented in Yu et al. (2010). A centralized CRN is a static and centralized single-hop wireless network established by fixed network infrastructure, such as a base station and static hosts. One of the major concerns of the RL approach is the convergence of the Q-value for each action in practice, and successful convergence is demonstrated in this section. Without convergence, the Q-values will fluctuate, and hence the optimal action will change from time to time. The objective of the DCS scheme is to maximize the probability of successful packet transmission, and hence throughput in static and centralized CRNs. Note that we discussed a RL-based DCS scheme for distributed CRNs in Section 3.3. The rest of this section provides discussions on the RL model of the DCS scheme, its implementation and the details of the implementation.

Table 6
A summary of the application of RL in wireless platforms.

Scheme	Reference	Type of wireless network	Wireless platform	Performance enhancement
Dynamic channel selection	Yu et al. (2010)	Cognitive radio	GNU Radio	Higher probability of successful packet transmission, and hence higher throughput in static and centralized CRN compared to random approach. The RL approach achieves the analytical and expected network performance
Medium access control	Gummeson et al. (2009, 2010)	Wireless sensor	TinyOS 2.0 node with CC2420 expansion board	RL-based adaptive link layer switches between radios depending on which radio offers better network performance, and it has been shown to provide lower energy consumption per successful packet transmission, percentage of packet loss, and cumulative energy consumption compared to existing radios in CC2420 and XE1205, respectively
Resource management	Istepanian et al. (2009)	Cellular	3.5 G (high speed downlink packet access, HSDPA)	Higher mean opinion score and peak signal-to-noise ratio compared to traditional rate control algorithm scheme, namely H.264
Application (sensing coverage)	Tham and Renaud (2005)	Wireless sensor	Crossbow Mica2 sensor mode	Two MARL approaches are applied, namely OptDRL and DVF. OptDRL provides comparatively higher degree of application-level performance, particularly higher sensing coverage; while DVF provides comparatively lower energy consumption

4.2.1. Reinforcement learning model

The state and action representations are similar to the RL approach for a distributed CRN in Section 3.3, and the reward representation for centralized CRNs is described below. The difference in the reward representation is due to the fact that, in centralized networks, it is not necessary to consider agent interaction and coordination, such as the effects of channel contention level, since there is a single agent only, which is the base station or the SU sender.

Reward: For every successful data packet transmission, there is a reward with positive constant value, otherwise a negative reward with negative constant value is incurred.

At each attempt to transmit a data packet, the SU sender chooses a channel for data transmission. The channels have different levels of PU channel utilization. Eq. (1) is rewritten as follows:

$$Q_{t+1}(a_t) \leftarrow (1-\alpha)Q_t(a_t)+\alpha r_{t+1}(a_t) \quad (5)$$

with the $\max_{a \in A} Q_t(s_{t+1},a)$ in Eq. (1) being omitted to indicate no dependency on the discounted rewards. The events and rules are not represented. It is assumed that there are two SUs, a sender and a receiver, to represent a centralized network. Hence, the state is not included in Eq. (5) as there is a single SU receiver only. A thorough investigation of scenarios involving multiple SU receivers is provided in Yau et al. (2010a). Similar trends are observed in state representation with single SU and multiple SUs, so the work shown in this section chooses to implement the scenario with a single SU receiver.

4.2.2. Implementation of the reinforcement learning model

The SU sender always has backlogged data ready to transmit. Both sender and receiver can communicate with each other using a channel chosen from the set of available channels. The PU traffic in each channel follows a Poisson process. There are three available channels at different frequency bands. An initial Q-value set of [0, 10, 5] indicates that channel 1, 2 and 3 are initialized to values 0, 10 and 5, respectively. A PU channel utilization level set of [0.9, 0.7, 0.2] indicates that channel 1, 2 and 3 have PU channel utilization levels of 0.9, 0.7 and 0.2, respectively.

Figure 6 shows that, using the initial Q-values of [0, 10, 5] and PU channel utilization levels of [0.9, 0.7, 0.2], the Q-values of the SU sender converge to certain values, hence making it possible for the RL approach to choose an optimal action as the number of SU transmission attempts increases. It is shown that the optimal action, which is channel 3, reaches the highest Q-value within the first 10 attempts. In Yu et al. (2010), it is also shown that the RL approach achieves significant throughput enhancement compared to the random approach, where the channel for data transmission is chosen randomly without learning.

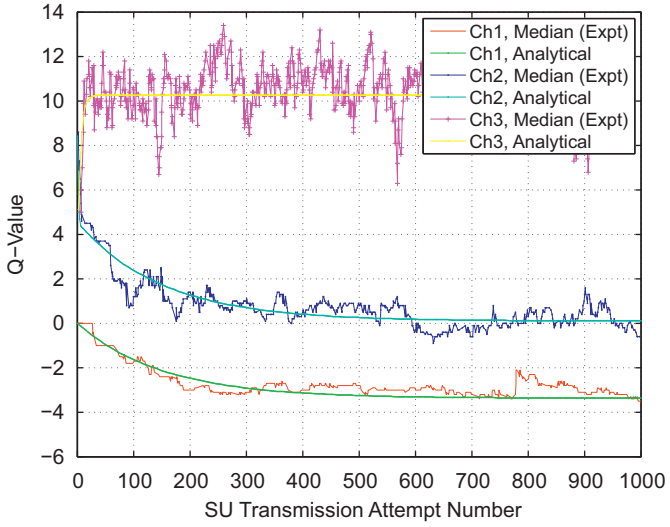

Fig. 6. Channel Q-values against number of SU transmission attempts (Yu et al., 2010). Both experimental and analytical results are shown.

4.2.3. Implementation description of the reinforcement learning model

To show how the RL approach works, we choose to explain the update of Q-value for channel 3, which Q-value is initialized with value of 5. When the SU sender chooses channel 3 to transmit a data packet to its receiver, the Q-value for this channel is updated. If a link layer acknowledgment is received for the data packet sent, using Eq. (5) with $\alpha=0.2$, reward of 15, and negative reward of –5, the update of the Q-value $Q_t(3)$ is $Q_{t+1}(3) \leftarrow 0.8 \times 5+0.2 \times 15=7$, otherwise the transmission is unsuccessful and the update of the Q-value is $Q_{t+1}(3) \leftarrow 0.8 \times 5+0.2 \times (-5)=3$. The Q-values are updated for every data packet transmission subsequently for the chosen channel during exploitation and exploration.

5. Open issues

In Section 2, we discussed possible future research directions relating to each element in the RL approach. This section discusses several open issues that can be pursued so that the application of RL can be extended to a wider range of schemes and to improve existing schemes that adopt the RL approach. Some open issues are:

- *Adaptation to an operating environment that undergoes abrupt changes.* Q-value provides an estimate of the reward for each action taken. As time goes by, the estimation improves so that an optimal action can be chosen. The learning rate α determines how quickly an agent adapts to the environment, with $\alpha=1$ replacing the learnt Q-value with the most recent estimated reward. In wireless networks, the environment may change abruptly. For instance, nodes may move in random directions at different speeds, may move out of range from each other, may be switched on and off, or an obstacle may exist. Thus, the level of α may be adjusted dynamically, and new *events* may be designed to detect occurrences that change abruptly so that appropriate action can be taken. For example, in MANETs, a mobile node may appear and disappear at any time so a rule is required to detect this event and set the Q-value to infinity when out of range and to an initial value when the node is in range.
- *Effects of exploration on stability*. In a multi-agent environment, exploration can contribute to instability since the agents switch their respective action from time to time. For instance, in the DCS scheme in distributed CRNs, exploration may cause a SU node pair to switch its operating channel constantly. There are two reasons for this. Firstly, when several agents explore at the same time, the Q-values become unstable and they do not portray the correct condition of the operating environment. For instance, in distributed CRNs, when two agents (SU node pairs) explore a particular action (channel), the Q-value (throughput) for the channel reduces for all agents and does not portray the exact level of contention (Yau et al., 2010b). Secondly, an agent that explores a particular action, and then exploits the other one in the following epoch causes the Q-values of both actions in itself and its neighbor agents to fluctuate. Adopting the assumption of a single collision domain in which all the agents can hear each other or within the communication range of each other, the RL approach in

Yau et al. (2010b) ensures an agent explores and updates its Q-values at the right time, and this has been shown to improve network stability by reducing the number of channel switches. The scenario that does not apply the assumption of a single collision domain remains an open issue.

- *Achieving short-term network performance enhancement*. The single-agent and multi-agent approaches provide long-term performance enhancement, but *short-term* enhancement (Barrett et al., 2002) may be necessary. For example, throughput and delay performance enhancement may need to be achieved in a short time frame to provide QoS guarantees to high priority data packets. Another example is the DCS scheme in distributed CRNs. Since there are many SUs competing for the data channels, short-term fairness may be necessary to provide fair throughput among the SUs.
- *Model-based single-hop and multiple-hop coordination-based MARL*. Model-based RL builds and constantly updates an internal model of the operating environment such as state transition probability for each possible pair of states, and makes decisions on action selection based on the model. The purpose is to improve the convergence rate to an optimal or near-optimal joint action. However, the drawbacks are computation complexity and higher memory requirement for storage of the model, especially as the number of states increases. In Dowling et al. (2005), Fu and Schaar (2009), Littman et al. (1995) and Shiang and Schaar (2010), an agent builds a state transition model using statistical information, such as the number of packets required for a successful transmission to its next hop (Dowling et al., 2005), and number of times a state transition of s_t^i to s_{t+1}^i occurs when action a_t^i is taken (Fu and Schaar, 2009; Shiang and Schaar, 2010). Additionally, in Shiang and Schaar (2010), the immediate reward $r_{t+1}^i(s_{t+1}^i)$, which is the queueing delay at agent i itself, may be estimated and computed based on the M/G/1 queuing model. Instead of updating the Q-value for a single state–action pair of the traditional RL approach, Q-values for any state–action pairs are updated based on the estimated state transition probability and immediate reward in Shiang and Schaar (2010), and this has been shown to enhance the convergence rate. Future research could be pursued for other models, and to address the drawbacks.
- *Achieving stable Q-values*. This includes detecting and responding to any fluctuations in the Q-values due to unforeseen circumstances. This enables the Q-values to converge to optimal or near-optimal action in single-agent and multi-agent approaches.
- *Effects of irrational agents on learning outcome*. In a multi-agent environment, several irrational agents, which may be caused by low residual energy or other factors, that take suboptimal or random actions may result in instability throughout the network. The irrational agents may affect the learning outcome of the rational agent. New events may be designed to detect and respond to irrational or malicious agents so that appropriate action can be taken as soon as possible.
- *Achieving heterogeneous learning environment*. In a multi-agent environment, each agent may represent the Q-values with different performance metrics in a particular network to enable heterogeneous learning objectives. As long as the optimal or near-optimal global Q-value is achieved, all agents may achieve their respective objectives.

6. Conclusions

In this article, we advocate the use of reinforcement learning (RL) to achieve context awareness and intelligence in wireless networks. In general, context awareness and intelligence enable each host to observe, learn, and respond appropriately in an efficient manner with respect to its complex and dynamic operating environment without adhering to a strict and static predefined set of rules. This capability is of paramount importance for general functionality and performance enhancement in various kinds of wireless networks including mobile ad hoc networks, wireless sensor networks, and the next generation wireless networks such as cognitive radio networks. RL has been successfully applied to routing, resource management, dynamic channel selection and other network functions demonstrating significant performance enhancement. RL has been shown to achieve performance enhancement for dynamic channel selection both in simulation and real implementation on a cognitive radio network platform. Hence, RL is an effective approach for achieving context awareness and intelligence in wireless networks. Through the definition of some optional elements including *state*, *action* and *reward*, RL is a suitable solution for many problems in wireless networks. Existing schemes that apply RL can be further enhanced using additional features not used in traditional RL including *events*, *rules*, and *agent interaction and coordination*. Two kinds of RL approaches are single-agent RL and multi-agent RL (MARL). Two categories of MARL are single-hop coordination-based MARL and multiple-hop coordination-based MARL. Certainly, there is a great deal of future work in the use of RL including the designs of events and rules, and multi-agent approaches, as well as the open issues raised in this paper.

References

Alexandri E, Martinez G, Zeghlache D. A distributed reinforcement learning approach to maximize resource utilization and control handover dropping in multimedia wireless networks. In: Proceedings of the PIMRC'02 13th international symposium on personal, indoor and mobile radio comm, Portugal, 2002a. p. 2249–53.

Alexandri E, Martinez G, Zeghlache D. Adaptive joint call admission control and access network selection for multimedia wireless systems. In: Proceedings of the WPMC'02 5th international symposium on wireless personal multimedia communications, Hawaii, US, vol. 3, 2002b. p. 1390–94.

Alexandri E, Martinez G, Zeghlache D. An intelligent approach to partition multimedia traffic onto multiple radio ccess networks. In: Proceedings of the VTC'02-Fall IEEE 56th vehicular technology conference, Vancouver, Canada, 2002c. p. 1086–90.

Arroyo-Valles R, Alaiz-Rodriguez R, Guerrero-Curieses A, Cid-Sueiro J. Q-Probabilistic routing in wireless sensor networks. In: Proceedings of the ISSNIP'07 third international conference on intelligent sensors, sensor network and information processing, Australia, 2007.

Barrett CL, Marathe MV, Engelhart DC, Sivasubramaniam A. Analyzing the short-term fairness of IEEE 802.11 in wireless multi-hop radio networks. In: Proceedings of the MASCOTS'02 10th international symposium on modeling, analysis and simulation of computers and telecommunication systems, USA, 2002. p. 137–44.

Baruah P, Urgaonkar R. Learning-enforced time domain routing to mobile sinks in wireless sensor fields. In: Proceedings of the LCN'04 29th annual IEEE international conference on local computer network, Tampa, FL, 2004.

Bernardo F, Agusti R, Perez-Romero J, Sallent O. A novel framework for dynamic spectrum management in multicell OFDMA networks based on reinforcement learning. In: Proceedings of the WCNC'09 wireless communications and network conference, Hungary, 2009a.

Bernardo F, Agusti R, Perez-Romero J, Sallent O. Distributed spectrum management based on reinforcement learning. In: Proceedings of the CROWNCOM'09 fourth international conference on cognitive radio oriented wireless network and communications, Hannover, Germany, 2009b.

Bianchi G. Performance analysis of the IEEE 802.11 distributed coordination function. IEEE Journal on Selected Areas in Communications 2000;18(3):535–47.

Bing X, Wahab MH, Yang Y, Zhong F, Sooriyabandara M. Reinforcement learning based spectrum-aware routing in multi-hop cognitive radio networks. In: Proceedings of the CROWNCOM'09 fourth international conference on cognitive radio oriented wireless network and communications, Germany, 2009.

Chang YH, Ho T, Kaelbling LP. Mobilized ad-hoc networks: a reinforcement learning approach. In: Proceedings of the ICAC'04 international conference on autonomic computing, USA, 2004.

Chanloha P, Usaga W. Call admission control in wireless DS-CDMA systems using actor-critic reinforcement learning. In: Proceedings of the ISWPC'07 second international symposium on wireless pervasive computing, San Juan, Puerto Rico, 2007.

Chen X, Zhao Z, Zhang H, Jiang T, Grace D., Inter-cluster connection in cognitive wireless mesh networks based on intelligent network coding. In: Proceedings of PIMRC'09 20th international symposium on personal, indoor and mobile radio communications, Japan, 2009. p. 1251–6.

Curran E, Dowling J. SAMPLE: statistical network link modeling in an on-demand probabilistic routing protocol for ad hoc networks. In: Proceedings of WONS'05 second annual conference on wireless on-demand network systems and services, Switzerland, 2005. p. 200–5.

Dong S, Agrawal P, Sivalingam K. Reinforcement learning based geographic routing protocol for UWB wireless sensor network. In: Proceedings of the GLOBECOM'07 global telecommunications conference, USA, 2007. p. 652–6.

Dowling J, Curran E, Cunningham R, Cahill V. Using feedback in collaborative reinforcement learning to adaptively optimize MANET routing. IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems and Humans 2005;35(3):360–72.

Egorova-Forster A, Murphy AL. Exploiting reinforcement learning for multiple sink routing in WSNs. In: Proceedings of the MOBHOC'07 IEEE international conference on mobile adhoc and sensor systems, 2007. p. 652–6.

El-Alfy ES, Yao YD, Heffes H. A model-based Q-learning scheme for wireless channel allocation with prioritized handoff. In: Proceedings of GLOBECOM'01 IEEE global telecommunications conference, San Antonio, 2001. p. 3668–72.

El-Alfy ES, Yao YD, Heffes H. A learning approach for prioritized handoff channel allocation in mobile multimedia networks. IEEE Transactions on Wireless Communications 2006;5(7):1651–60.

Felice MD, Chowdhury KR, Meleis W, Bononi L. To sense or to transmit: a learning-based spectrum management scheme for cognitive radio mesh networks. In: Proceedings of WIMESH'10 fifth IEEE workshop on wireless mesh network, Boston, MA, 2010.

Forster A, Murphy A. FROMS: feedback routing for optimizing multiple sinks in WSN with reinforcement learning. In: Proceedings of ISSNIP'07 third international conference on intelligent sensors, sensor network and information processing, 2007. p. 371–6.

Forster A, Murphy A. CLIQUE: role-free clustering with Q-learning for wireless sensor networks. In: Proceedings of ICDCS 29th IEEE international conference on distributed computing systems, 2009. p. 441–9.

Fu F, Schaar Mvd. Learning to compete for resources in wireless stochastic games. IEEE Transactions on Vehicular Technology 2009;58(4):1904–19.

Fu P, Li J, Zhang D. Heuristic and distributed QoS route discovery for mobile ad hoc networks. In: Proceedings of CIT'05 fifth international conference on computer and information technology, Shanghai, China, 2005.

Galindo-Serrano A, Giupponi L. Distributed Q-learning for aggregated interference control in cognitive radio networks. IEEE Transactions on Vehicular Technology 2010a;59(4):1823–34.

Galindo-Serrano A, Giupponi L. Decentralized Q-learning for aggregated interference control in completely and partially observable cognitive radio networks. In: Proceedings of CCNC'10 seventh IEEE consumer communications and networking conference, 2010b.

Gelenbe E, Gellman M. Can routing oscillations be good? The benefits of route-switching in self-aware networks. In: Proceedings of MASCOTS 15th international symposium modeling, analysis, and simulation of computer and telecommunication system, Istanbul, 2007. p. 343–52.

Giupponi L, Agusti R, Perez-Romero J, Roig OS. A novel approach for joint radio resource management based on fuzzy neural methodology. IEEE Transactions on Vehicular Technology 2008;57(3):1789–805.

Giupponi L, Galindo-Serrano AM, Dohler M. From cognition to docition: the teaching radio paradigm for distributed & autonomous deployments. Computer Communications 2010;33(17):2015–20.

GNU Radio. Available 〈http://gnuradio.org/trac〉.

Gonzalez-Valenzuela S, Vuong ST, Leung VCM. A mobile-directory approach to service discovery in wireless ad hoc networks. IEEE Transactions on Mobile Computing 2008;7(10):1242–56.

Gummeson J, Ganesan D, Corner MD, Shenoy P. An adaptive link layer for range diversity in multi-radio mobile sensor networks. In: Proceedings of INFOCOM'09, Rio de Janeiro, 2009. p. 154–62.

Gummeson J, Ganesan D, Corner MD, Shenoy P. An adaptive link layer for heterogeneous multi-radio mobile sensor networks. IEEE Journal on Selected Areas in Communications 2010;28(7):1094–104.

Hu T, Fei Y. QELAR: a machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks. IEEE Transactions on Mobile Computing 2010;9(6):796–809.

Henkel D, Brown TX. Towards autonomous data ferry route design through reinforcement learning. In: Proceedings of the WoWMoM'08 international symposium on a world of wireless, mobile and multimedia networks, USA, 2008.

Istepanian RSH, Philip NY, Martini MG. Medical QoS provision based on reinforcement learning in ultrasound streaming over 3.5 G wireless systems. IEEE Journal on Selected Areas in Communications 2009;27(4):566–74.

Jabbarihagh M, Lahouti F. A decentralized approach to network coding based on learning. In: Proceedings of the ITWITWN'07 workshop on infomation theory for wireless networks, Solstrand, 2007.

Joshi T, Ahuja D, Singh D, Agrawal DP. SARA: stochastic automata rate adaptation for IEEE 802.11 networks. IEEE Transactions on Parallel and Distributed Systems 2008;19(11):1579–90.

Kok JR, Vlassis N. Collaborative multiagent reinforcement learning by payoff propagation. Journal of Machine Learning Research 2006;7:1789–828.

Lee M, Ye X, Johnson S, Marconett D, Chaitanya VSK, Vemuri R, Yoo SJB. Cognitive security management with reputation based cooperation schemes in heterogeneous networks. In: Proceedings of the CICS'09 symposium on computational intelligence in cyber security, USA, 2009. p. 19–23.

Li H, Grace D, Mitchell PD. Collision reduction in cognitive radio using multi-channel 1-persistent CSMA combined with reinforcement learning. In: Proceedings of the CROWNCOM'10 fifth international conference on cognitive radio oriented wireless network and communications, France, 2010.

Liang X, Balasingham I, Byun SS. A reinforcement learning base routing protocol with QoS support for biomedical sensor networks. In: Proceedings of the ISABEL'08 first international symposium applied sciences on biomedical and communication technology, Aalborg, Denmark, 2008a.

Liang X, Balasingham I, Byun SS. A multi-agent reinforcement learning based routing protocol for wireless sensor networks. In: Proceedings of the ICWCS'08 IEEE international symposium on wireless communication systems, Reykjavik, Iceland, 2008b. p. 552–7.

Liang X, Balasingham I, Leung VCM. Cooperative communications with relay selection for QoS provisioning in wireless sensor networks. In: Proceedings of the GLOBECOM'09 IEEE global telecommunication conference, Honolulu, Hawaii, 2009.

Lilith N, Dogancay K. Dynamic channel allocation for mobile cellular traffic using reduced-state reinforcement learning. In: Proceedings of the WCNC'04 IEEE wireless communications and networking conference, Atlanta, GA, 2004.

Lilith N, Dogancay K. Reinforcement learning-based dynamic guard channel scheme with maximum packing for cellular telecommunications systems. In: Proceedings of the WiCom'07 international conference on wireless communications, networking and mobile computing, 2007.

Littman ML, Cassandra AR, Kaelbling LP. Learning policies for partially observable environments: scaling up. In: Proceedings of the ICML'95 12th international conference on machine learning, USA, 1995. p. 362–70.

Liu NX, Zhou X, Baras JS. Adaptive hierarchical resource management for satellite channel in hybrid MANET-satellite–internet network. In: Proceedings of the VTC2004-Fall IEEE 60th vehicular technology conference, LA, USA, vol. 6, 2004. p. 4027–31.

Liu Z, Elhanany I. RL-MAC: a QoS-aware reinforcement learning based MAC protocol for wireless sensor networks. In: Proceedings of the ICNSC international conference on networking, sensing and control, USA, 2006. p. 768–73.

Long C, Zhang Q, Li B, Yang H, Guan X. Non-cooperative power control for wireless ad hoc networks with repeated games. IEEE Journal on Selected Areas in Communications 2007;25(6):1101–11.

Martinez-Bauset J, Gimenez-guzman JM, Pla V. Optimal admission control in multimedia mobile networks with handover prediction. IEEE Wireless Communications 2008;15(5):38–44.

Mitola III J, Maguire GQJ. Cognitive radio: making software radios more personal. IEEE Personal Communications 1999;6(4):13–8.

Moradi S, Rad AHM, Wong VWS. A novel scheduling algorithm for video traffic in high-rate WPANs. In: Proceedings of the GLOBECOM'07 global telecommunications conference, 2007. p. 742–7.

Murphy KP. A survey of POMDP solution techniques. Technical Report, University of Carlifornia Berkeley, USA, 2000. [Online]. Available: 〈http://www.cs.ubc.ca/~murphyk/mypapers.html〉.

Naddafzadeh-Shirazi G, Kong PY, Tham CK. Cooperative retransmissions using Markov decision process with reinforcement learning. In: Proceedings of the PIMRC'09 IEEE 20th international symposium on personal, indoor and mobile radio communications, Tokyo, Japan, 2009.

Naddafzadeh-Shirazi G, Kong PY, Tham CK. Distributed reinforcement learning frameworks for cooperative retransmission in wireless networks. IEEE Transactions on Vehicular Technology 2010;59(8):4157–62.

Niyato D, Hossain E. Dynamics of network selection in heterogeneous wireless networks: an evolutionary game approach. IEEE Transactions on Vehicular Technology 2009;58(4):2008–17.

Ouferhat N, Mellouk A. Energy and delay efficient state dependent routing algorithm in wireless sensor networks. In: Proceedings of the LCN ninth IEEE international workshop on wireless local networks, Zurich, Switzerland, 2009.

Pandana C, Liu KJR. A near-optimal reinforcement learning scheme for energy efficient point-to-point wireless communications. In: Proceedings of the GLOBECOM'04 global telecommunications confernce, vol. 2, 2004. p. 763–7.

Pandana C, Liu KJR. Near-optimal reinforcement learning framework for energy-aware sensor communications. IEEE Journal on Selected Areas in Communications 2005;23(4):788–97.

Renaud JC, Tham CK. Coordinated sensing coverage in sensor networks using distributed reinforcement learning. In: Proceedings of the ICON 14th IEEE international conference on networks, 2006.

Salodkar N, Karandikar A, Borkar VS. A stable online algorithm for energy-efficient multiuser scheduling. IEEE Transactions on Mobile Computing 2010;9(10):1391–406.

Saoseng JY, Tham CK. Coordinated rate control in wireless sensor network. In: Proceedings of the ICCS'06 10th IEEE Singapore international conference on communication systems, Singapore, 2006.

Seah MWM, Tham CK, Srinivasan V, Xin A. Achieving coverage through distributed reinforcement learning in wireless sensor networks. In: Proceedings of the ISSNIP third international conference on intelligent sensors, sensor network and information processing, Australia, 2007. p. 425–30.

Shiang HP, Schaar Mvd. Online learning in autonomic multi-hop wireless networks for transmitting mission-critical applications. IEEE Journal on Selected Areas in Communications 2010;28(5):728–41.

Sridhar P, Nanayakkara T, Madni AM, Jamshidi M. Dynamic power management of an embedded sensor network based on actor-critic reinforcement based

learning. In: Proceedings of the ICIAFS third international conference on information and automation for sustainability, Australia, 2007. p. 76–81.
Sutton RS, Barto AG. Reinforcement Learning: an Introduction. MIT Press, Cambridge, MA,; 1998.
Tao J, Grace D, Liu Y. Cognitive radio spectrum sharing schemes with reduced spectrum sensing requirements. In: Proceedings of the IET seminar cognitive radio and software defined radios: technology and techniques, UK, 2008.
Tao J, Grace D, Mitchell PD. Improvement of pre-partitioning on reinforcement learning based spectrum sharing. In: Proceedings of the CCWMC'09 IET international communication conference on wirelesss mobile and computing, UK, 2009.
Tao T, Tagashira S, Fujita S. LQ-routing protocol for mobile ad-hoc networks. In: Proceedings of the ICIS'05 fourth annual ACIS international conference on computer and information science, South Korea, 2005.
Tham CK, Renaud JC. Multi-agent systems on sensor networks: a distributed reinforcement learning approach. In: Proceedings of the ISSNIP'05 international conference on intelligent sensors, sensor network and information processing, Melbourne, Australia, 2005.
Udenze A, McDonald-Maier K. Direct reinforcement learning for autonomous power configuration and control in wireless networks. In: Proceedings of the AHS'09 NASA/ESA conference on adaptive hardware and systems, San Francisco, CA, 2009.
Usaha W. A reinforcement learning approach for path discovery in MANETs with path caching strategy. In: Proceedings of the ISWCS'04 first international symposium on wireless communication systems, Mauritius, 2004. p. 220–4.
Usaha W, Karnkamon M. Preventing malicious nodes in ad hoc networks using reinforcement learning. In: Proceedings of the ISWCS'05 second international symposium on wireless communication systems, Siena, Italy, 2005.
Usaha W, Maneenil K. Identifying malicious nodes in mobile ad hoc networks using a reputation scheme based on reinforcement learning. In: Proceedings of the TENCON'06 IEEE region 10 conference, Hong Kong, China, 2006.
Vengerov D, Bambos N, Berenji HR. A fuzzy reinforcement learning approach to power control in wireless transmitters. IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics 2005;35(4):768–78.
Venkatraman P, Hamdaoui B, Guizani M. Opportunistic bandwidth sharing through reinforcement learning. IEEE Transactions on Vehicular Technology 2010;59(6):3148–53.
Vucevic N, Perez-Romero J, Sallent O, Agusti R. Reinforcement learning for active queue management in mobile all-IP networks. In: Proceedings of the PIMRC 18th international symposium on personal, indoor and mobile radio communications, Athens, 2007.
Vucevic N, Perez-Romero J, Sallent O, Agusti R. Joint radio resource management for LTE-UMTS coexistence scenarios. In: Proceedings of the PIMRC 20th international symposium on personal, indoor and mobile radio communications, Japan, 2009.
Xue Y, Lin Y, Feng Z, Cai H, Chi C. Autonomic joint session scheduling strategies for heterogeneous wireless networks. In: Proceedings of the WCNC'08 wireless communications and networking conference, Las Vegas, 2008.
Yagan D, Tham CK. Adaptive QoS provisioning in wireless ad hoc networks: a semi-MDP approach. In: Proceedings of the WCNC'05 IEEE wireless communications and networking conference, New Orleans, LA, vol. 4, 2005. p. 2238–44.
Yang M, Grace D. Cognitive radio with reinforcement learning applied to heterogeneous multicast terrestrial communication systems. In: Proceedings of the CROWNCOM fourth international conference on cognitive radio oriented wireless networks and communications, Germany, 2009.
Yau KLA, Komisarczuk P, Teal PD. Achieving context awareness and intelligence in cognitive radio networks using reinforcement learning for stateful applications. Technical Report ECSTR10-01, Victoria University of Wellington, New Zealand, 2010a.
Yau KLA, Komisarczuk P, Teal PD. Context awareness and intelligence in distributed cognitive radio networks: a reinforcement learning approach. In: Proceedings of the AusCTW 11th Australian communications theory workshop, Australia, 2010b. p. 35–42.
Yau KLA, Komisarczuk P, Teal PD. Achieving efficient and optimal joint action in distributed cognitive networks using payoff propagation. In: Proceedings of the ICC international conference communications, South Africa, 2010c.
Yau KLA, Komisarczuk P, Teal PD. Enhancing network performance in distributed cognitive radio networks using single-agent and multi-agent reinforcement learning. In: Proceedings of the LCN 35th conference on local computer networks, USA, 2010d.
Yu FR, Wong VWS, Leong VCM. A new QoS provisioning method for adaptive multimedia in cellular wireless networks. In: Proceedings of the INFOCOM'04 23rd annual joint conference of the IEEE computer and communications societies, Hong Kong, 2004a.
Yu FR, Wong VWS, Leong VCM., Efficient QoS provisioning for adaptive multimedia in mobile communication networks by reinforcement learning. In: Proceedings of the BROADNETS'04 first international conference on broadband networks, San Jose, CA, 2004b.
Yu FR, Wong VWS, Leong VCM. A new QoS provisioning method for adaptive multimedia in wireless networks. IEEE Transactions on Vehicular Technology 2008;57(3):1899–909.
Yu R, Sun Z, Mei S. Packet scheduling in broadband wireless networks using neuro-dynamic programming. In: Proceedings of the VTC'07-Spring IEEE 65th vehicular technology conference, Dublin, 2007. p. 2776–80.
Yu R, Dmochowski P, Komisarczuk P. Analysis and implementation of reinforcement learning on a GNU radio cognitive radio platform. In: Proceedings of the CROWNCOM fifth international conference on cognitive radio oriented wireless network and communications, France, 2010.
Zaine S, Mellouk A. Dynamic routing optimization based on real time adaptive delay estimation for wireless networks. In: Proceedings of the ISCC'10 IEEE symposium on computers and communications, Riccione, Italy, 2010.
Zhang Y, Fromherz M. Constrained flooding: a robust and efficient routing framework for wireless sensor networks. In: Proceedings of the AINA'06 20th international conference on advanced information networking and applications, Vienna, Austria, 2006.
Zhou Y, Yun M, Kim T, Arora A, Choi HA. RL-based queue management for QoS support in multi-channel multi-radio mesh networks. In: Proceedings of the NCA eighth IEEE international symposium on network computing and applications, USA, 2009. p. 306–9.