Feedback-Driven Interactive Learning in Dynamic Wireless Resource Management for Delay Sensitive Users

Hsien-Po Shiang and Mihaela van der Schaar

Department of Electrical Engineering (EE), University of California Los Angeles (UCLA) Los Angeles, CA
{hpshiang, mihaela} @ee.ucla.edu

[*]**Abstract – In this paper, we study the problem of dynamic resource management for delay sensitive users over wireless networks. We focus on a decentralized setting, where autonomous users make self-interested decisions in order to maximize their utility functions as evaluated based on information feedback. In this paper, two types of information feedback are discussed. One is the *private information feedback* between a transmitter-receiver pair. The other is the *public information feedback* among users (i.e. different transmitter-receiver pairs). Due to the informationally-decentralized nature of the wireless network, a user cannot have complete information about the transmission actions of its interfering neighbors. However, the user can model implicitly or explicitly the transmission strategies of its major interference sources based on the information feedback. In this paper, we provide an interactive learning framework for distributed power control of delay sensitive users over multi-carrier wireless networks. Specifically, the user can adopt corresponding interactive learning schemes to explicitly model the other users' strategies if public information feedback is available, or to implicitly model the impact of other users' actions on its utility if only private information is available. Based on these models, the user creates beliefs and is able to strategically adapt its decisions to maximize its utility. We determine the performance upper bounds for the user's utility when learning from private or public information feedback and investigate the cost-performance tradeoffs resulting from the information feedback gathered with different frequencies and from various users. The simulation results show that the proposed adaptive interactive learning approach significantly improves the energy efficiency of delay sensitive users compared to schemes that perform myopic best response.**

Index Terms: interactive learning; information feedback; delay sensitive applications; power control.

I. INTRODUCTION

Dynamic resource management is an important problem in wireless networks. Prior literature has investigated dynamic resource management for path selection (routing) [1], time sharing [2], frequency channel selection [3][4], power allocation [5][8][13], etc. In this paper, we focus on the non-cooperative decentralized setting, where autonomous users make decisions on accessing resources based on their current knowledge about their opponents as determined from information feedback. Such information feedback is essential for decentralized dynamic resource management, since in informationally-decentralized wireless networks, it is impossible for a user to know the exact actions of the other users sharing the network. Hence, it is important to investigate how users can dynamically adapt their current decisions to maximize the expected utility based on available information feedback. We focus on the joint *power-spectrum* allocation for dynamic resource management in wireless networks, since the interference at the physical layer results in a strong coupling between the transmission actions (i.e. the power/frequency channel selections) of the competing users. However, the proposed solution can also be used in other decentralized dynamic resource management problems.

Joint power and spectrum resource allocation research has attracted a lot of attention in recent years [7]-[12]. For the multi-user case to maximize the overall throughput, the resource allocation problem becomes very complicated since the wireless mutual interference among users results in a non-convex optimization problem [7]. The computational complexity of the centralized approaches becomes prohibitive as the number of users increases. Moreover, the centralized approaches require the propagation of global control information back and forth to a common coordinator, thereby incurring heavy signaling overhead [5]. Hence, decentralized solutions, such as the "iterative water filling" [8], are more desirable in practice.

Recently, game-theoretic concepts have been applied to deal with the decentralized resource allocation problem [9]-[13] using various utility functions. For example, in [9], non-cooperative power control games were constructed where each user possesses an energy-efficient utility function. The existence and uniqueness of Nash equilibrium in such non-cooperative game was extensively studied. In [9][10], other than maximizing the throughput, users maximize a ratio of throughput over the transmitted power (measured in bits/joule). In [11][12], a pricing mechanism was employed to provide Pareto-efficient solutions [20] by adopting an additional penalty term associated with the power consumption in the utility function. In [13], a reinforcement learning approach for the non-cooperative game is proposed and the convergence property of the reinforcement approach was studied.

In short, previous research mainly concentrates on studying the existence and performance of Nash equilibrium in non-cooperative games or developing efficient algorithms to approach the Pareto boundary. However, prior research does not consider the users' availability of information feedback from various users and ignores the performance degradation when the actions of the other users are not accurately modeled. Note that

[*] The work was supported by ONR and NSF CAREER 0541867.

without a central coordinator, multiple users sharing the same wireless network need to manage their local resources based on the available information feedback. Hence, the best response strategy of a selfish user making decisions in the non-cooperative game based on “limited” (incomplete) information feedback [5] still needs to be determined. Intuitively, a “foresighted” user with more information should be able to gain more benefits in such a non-cooperative game. However, such information feedback is not costless. In practical systems, heavy signaling overhead can degrade the users’ performance [17]. Therefore, it is important to investigate what is the benefit that a user can derive from gathering more information feedback, which allows it to better model the competing wireless users, while *explicitly* considering the cost of feeding back the information.

In this paper, we investigate two types of information feedback for autonomous self-interested users (transmitter-receiver pairs) participating in the power control game. The transmitters will select the transmitting power levels and the frequency channels by maximizing the utility function based on two types of information feedback:

1) *Private information feedback* – To evaluate the utility function, transmitters actually require their receivers to provide important channel state information, the Signal-to-Interference-Noise Ratio (SINR). The SINR value contains the aggregate effect of other users’ actions and this value can only be measured at the receiver side. Such information needs to be fed back to the transmitter to make decisions. This information feedback between the transmitter-receiver pair is referred to as the private information feedback.
2) *Public information feedback* – When non-cooperative users have incentives to exchange information (depending on the communication protocols, such as in [18]), explicit information feedback about the other users’ actions enables a user to directly model the other users efficiently and hence, improve the accuracy of the utility evaluation resulting from taking different actions. Even when users are non-cooperative, they can still reveal their action information to others in order to maximize their own utilities [21]. This explicit information feedback among users is referred to as the public information feedback.

Note that the private information feedback contains implicit information about the actions of the other users in the network. On the other hand, by gathering public information feedback, users can explicitly model their opponents. Due to the informationally-decentralized nature of the wireless network, when a user makes decisions, the user does not know the exact transmission actions that its interfering neighbors will take. If a user is foresighted, meaning that it can predict the exact actions of its competing users by exploiting the experienced information feedback, its performance can be improved [3][21] . This requires the user to learn the transmission strategies of its major interferers through interactive learning [19] based on the available information feedback. Figure 1 illustrates the differences of the conventional distributed power control and the proposed power control using interactive learning. We discuss two classes of interactive learning schemes – payoff-based learning and model-based learning, which require different types of information feedback. In this paper, we assume that the information feedback is truthful and error-free[1], and investigate how to adapt the information feedback to enable a user to maximize its utility in different network scenarios through interactive learning.

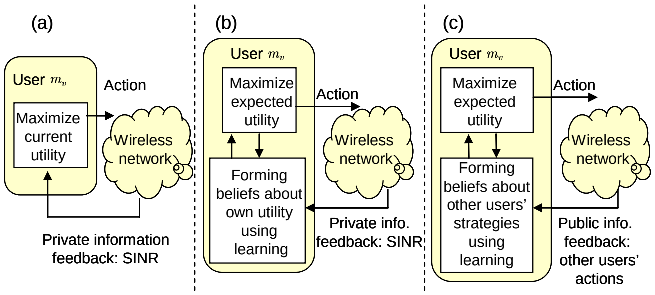
(a)
User m_v
Maximize current utility
Action
Wireless network
Private information feedback: SINR
(b)
User m_v
Maximize expected utility
Forming beliefs about own utility using learning
Action
Wireless network
Private info. feedback: SINR
(c)
User m_v
Maximize expected utility
Forming beliefs about other users’ strategies using learning
Action
Wireless network
Public info. feedback: other users’ actions

Fig. 1 (a) Conventional distributed power control. (b) Payoff-based interactive learning with private information feedback. (c) Model-based interactive learning with public information feedback.

We focus on the problem of delay-sensitive applications sharing the same wireless network. Due to the delay sensitivity, the utility of a user is dramatically impacted by the applications of other users. This provides the user an additional incentive to adopt a better learning scheme, since it cannot wait a long time to transmit the packets. To cope with the delay sensitivity, we need to consider not only the impact of the effective throughput over the wireless network, but also the source traffic characteristics, including the source rates and the delay deadlines of the applications.

In summary, this paper aims to make the following contributions:

1) **Feedback-driven interactive learning framework.** We develop a feedback-driven learning framework for distributed power control of delay sensitive users. Depending on the available information feedback, a user can form beliefs using interactive learning about what should be its expected future utility for the various actions or about the transmission strategies of its major interferers based on which it can compute the impact of its actions on its expected future utility.

2) **Cost-performance tradeoff of interactive learning.** We characterize the cost of information feedback by explicitly considering – a) from whom (i.e. from which transmitters or receivers) this information is obtained, and b) how often such information is obtained (i.e. the frequency of getting feedback). We quantify the cost-performance tradeoff when learning from different information feedback and show how to adapt the information feedback to maximize the learning efficiency.

3) **Analytical upper bounds based on interactive learning.** We also quantify the utility upper bounds that

[1] In this paper we will assume that the public information is accurately transmitted. However, if it is believed that malicious users are presented in the system, mechanism design can be used to compel users to declare their information truthfully (see [6]).

can be achieved by a user through learning based on private or public information feedback.

4) **Outperforming the Nash equilibrium performance in the power control game.** We consider learning solutions based on both the private and public information feedback, which maximize the expected user's utility rather than optimizing myopically the immediate (current) utility. These learning solutions outperform the Nash equilibrium performance, which is achieved when users deploy myopic best response such as iterative water-filling [8].

The paper is organized as follows. In Section II, we discuss the considered network settings and formulate the studied informationally-decentralized dynamic resource management problem among wireless users competing for resources with incomplete information. In Section III, we characterize the information feedback and discuss the cost-performance tradeoff of the information feedback. Based on the type of information feedback, we introduce two classes of interactive learning solutions and discuss how to adjust the information feedback to improve the learning efficiency. In Section IV, payoff-based learning is discussed, which employs only private information feedback. In Section V, we introduce model-based learning, which requires public information feedback. Section VI presents simulation results and Section VII concludes the paper.

II. Network Settings and Problem Formulation

A. Network settings

We assume that there are V users ($m_1, \ldots, m_V$) that are simultaneously transmitting delay sensitive applications over the same wireless infrastructure. A network user m_v is composed of a source node n_v^s (transmitter) and a destination node n_v^d (receiver) that can establish a direct communication connection, i.e. $m_v = \{n_v^s, n_v^d\}$. We assume that there are multiple frequency channels for users to transmit their applications and $\mathcal{F}$ is the set of all channels. An illustrative network example is depicted in Figure 2.

B. Actions and strategies

We consider a fully distributed setting where each user attempts to maximize its own utility function by selecting the optimal frequency channels and transmitted power levels in the selected channels. We assume that only frequency channels in the set $\mathcal{F}_v \subseteq \mathcal{F}$ are available to the user m_v. Network user m_v transmits its application through one of the available frequency channels $f_v \in \mathcal{F}_v$ with a power level $0 \leq P_v \leq P_v^{\max}$. In this paper, we assume that the transmit power level can take a discrete set of values in the set $\mathcal{P}_v$. Hence, we define the action of a user m_v as $A_v = [f_v, P_v] \in \mathcal{A}_v = \mathcal{F}_v \times \mathcal{P}_v$. We assume that $S_v(A_v)$ represents the probability that a user m_v takes A_v as its action. The *strategy*[2] of user m_v is defined as a probability distribution $\mathbf{S}_v = [S_v(A_v),\ \text{for } A_v \in \mathcal{A}_v] \in \mathcal{S}_v$, where $\mathcal{S}_v$ is a set of probability distributions over all feasible actions $A_v \in \mathcal{A}_v$.

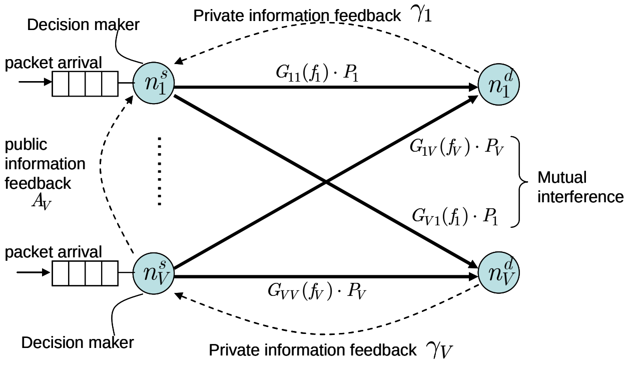

Fig. 2 System diagram of the dynamic joint power-spectrum resource allocation.

Let $G_{vv'}(f_v)$ represent the channel gain from the transmitter $n_{v'}^s$ of the user $m_{v'}$ to the receiver n_v^d of the user m_v, which is related to the distance of the two nodes and channel characteristics. Let $\mathbf{G}_v = \{G_{vv'}(f_v), \forall n_{v'}^s, \forall f_v\}$ denote a set of channel gains from all the transmitters $n_{v'}^s$ to the receiver n_v^d of the user m_v. The SINR γ_v experienced by user m_v in frequency channel f_v depends on the user's action A_v and the actions of all the other users, denoted as A_{-v}:

$$\gamma_v(A_v, A_{-v}) = \frac{G_{vv}(f_v)P_v}{N_{f_v} + \sum_{v' \neq v, f_{v'} = f_v} G_{vv'}(f_v)P_{v'}}, \qquad (1)$$

where N_{f_v} represents the AWGN noise level in the frequency channel f_v. The term $\sum_{v' \neq v, f_{v'} = f_v} G_{vv'}(f_v)P_{v'}$ represents the mutual interference coupling from the other users. The effective throughput available at a transmitter n_v^s depends on the experienced SINR γ_v and it is denoted as $B_v(A_v, A_{-v}) = T_v(f_v)(1 - p_v(\gamma_v))$, where $T_v(f_v)$ and $p_v(\gamma_v)$ represent the maximum transmission rate and packet error rate of user m_v using the frequency channel f_v.

C. Delay sensitive applications

We assume that users are transmitting delay sensitive applications. The packet arrival process of a user m_v is assumed to be Poisson with the mean arrival rate λ_v. The delay deadline of the packets of user m_v is d_v. We assume that each user maintains a buffer at its transmitter and that the arriving packets which cannot be transmitted

[2] The strategy defined in this paper can be regarded as a mixed strategy and the action defined in this paper can be regarded as a pure strategy in game theory [20].

immediately will be queued in the buffer. The effective throughput $B_v(A_v, A_{-v})$ is independent of the packet arrival process. Hence, there will be queuing delay and transmission delay. We denote the total delay as D_v, which is a random variable depending on both arrival rate λ_v and the effective throughput $B_v(A_v, A_{-v})$. The packet loss rate is defined as the probability when this delay exceeds the packet delay deadline, i.e. $\mathrm{Prob}\{D_v(\lambda_v, B_v(A_v, A_{-v})) > d_v\}$. Therefore, the rate of successfully received packets is $\lambda_v \mathrm{Prob}\{D_v(\lambda_v, B_v(A_v, A_{-v})) \le d_v\}$.

D. Utility function definition

We assume the users attempt to maximize their energy-efficient utility functions (measured in bits/joule) similar to [9]. The difference is that we also consider the packet loss due to the expiration of the delay deadline for delay sensitive applications. The utility function of a user m_v is

$$u_v(A_v, A_{-v}) = \frac{\lambda_v \mathrm{Prob}\{D_v(\lambda_v, B_v(A_v, A_{-v})) \le d_v\}}{P_v}. \quad (2)$$

The utility function reflects the expected number of packets that is successfully received (rather than transmitted as in [9]) per joule of energy consumed for delay sensitive users. More details about how this utility function can be computed in a practical communication setting can be found in Appendix I. Figure 3 illustrates the utility function of a user m_v using different power $0 \le P_v \le P_v^{\max}$ in a selected frequency channel f_v with fixed interference. We denote the power of user m_v that maximizes the utility function when transmitting in channel f_v as $P_v^{tar}(f_v)$.

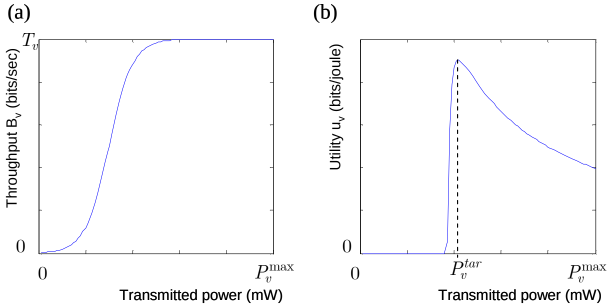

Fig. 3 (a) Throughput B_v vs. P_v in a selected frequency channel f_v with fixed interference.

(b) Utility u_v vs. P_v in a selected frequency channel f_v with fixed interference.

E. Problem formulation

Let $\mathbf{A}_{-v}^{myop}$ represent the latest actions of the other users observed by a user m_v in the network. Conventionally, the user m_v adopts a myopic distributed optimization, which can be formulated as:

$$A_v^{myop} = [f_v^{myop}, P_v^{myop}] = \arg\max_{A_v \in \mathcal{A}_v} u_v(A_v, \mathbf{A}_{-v}^{myop}). \quad (3)$$

In [9], it was shown that the myopic best response A_v^{myop} converges to the Nash equilibrium under certain conditions on channel gains. However, if a foresighted user m_v knows the exact response actions of other users $\mathbf{A}_{-v}^{fors}(A_v)$, a better performance can be achieved [20]. Let $\mathbf{A}_{-v}^{fors}(A_v)$ represent the actions of the other users given that the action A_v is taken by user m_v. The optimization performed by a foresighted user can be formulated as [20]:

$$A_v^{fors} = [f_v^{fors}, P_v^{fors}] = \arg\max_{A_v \in \mathcal{A}_v} u_v(A_v, \mathbf{A}_{-v}^{fors}(A_v)). \quad (4)$$

Let us assume that only one user is foresighted, and all the other users in the network still adopt a myopic best response. Given the exact response actions $\mathbf{A}_{-v}^{fors}(A_v)$, the foresighted decision making based on the complete information of the other users will converge to the Stackelberg equilibrium [20] and the optimal utility is denoted as

$$U_v(\mathbf{A}_{-v}^{fors}(A_v)) = \max_{A_v \in \mathcal{A}_v} u_v(A_v, \mathbf{A}_{-v}^{fors}(A_v)). \quad (5)$$

However, due to the informationally-decentralized nature of the wireless networks, it is impossible for each user to know in practice the exact response actions $\mathbf{A}_{-v}^{fors}(A_v)$. Hence, accurately modeling the actions $\mathbf{A}_{-v}^{fors}(A_v)$ based on the information feedback is necessary.

Definition 1: Denote the information feedback of user m_v at time slot t as $\mathcal{I}_v^t$, regardless whether the information feedback is private or public. We define the *observed information history* of user m_v at time slot t as $o_v^t = \{\mathcal{I}_v^t, o_v^{t-1}\}$.

Assume that the strategy of user m_v at time slot t is denoted as $\mathbf{S}_v^t$. We use the notation $\mathbf{M}_{-v}$ to indicate the set of all users except user m_v. The strategy of all users in the network except user m_v is $\mathbf{S}_{-v}^t = \{\mathbf{S}_u^t, \text{for } m_u \in \mathbf{M}_{-v}\}$.

Definition 2: Since the exact response actions of other users $\mathbf{A}_{-v}^{fors}$ are not available to user m_v in real time, user m_v estimates $\mathbf{A}_{-v}^{fors}$ by building a *belief* on the other users' strategies $\mathbf{S}_{-v}^t$. The belief of user m_v is defined as $\tilde{\mathbf{S}}_{-v}^t(A_v) = \{\tilde{S}_{-v}^t(A_{-v} \mid A_v), \text{ for all } A_v \in \mathcal{A}_v\}$, where $\tilde{S}_{-v}^t(A_{-v} \mid A_v)$ [3] are the estimated strategies of the other users given that user m_v decides to take action A_v.

In other words, user m_v estimates the other users'

[3] $\tilde{\mathbf{S}}_{-v}^t(A_v) = \{\tilde{\mathbf{S}}_u^t(A_u \in \mathcal{A}_u \mid A_v), \text{for } m_u \in \mathbf{M}_{-v}\}$ and $\tilde{\mathbf{S}}_u^t(A_u \in \mathcal{A}_u \mid A_v) = [\tilde{S}_u^t(A_u \mid A_v), \text{ for } A_u \in \mathcal{A}_u]$ represents the conditional probability distribution when user m_v takes the action A_v.

strategies $\tilde{\mathbf{S}}_{-v}^t(A_v)$ for each of its action $A_v \in \mathcal{A}_v$ [4].

Definition 3: Assume Λ_v represents the interactive learning scheme adopted by user m_v. A *learning scheme* Λ_v is defined as a method that allows user m_v to build a belief $\tilde{\mathbf{S}}_{-v}^t = \Lambda_v(o_v^t)$ [5] based on the observed information history o_v^t, in order to estimate the actions of the other users $\mathbf{A}_{-v}^{fors}$.

Specifically, by learning from the observed information history o_v^t, user m_v builds its *belief* $\tilde{\mathbf{S}}_{-v}^t$ on the other users' strategies and determine its own best response strategy $\mathbf{S}_v^t$. Figure 4 illustrates how a delay sensitive user makes decisions based on the observed information history o_v^t and the mutual interference coupling in the dynamic wireless environment. The problem in equation (4) can be now reformulated as:

$$\mathbf{S}_v^t(\tilde{\mathbf{S}}_{-v}^t) = \arg\max_{\mathbf{S}_v \in \mathcal{S}_v} E_{(\mathbf{S}_v, \tilde{\mathbf{S}}_{-v}^t)}[u_v(\mathbf{S}_v, \tilde{\mathbf{S}}_{-v}^t)]\,. \qquad (6)$$

Based on the determined $\mathbf{S}_v^t$, user m_v selects an action A_v at time slot t.

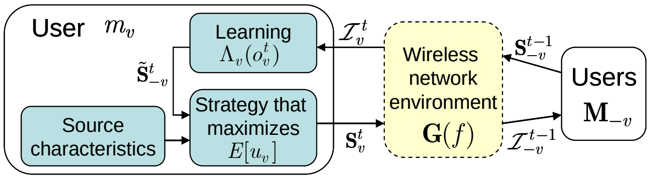

Fig. 4 Interactions among users and the foresighted decision making based on information feedback.

F. Learning efficiency

The performance of an interactive learning approach depends on how accurate the belief $\tilde{\mathbf{S}}_{-v}^t = \Lambda_v(o_v^t)$ can predict the actions $\mathbf{A}_{-v}^{fors}$. A more accurate prediction of $\mathbf{A}_{-v}^{fors}$ can lead to a better learning efficiency. We define the learning efficiency $J_v(\Lambda_v(o_v^t))$ of the learning approach Λ_v (based on the observed information history o_v^t) by quantifying its impact on the expected utility, i.e.

$$J_v(\Lambda_v(o_v^t)) \triangleq E_{(\mathbf{S}_v^t, \tilde{\mathbf{S}}_{-v}^t)}[u_v(\mathbf{S}_v^t, \Lambda_v(o_v^t))], \qquad (7)$$

where

$$E_{(\mathbf{S}_v^t, \tilde{\mathbf{S}}_{-v}^t)}[u_v(\mathbf{S}_v^t, \Lambda_v(o_v^t))] = \sum_{A_v \in \mathcal{A}} \left(S_v^t(A_v) \left(\sum_{A_{-v} \in \mathcal{A}^{V\text{-}1}} \tilde{S}_{-v}^t(A_{-v} \mid A_v) u_v(A_v, A_{-v}) \right) \right). \qquad (8)$$

The notation $\tilde{S}_{-v}^t(A_{-v} \mid A_v)$ is used to represent the joint probability that the users $m_u \in \mathbf{M}_{-v}$ take actions A_{-v}, given that user m_v took the action A_v.

Since the belief $\tilde{\mathbf{S}}_{-v}^t$ is only a prediction for $\mathbf{A}_{-v}^{fors}$, we define the Price of Imperfect Belief (PIB) for using the learning scheme Λ_v based on the observed information history o_v^t as the performance difference between the Stackelberg equilibrium [21] $U_v(\mathbf{A}_{-v}^{fors})$ (where the user m_v knows the exact response of the other users) and the practical learning efficiency $J_v(\Lambda_v(o_v^t))$, i.e.

$$\Delta_P(\Lambda_v(o_v^t)) \triangleq U_v(\mathbf{A}_{-v}^{fors}) - J_v(\Lambda_v(o_v^t))\,. \qquad (9)$$

In the next sections, we quantify the cost of the information feedback $\mathcal{I}_v^t$ and study two classes of interactive learning approaches Λ_v^{priv} and Λ_v^{pub} based on different types of information feedback.

III. INFORMATION FEEDBACK FOR INTERACTIVE LEARNING

A. Characterization of information feedback

In this paper, we define the *entire* information history from all users until time slot t as

$$h^t = \{\gamma_v^s, \mathbf{G}_v^s, A_v^s, \text{for } v = 1,...,V, s = 0,...,t\}\,. \qquad (10)$$

Note that a user m_v observes only a subset of the entire history through information feedback, i.e. $o_v^t \subseteq h^t$. The observed information history o_v^t can be characterized in three distinct categories:

- **Types of information feedback** – As mentioned before, there are two types of information that a user m_v can observe at a certain time slot t, i.e. the private information feedback $\mathcal{I}_v^{t,priv} = \{\gamma_v^{t-1}\}$ or the public information feedback $\mathcal{I}_{-v}^{t,pub} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, \text{ for } m_u \in \mathbf{M}_{-v}\}$. Recall that $o_v^t = \{\mathcal{I}_v^t, o_v^{t-1}\}$ in Definition 1.
- **Information zone** – We define the information zone $\mathbf{V}_v^t$ as a set of users that are able to feed back information to the transmitter of user m_v at time slot t. In the wireless communication networks, the information from further users is less significant, since the effect of mutual interference coupling decreases ($G_{vv'}$ decreases in equation (1)) as the distance increases [14]. Hence, user m_v can selectively collect the information only from a set of neighboring (e.g. within an information horizon as in [17]) users $m_u \in \mathbf{V}_v^t$, i.e.

[4] Based on different types of information feedback, user m_v may implicitly model the other users by only estimating the aggregate effect of the other users. See Section IV for more detail.

[5] For representation convenience, we use the simplified notation $\mathbf{A}_{-v}^{fors}$ to represent $\mathbf{A}_{-v}^{fors}(A_v), A_v \in \mathcal{A}_v$ as the exact response actions of other users. And also use $\tilde{\mathbf{S}}_{-v}^t$ to represent $\tilde{\mathbf{S}}_{-v}^t(A_v)$ in the rest of the paper.

$\mathcal{I}_{-v}^{t,pub} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, \text{ for } m_u \in \mathbf{V}_v^t\}$. Since the information zone of the private information feedback only contains user m_v itself, we define $|\mathbf{V}_v^t| = 0$ for $\mathcal{I}_v^{t,priv} = \{\gamma_v^{t-1}\}$.

- **Information feedback frequency** – In our problem formulation in equation (6), user m_v can obtain the information feedback and make decisions during every time slot. However, in practice, user m_v can obtain the information feedback at different time scales. Assume that user m_v observes the information feedback for every τ_v time slots ($\tau_v \in \mathbb{Z}^+$). Define $\omega_v = 1/\tau_v$ as the frequency of the information feedback, $0 \le \omega_v \le 1$. Let $\omega_v = 0$ represent the case when no information feedback is obtained. Let $\mathbf{T}_v^t$ represent the set of time slots before time slot t at which the user m_v obtains information and makes decisions, i.e. $\mathbf{T}_v^t = \{(s_v^0 + k\tau_v), k = 0,1,...,K_v^t\}$, where s_v^0 is the initial time slot that a user m_v obtains information and starts making decisions. The number of decisions made by user m_v up to time t equals $K_v^t = \lfloor (t - s_v^0)/\tau_v \rfloor$, where $\lfloor \bullet \rfloor$ is the floor operation. The observed information history now becomes $o_v^t = \{\mathcal{I}_v^s, \text{for } s \in \mathbf{T}_v^t\}$.

B. *Cost-performance tradeoff when adjusting the information feedback*

Let us denote the information feedback overhead of user m_v as $\sigma_v(\omega_v, |\mathbf{V}_v|)$ [6], which is a function of the information feedback frequency ω_v and the number of the neighboring users $|\mathbf{V}_v|$. In general, with more frequent information feedback (i.e. a larger ω_v) or feedback from more users (i.e. a larger $|\mathbf{V}_v|$), a user can obtain more information from the entire information history h^t and hence, this results in a more accurate belief. On the other hand, a large information overheads $\sigma_v(\omega_v, |\mathbf{V}_v|)$ can degrade the learning efficiency $J_v(\Lambda_v(o_v^t))$.

In this paper, we assume that the packet transmission and the information feedback are multiplexed in the same frequency channel. Hence, considering the information overhead, the effective throughput can be represented as $B_v'(\sigma_v, A_v, A_{-v}) = B_v(A_v, A_{-v}) \times \theta(\sigma_v)$, where $0 < \theta(\sigma_v) \le 1$ represents the fraction of time dedicated to the packet transmission, and it is a decreasing function of σ_v . Given $\mathbf{A}_{-v}^{fors}$, the utility function in equation (2) can be derived as (see Appendix I for more detail):

$$u_v(\sigma_v, A_v, A_{-v}) = \begin{cases} \dfrac{\lambda_v}{P_v}\left(1 - \dfrac{1}{F_v(\sigma_v, \gamma_v(A_v, A_{-v}))}\right), \\ \text{if } B_v'(\sigma_v, A_v, A_{-v})/L_v > \lambda_v \\ 0 \quad ,\text{otherwise} \end{cases} \tag{11}$$

$$F_v(\sigma_v, \gamma_v(A_v, A_{-v})) \equiv \exp(B_v'(\sigma_v, \gamma_v(A_v, A_{-v}))\frac{d_v}{L_v} - \lambda_v d_v)$$

$$. \tag{12}$$

L_v represents the average packet length of user m_v . Note that both $B_v'(\sigma_v, A_v, A_{-v})$ and $F_v(\sigma_v, \gamma_v(A_v, A_{-v}))$ are decreasing functions of σ_v . Hence, the utility function is a non-increasing function of σ_v .

Intuitively, if σ_v is large, the belief $\tilde{\mathbf{S}}_{-v}^t$ provides an accurate model on $\mathbf{A}_{-v}^{fors}$. On the other hand, if σ_v is small, the belief $\tilde{\mathbf{S}}_{-v}^t$ provides an inaccurate model on $\mathbf{A}_{-v}^{fors}$. By having more information $o_v^t \subseteq h^t$, increasing σ_v can improve the learning efficiency.

Proposition 1. Optimal information feedback overhead: For a given learning scheme Λ_v , there exists at least one optimal information feedback overhead σ_v^* such that

$$\sigma_v^*(\Lambda_v) = \arg\min_{\sigma} \Delta_P(\Lambda_v(o_v^t(\sigma))) \,. \tag{13}$$

Proof: Note that minimizing Δ_P is the same as maximizing $J_v(\Lambda_v(o_v^t))$. Since $0 \le J_v(\Lambda_v(o_v^t)) \le U(\mathbf{A}_{-v}^{fors})$ is bounded, there must exist a minimum value with a certain σ_v^* .

Based on Proposition 1, we propose an adaptive interactive learning that adapts the information feedback parameters for user m_v to improve its learning efficiency J_v . Figure 5 presents the system block diagram of our adaptive interactive learning framework. Due to the consideration of the source characteristics, the interactive learning framework is operated at the application layer. The goal of user m_v in the adaptive interactive learning framework is to build the belief $\tilde{\mathbf{S}}_{-v}^t$ based on o_v^t for determining the best response strategy $\mathbf{S}_v^t$ and adjust the information feedback $\mathcal{I}_v^{t+1}(\sigma_v)$ to improve the learning efficiency $J_v(\Lambda_v(o_v^t))$. In the following sections, we will discuss the adaptive interactive learning schemes based on different types of information feedback in more details.

IV. Interactive Learning with Private Information Feedback

In the case where user m_v only observes the private information feedback $\mathcal{I}_v^{t,priv}$, it can only model the aggregate effect of other users' actions through the experienced SINR value γ_v . Hence, it cannot model the exact response actions of the other users $\mathbf{A}_{-v}^{fors}$ explicitly. Note that the observed information history in this case is $o_v^t(\omega_v) = \{\gamma_v^{s-1}, s \in \mathbf{T}_v^t(\omega_v)\}$. Based on this

[6] Note that for private information feedback, the information overhead σ_v only depends on ω_v ($|\mathbf{V}_v|$=0).

observed information history $o_v^t(\omega_v)$, user m_v is aware of its past actions $A_v^{s-1}, s \in \mathbf{T}_v^t$ and the past resulting utilities $u_v(A_v^{s-1}, \gamma_v^{s-1}), s \in \mathbf{T}_v^t$. Let $\tilde{u}_v(A_v, o_v^t(\omega_v))$ represent the estimated utility of user m_v if the action A_v is taken. Instead of predicting the exact response actions $\mathbf{A}_{-v}^{fors}$ explicitly, user m_v builds a belief on the utility and determines its best strategy $\mathbf{S}_v^t$ based on its past experienced action-utility pairs $[A_v^{s-1}, u_v(A_v^{s-1}, \gamma_v^{s-1})], s \in \mathbf{T}_v^t$. Hence, user m_v does not try to estimate the probability $\tilde{S}_{-v}^t(A_{-v} \mid A_v)$ in equation (8). Instead, user m_v builds directly its belief on what will be the average utility impact that it will experience if it takes action A_v, i.e. $\tilde{u}_v(A_v, o_v^t(\omega_v))$ substitutes the term $\sum_{A_{-v} \in \mathcal{A}^{V-1}} \tilde{S}_{-v}^t(A_{-v} \mid A_v) u_v(A_v, A_{-v})$ in equation (8).

Let $\mathbf{S}_v^t(\omega_v) = \Lambda_v^{priv}(o_v^t(\omega_v))$ be the strategy of user m_v at time slot t learned from the observed information history $o_v^t(\omega_v)$. From equation (7), the learning efficiency of user m_v is

$$J_v(\Lambda_v^{priv}(o_v^t(\omega_v))) = \sum_{A_v \in \mathcal{A}} S_v^t(A_v) \tilde{u}_v(A_v, o_v^t(\omega_v)). \quad (14)$$

To minimize Δ_P in equation (9), the best response strategy is:

$$\mathbf{S}_v^t(\omega_v) = \arg\max_{\mathbf{S}_v \in \mathcal{S}_v} \sum_{A_v \in \mathcal{A}} S_v(A_v) \tilde{u}_v(A_v, o_v^t(\omega_v)). \quad (15)$$

The payoff-based learning based on private information feedback can be represented equation (15). After the strategy $\mathbf{S}_v^t$ is determined, the action of user m_v at time slot t is determined by

$$A_v^t = Rand(\mathbf{S}_v^t), \quad (16)$$

where $Rand(\mathbf{S}_v^t)$ represents a random selection based on the probabilistic strategy $\mathbf{S}_v^t \in \mathcal{S}_v$. Payoff-based learning [19] provides a method to learn the strategy $\mathbf{S}_v^t$ from the past experienced action-utility pairs $[A_v^{s-1}, u_v(A_v^{s-1}, \gamma_v^{s-1})], s \in \mathbf{T}_v^t$. A simple example of a payoff-based learning method will be provided in Section IV.*A*.

If the private information feedback is costless (i.e. $B_v' = B_v$ in equation (11)), the utility upper bound of the payoff-based learning can be calculated based on the resulting strategy $\mathbf{S}_v^* = [S_v^*(A_v), \text{ for all } A_v \in \mathcal{A}_v]$ at convergence.

Proposition 2. Performance upper bound with private information feedback: For a payoff-based learning with private information feedback, if the information feedback is costless, the upper bound of the learning efficiency $J_v(\Lambda_v^{priv})$ is $(1 - \varepsilon_v(\Lambda_v^{priv}))U_v(\mathbf{A}_{-v}^{fors})$, with $0 \le \varepsilon_v(\Lambda_v^{priv}) < 1$, and

$$\varepsilon_v(\Lambda_v^{priv}) = \frac{1}{U_v(\mathbf{A}_{-v}^{fors})} \sum_{A_v \in \mathcal{A}} g(A_v) u_v(A_v, \mathbf{A}_{-v}^{fors}), \text{ where}$$

$$g(A) = \begin{cases} 1 - S_v^*(A), \text{ for } A = A_v^{fors} \\ -S_v^*(A), \text{ otherwise} \end{cases}. \quad (17)$$

Proof: By substituting equation (14) into equation (9), the PIB becomes $\Delta_P(\Lambda_v^{priv}) = \varepsilon_v(\Lambda_v^{priv}) U_v(\mathbf{A}_{-v}^{fors})$. Since $u_v(A_v, \mathbf{A}_{-v}^{fors})$ has costless information feedback, substituting $\tilde{u}_v(A_v, o_v^t(\omega_v))$ by $u_v(A_v, \mathbf{A}_{-v}^{fors})$ provides a lower bound on $\Delta_P(\Lambda_v^{priv})$, which is $\sum_{A_v \in \mathcal{A}} g(A_v) u_v(A_v, \mathbf{A}_{-v}^{fors})$.

In order to increase the learning efficiency $J_v(\Lambda_v^{priv})$, user m_v needs to increase the accuracy of the best response strategy $\mathbf{S}_v^*$ such that it approaches A_v^{fors}. Next, let us give a simple example using a well-known reinforcement learning solution [19].

A. Reinforcement learning based on private information feedback

In this subsection, let us assume $\omega_v = 1$. By applying typical reinforcement learning, user m_v models its best response strategy $\mathbf{S}_v^t$ as

$$S_v^t(A_v) = \frac{r_v^t(A_v)}{\sum_{A_v \in \mathcal{A}_v} r_v^t(A_v)}, \quad (18)$$

where $r_v^t(A_v)$ represents the *propensity* [19] of user m_v choosing an action A_v at time slot t. Let us define $\mathbf{r}_v^t = [r_v^t(A_v), \text{ for } A_v \in \mathcal{A}_v]$ as a vector of propensity of all feasible actions. The user updates $\mathbf{r}_v^t$ based on the experienced utility, $u_v(A_v^{t-1}, \gamma_v^{t-1})$ when the action A_v^{t-1} is taken at time slot $t-1$. Here, we adopt the cumulative payoff matching [19]:

$$\mathbf{r}_v^t = \alpha \mathbf{r}_v^{t-1} + (1-\alpha) u_v(A_v^{t-1}, \gamma_v^{t-1}) \mathbf{I}(A_v^{t-1}), \quad (19)$$

where α is the discount factor for the history value of the cumulative propensity. $\mathbf{I}(A_v^t) = [I(A = A_v^t), \text{ for } A \in \mathcal{A}_v]$ represents an indicator vector such that

$$I(A = A_v^t) = \begin{cases} 1, \text{ if } A = A_v^t \\ 0, \text{ if } A \ne A_v^t \end{cases}. \quad (20)$$

B. Adaptive reinforcement learning

The reinforcement learning in the previous subsection fixes $\omega_v = 1$, i.e. user m_v obtains information feedback at each time slot. From Proposition 1, we know that by adjusting information feedback frequency ω_v to ω_v^*, user m_v can minimize its PIB Δ_P. Hence, we introduce the adaptive reinforcement learning[7] that adjusts ω_v to maximize the learning efficiency $J_v(\Lambda_v^{priv})$. Specifically, for $\omega_v < 1$, user m_v will not receive the private information feedback at each time slot with probability $1-\omega_v$. If there is no information feedback, user m_v takes the *baseline action* A_v^{base}, which is the past action that ever provides the best payoff value. Smaller ω_v means that the user is more reluctant to deviate from its baseline action and leads to a lower information feedback overhead. With probability ω_v, the user will receive the information feedback and perform the same reinforcement learning as in the previous subsection. After user m_v selects an action A_v^t, it compares the payoff value u_v and then updates the record of the baseline action A_v^{base} and the baseline payoff value u_v^{base}:

$$A_v^{base} = \begin{cases} A_v^{t-1}, \text{ if } u_v(A_v^{t-1}, \gamma_v^{t-1}) > u_v^{base} \\ A_v^{base}, \text{ otherwise} \end{cases}. \quad (21)$$

$$u_v^{base} = \max(u_v^{base}, u_v(A_v^{t-1}, \gamma_v^{t-1})). \quad (22)$$

Finally, user m_v evaluates the learning efficiency $J(\Lambda_v(o_v^t))$ and changes the information feedback frequency ω_v by $\triangle\omega_v$ until the maximum $J(\Lambda_v(o_v^t))$ is found. The details of the proposed adaptive reinforcement learning can be found in Algorithm 1.

V. INTERACTIVE LEARNING WITH PUBLIC INFORMATION FEEDBACK

Unlike the payoff-based learning, when user m_v observes public information feedback $\mathcal{I}_{-v}^{t,pub} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, \text{ for } m_u \in \mathbf{M}_{-v}\}$, the observed information history is $o_v^t = \{\mathcal{I}_{-v}^{s,pub}, s \in \mathbf{T}_v^t\}$. Based on this, user m_v can directly model the strategy of other users and build belief $\tilde{\mathbf{S}}_{-v}^t$ on it explicitly.

Let $\tilde{\mathbf{S}}_{-v}^t(\sigma_v) = \Lambda_v^{pub}(o_v^t(\sigma_v))$. From equation (7), the learning efficiency is

$$J_v(\Lambda_v^{pub}(o_v^t(\sigma_v))) = \sum_{A_v \in \mathcal{A}} \left(S_v^t(A_v) \left(\sum_{A_{-v} \in \mathcal{A}^{V\text{-}1}} \tilde{S}_{-v}^t(A_{-v} \mid A_v) u_v(\sigma_v, A_v, A_{-v}) \right) \right). \quad (23)$$

To minimize the Δ_P in equation (9), the best response strategy of user m_v is to take the action ($\mathbf{S}_v^t = \mathbf{I}(A_v^t)$):

$$A_v^t(\sigma_v) = \arg\max_{A_v \in \mathcal{A}_v} E_{\tilde{\mathbf{S}}_{-v}^t}[u_v(A_v, \tilde{\mathbf{S}}_{-v}^t(\sigma_v))]. \quad (24)$$

Algorithm 1 Adaptive reinforcement learning with private information feedback

```
For user $m_v$ at time slot $t$, assume $\mathrm{U}(0,1)$ represents a uniform distribution from 0 to 1.
Initialization: Set $J_v^{prev} = 0$, $\omega_v = 1$, $\triangle\omega_v = 0.05$.
Step 1. If $Rand(\mathbf{U}(0,1)) < 1-\omega_v$, keep using action
        $A_v^t = A_v^{base}$, $t \leftarrow t+1$, and repeat Step 1,
        otherwise go to Step 2.
Step 2. Calculate $u_v(A_v^{t-1}, \gamma_v^{t-1})$ from previous action
        $A_v^{t-1} = [f_v^{t-1}, P_v^{t-1}]$ and the private
        information feedback $\mathcal{I}_v^{t,priv} = \{\gamma_v^{t-1}\}$.
Step 3. Update the propensity $\mathbf{r}_v^t$ and the strategy $\mathbf{S}_v^t$.
Step 4. Determine the action from $A_v^t = Rand(\mathbf{S}_v^t)$.
Step 5. Update the baseline action $A_v^{base}$ and baseline
        payoff value $u_v^{base}$ as in equation (21) and (22)
Step 6. Evaluate $J_v$. If $J_v > J_v^{prev}$, then
        if $\omega_v - \triangle\omega_v > 0$, $\omega_v \leftarrow \omega_v - \triangle\omega_v$,
        else if $\omega_v - \triangle\omega_v \le 0$, keep $\omega_v$.
        Otherwise, if $\omega_v + \triangle\omega_v \le 1$, $\omega_v \leftarrow \omega_v + \triangle\omega_v$,
        else if $\omega_v - \triangle\omega_v > 1$, keep $\omega_v$.
Step 7. Set $J_v^{prev} \leftarrow J_v$, $t \leftarrow t+1$, and go back to
        Step 1.
```

Model-based learning [19] provides a method to build the belief on $\tilde{\mathbf{S}}_{-v}^t(\sigma_v)$ of other users' actions from the past experienced public information $A_u^{s-1}, s \in \mathbf{T}_v^t$. We present the action learning that performs equation (24) as an example in Section V.*A*.

Similarly, if the public information feedback is costless (i.e. $B_v' = B_v$ in equation (11)), the utility upper bound of the model-based learning can be calculated as discussed below.

Proposition 3. Performance upper bound with public information feedback: For the model-based learning based on the public information feedback, if the information feedback is costless, the upper bound of the learning efficiency $J_v(\Lambda_v^{pub})$ is $U_v(\mathbf{A}_{-v}^{fors})$.

Proof: Substitute equation (24) into equation (23) and substitute $u_v(\sigma_v, A_v, A_{-v})$ by $u_v(A_v, \mathbf{A}_{-v}^{fors})$. And this provides an upper bound on $J_v(\Lambda_v^{pub})$, since $u_v(A_v, \mathbf{A}_{-v}^{fors}) \ge u_v(\sigma_v, A_v, A_{-v})$. Equation (23) then becomes

$$\max_{A_v \in \mathcal{A}_v} \left(u_v(A_v, \mathbf{A}_{-v}^{fors}) \left(\sum_{A_{-v} \in \mathcal{A}^{V\text{-}1}} \tilde{S}_{-v}^t(A_{-v} \mid A_v) \right) \right). \quad (25)$$
$$= u_v(A_v^{fors}, \mathbf{A}_{-v}^{fors}) = U_v(\mathbf{A}_{-v}^{fors})$$

The reason why the model-based learning with public information feedback has a higher upper bound compared

[7] In [13], the authors focused on developing a reinforcement learning algorithm that guarantees convergence without considering the cost of the private information feedback. Our AR scheme employs reinforcement learning *while* considering the cost of the information feedback and also adapts the information feedback frequency to maximize the user's utility.

to the payoff-based learning with private information feedback is because it enables the user to explicitly model the actions of other users and hence, the user can directly choose the action that maximizes its expected utility. Next, we provide a simple model-based learning – action learning, which is similar to the well-known fictitious play [19].

A. *Action learning based on public information feedback*

Recall that in order to build the belief $\tilde{\mathbf{S}}_{-v}^{t}$ from $o_v^t = \{\mathcal{I}_{-v}^{s,pub}, s \in \mathbf{T}_v^t\}$, user m_v maintains a set of strategy vectors $\tilde{S}_{-v}^{t}(A_{-v} \mid A_v) = \{\tilde{S}_u^t(A_u \in \mathcal{A}_u \mid A_v), \text{for } m_u \in \mathbf{M}_{-v}\}$ for all possible actions $A_v \in \mathcal{A}_v$, where $\tilde{S}_u^t(A_u \in \mathcal{A}_u \mid A_v) = [\tilde{S}_u^t(A_u \mid A_v), \text{ for } A_u \in \mathcal{A}_u]$ represents the estimated strategy of the user $m_u \in \mathbf{M}_{-v}$ given that user m_v taking action A_v at time slot t. Hence, in the action learning, whenever action A_v is taken by the user m_v, we set

$$\tilde{S}_u^t(A_u \mid A_v) = \frac{r_u^t(A_u \mid A_v)}{\sum_{A \in \mathcal{A}_u} r_u^t(A \mid A_v)}, \tag{26}$$

where $r_u^t(A_u \mid A_v)$ is the *propensity* of user m_u at time t. The propensity represents the number of times that user m_u takes action A_u given that user m_v took action A_v. Hence, whenever the action A_v is taken by user m_v, the vector $\mathbf{r}_u^t(A_u \in \mathcal{A}_u \mid A_v) = [r_u^t(A_u \mid A_v), \text{ for all } A_u \in \mathcal{A}_u]$ is updated by:

$$\mathbf{r}_u^t(A_u \in \mathcal{A}_u \mid A_v) = \mathbf{r}_u^{t-1}(A_u \in \mathcal{A}_u \mid A_v) + \mathbf{I}(A_u^{t-1}) . \tag{27}$$

Then, the probability $\tilde{S}_u^t(A_u \mid A_v)$ represents the *empirical frequency* that user m_u will take an action $A_u \in \mathcal{A}_u$ given that user m_v took an action A_v.

Next, we show how to maximize $E_{\tilde{\mathbf{S}}_{-v}^t}[u_v(A_v, \tilde{\mathbf{S}}_{-v}^t(\sigma_v))]$ in equation (24) analytically given the belief $\tilde{\mathbf{S}}_{-v}^t$. First, we show the necessary condition for user m_v to maximize its utility function.

<u>*Proposition 4. Target SINR values:*</u> For a certain frequency channel f, in order to maximize $u_v(f)$, user m_v needs to transmit at the target SINR value $\gamma_v^{tar}(f)$, which is the unique positive solution of $\gamma \frac{\partial B_v'(\gamma)}{\partial \gamma} = \frac{L_v}{d_v}(F_v(\gamma) - 1)$ ($F_v(\gamma)$ is in equation (12)).

<u>*Proof:*</u> See Appendix II.

Proposition 4 suggests that if user m_v is using the frequency channel f, it should adapt the target power level $P_v^{tar}(f)$ accordingly to the interference from the other users using the same frequency channel to support the target SINR value $\gamma_v^{tar}(f)$. Since the power level in our setting is discrete, we choose the $P_v^{tar}(f) \in \mathcal{P}_v$ as the power that provides the nearest SINR value to $\gamma_v^{tar}(f)$. If the target SINR $\gamma_v^{tar}(f)$ requires a power higher than $P_v^{\max}$ (when the interference in the channel is too high), then set $P_v^{tar}(f)$ to $P_v^{\max}$.

Next, given the target $P_v^{tar}(f)$, we further determine the optimal frequency channel selection of the user m_v.

<u>*Proposition 5. Optimal actions given the target SINR values:*</u> Let $F_v^{tar}(f) = F_v(f, \gamma_v^{tar}(f))$ in equation (12). Given the corresponding target $P_v^{tar}(f)$, the optimal action A_v^* of a user m_v is

$$f_v^* = \arg\min_{f \in \mathcal{F}_v} \{P_v^{tar}(f) \frac{F_v^{tar}(f)}{F_v^{tar}(f) - 1}\} \quad \text{and}$$
$$P_v^* = P_v^{tar}(f_v^*) . \tag{28}$$

<u>*Proof:*</u> From Proposition 4, maximizing $u_v = \frac{\lambda_v}{P_v}(1 - \frac{1}{F_v})$ leads to equation (28).

In summary, user m_v selects the frequency channel f_v^* and power level P_v^* to support the target SINR $\gamma_v^{tar}(f_v^*)$, which maximizes the utility function in equation (2). This requires user m_v to estimate the interference from other users, which can be computed by user m_v based on its belief $\tilde{\mathbf{S}}_{-v}^t$. Specifically, denote the estimated interference of user m_v as $\Omega_v(A_v)$, when the action A_v is taken. Given $\tilde{\mathbf{S}}_{-v}^t$, $\Omega_v(A_v)$ can be computed as:

$$\Omega_v(A_v) = \sum_{\substack{u \neq v \\ A_u \in \mathcal{A}_u}} G_{uv}(f_v)[\tilde{S}_u^t(A_u \mid A_v) P_u I(f_u = f)] . \tag{29}$$

Then, the resulting SINR value $\gamma_v(A_v)$ is ($A_v = [f_v, P_v]$):

$$\gamma_v(f_v, P_v) = \frac{G_{vv}(f_v) P_v}{N_{f_v} + \Omega_v(A_v)} . \tag{30}$$

By applying Proposition 4, we calculate the target power $P_v^{tar}(f)$ in different frequency channels:

$$P_v^{tar}(f) = \min_{P \in \mathcal{P}_v} \left| \gamma_v^{tar}(f) - \gamma_v(f, P) \right| . \tag{31}$$

Then we apply Proposition 5 to determine $A_v^t = [f_v^*, P_v^*]$.

B. *Adaptive action learning*

For the action learning in the previous subsection, the public information feedback $\mathcal{I}_{-v}^{t,pub} = \{\mathbf{G}_u^{t-1}, A_u^{t-1},\ m_u \in \mathbf{M}_{-v}\}$ is required from every user in the network, during each time slot. This results in heavy information overhead. Moreover, the overall action space $\mathcal{A}^{V-1}$ makes the computational complexity prohibitive to model all the users in the network. To approach the upper bound $U_v(\mathbf{A}_{-v}^{fors})$ of the model-based learning efficiency, we need to adjust the information overhead $\sigma_v(\omega_v, |\mathbf{V}_v|)$ by changing the information feedback parameters ω_v and $|\mathbf{V}_v|$.

Hence, in our proposed active action learning, to reduce the overhead, we classify the neighboring users of user m_v into H groups ($1 \le H \le |\mathbf{M}_{-v}|$) and assign different information feedback frequency ω_v^i to different groups (i.e. $1 \ge \omega_v^1 \ge \omega_v^2 \ge ... \ge \omega_v^H \ge 0$). For the dynamic power/spectrum management problem in this paper, the neighboring users can be classified based on their average channel gains $\bar{G}_{uv}$ over the frequency channels, i.e. $\bar{G}_{uv} = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} G_{uv}(f)$ (from the transmitter of the neighboring user m_u to the receiver of the foresighted user m_v), since these channel gains directly impact the user's utility (see equation (1) and (2)). For instance, a neighboring user m_u with a larger channel gain $\bar{G}_{uv}$ will have more impact on u_v.

Let X_v^i represents the number of users in the group $H_i, i = 1,...,H$. Assume the neighboring users are relabeled according to its average channel gain value, i.e. $\bar{G}_{[1]v} \ge \bar{G}_{[2]v} \ge ... \ge \bar{G}_{[V-1]v}$. Then,

$$m_{[u]} \in H_i, \text{ iff } \sum_{j=1}^{i-1} X_v^j \le [u] \le \sum_{j=1}^{i} X_v^j . \qquad (32)$$

In Algorithm 2, we provide our adaptive action learning approach for the extreme case when $H = 2$ as an example. In this case, we only need to adapt $|\mathbf{V}_v|$ ($X_v^1 = |\mathbf{V}_v|$ and $X_v^2 = V - 1 - |\mathbf{V}_v|$). If the neighboring users $m_u \in \mathbf{V}_v$, we set $\omega_v = 1$, otherwise, $\omega_v = 0$. Meaning that user m_v only needs to model the users in set $\mathbf{V}_v$ based on $\mathcal{I}_{-v(\mathbf{V}_v^t)}^{t,pub} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, \ m_u \in \mathbf{V}_v\}$. In Table I, we compare the two proposed interactive learning algorithms.

TABLE I
COMPARISONS OF THE PROPOSED LEARNING ALGORITHMS

	Info. feedback	Build belief on	Adapt to	Performance upper bounds
Adaptive Reinforcement Learning (payoff-based)	Private	Own utility $\tilde{u}_v$	Other users' actions A_{-v}, information feedback frequency ω_v	$(1-\varepsilon_v) \times U_v(\mathbf{A}_{-v}^{fors})$
Adaptive Action Learning (model-based)	Public	Other users' strategies $\tilde{\mathbf{S}}_{-v}$	Other users' actions A_{-v}, number of neighbor users $\vert\mathbf{V}_v\vert$	$U_v(\mathbf{A}_{-v}^{fors})$

Algorithm 2 Adaptive action learning (H =2) with public information feedback

```
For user $m_v$ at time slot $t$,
Initialization: Set $J_v^{prev} = 0$, $|\mathbf{V}_v| = |\mathbf{M}_{-v}|$,
                $\triangle|\mathbf{V}_v| = 1$.
Step 1. Observe the public information feedback
        $\mathcal{I}_{-v(\mathbf{V}_v)}^{t,pub} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, \ m_u \in \mathbf{V}_v\}$ fed back
        from the users $m_u \in \mathbf{V}_v$.
Step 2. Update the propensity $\mathbf{r}_u^t$ for users $m_u \in \mathbf{V}_v$
        and calculate the strategy vector $\tilde{\mathbf{S}}_{-v}^t(A_v)$.
Step 3. Calculate the target power $P_v^{tar}(f)$ from
        equation (31) and find the action
        $A_v^t = [f_v^*, P_v^*]$ using Proposition 5.
Step 4. Evaluate $J_v$. If $J_v > J_v^{prev}$, then
        if $|\mathbf{V}_v| - \triangle|\mathbf{V}_v| > 0, |\mathbf{V}_v| \leftarrow |\mathbf{V}_v| - \triangle|\mathbf{V}_v|$,
        else if $|\mathbf{V}_v| - \triangle|\mathbf{V}_v| \le 0$, keep $|\mathbf{V}_v|$.
        Otherwise, if $|\mathbf{V}_v| + \triangle|\mathbf{V}_v| \le |\mathbf{M}_{-v}|$,
            $|\mathbf{V}_v| \leftarrow |\mathbf{V}_v| + \triangle|\mathbf{V}_v|$,
        else if $|\mathbf{V}_v| + \triangle|\mathbf{V}_v| > |\mathbf{M}_{-v}|$, keep $|\mathbf{V}_v|$.
Step 5. Set $J_v^{prev} \leftarrow J_v$, $t \leftarrow t + 1$, and go back to
        Step 1.
```

VI. SIMULATION RESULTS

We simulate an ad hoc wireless network environment shown in Figure 6 with 5 users (distinct transmitter-receiver pairs) and 3 frequency channels. The frequency channels are accessible for all the users, i.e. $\mathcal{F}_v = \mathcal{F}$, for $\forall m_v$. Each user can choose its power level P_v from a set $\mathcal{P} = \{20, 40, 60, 80, 100\}$ (mW). Hence, there are a total of 15 actions A_v for users to adapt. At the physical layer, we model the channel gain between different network nodes using $G_{vv'} = K_0 ({dis_{vv'}}/{dis_0})^{-\alpha}$ for all frequency channels, where $dis_{vv'}$ represents the distance from the transmitter of the user m_v to the receiver of the user $m_{v'}$, and $K_0 = 5 \times 10^{-4}$, $N_f = 1 \times 10^{-5}$, $dis_0 = 10$, $\alpha = 2$ are constants. For the application layer parameters, we set the average packet length L_v = 1000 bytes, input rate R_v = 500 Kbps ($\lambda_v = R_v / L_v$), and delay deadline $d_v = 200$ msec for all the users. The effective transmission rate $B_v'(\gamma_v) = T(1 - p_v(\gamma_v))\theta(\sigma_v)$, where $p_v(\gamma_v)$ represents the packet error rate (see Appendix I).

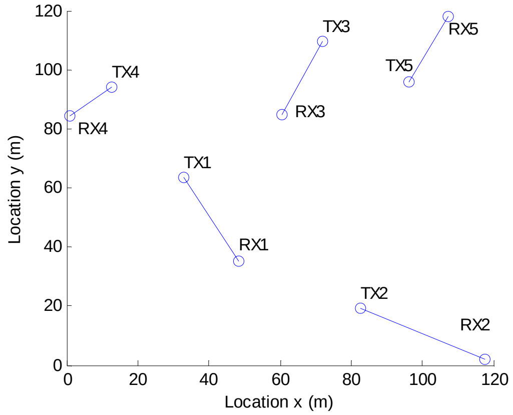

Fig. 6 Topology settings for the simulation.

A. Comparison among different learning approaches based on information feedback

We show the simulation results using five different schemes when the physical transmission rate are T = 700 Kbps and 2100 Kbps in Table II and III, respectively. The five schemes are – 1) the centralized optimal (CO) 2) the theoretical upper bound $U(\mathbf{A}_{-v})$ (UB) 3) myopic best response without learning (NE), 4) user m_1 adopting adaptive reinforcement learning with private information feedback in Algorithm 1 (AR), and 5) user m_1 adopting adaptive action learning with public information feedback in Algorithm 2 (AA). The CO scheme provides the global optimal results for the overall utilities. In the NE scheme, each user attempt to maximize its current utility function based on the actions they observe in the previous time slot as in equation (3). The UB is computed from equation (4) for m_1 given the exact response of the other four users ($u_1 = U(\mathbf{A}_{-1}^{fors})$). Since, the user m_1 is in the middle of the topology, we select m_1 to be the foresighted user who learns from the information feedback. Each simulation result is averaged over 500 time slots in the dynamic network settings with mutual interference in equation (1).

Table II shows that user m_1 stays in channel 1 in both the CO and UB scheme while the other four users using the rest of two channels. However, since users are self-interested, NE scheme shows that user m_5 also attempts to transmit in channel 1 and hence, the utility u_1 decreases and forces user m_1 to increase its power level. If user m_1 becomes foresighted, as shown in the AR scheme, it will keep using the highest power level to prevent user m_5 from using its channel. The resulting utility u_1 is higher than the NE scheme. Using the AA scheme, users are able to exploit the spectrum more efficiently, due to the ability that the users can better model the strategies of other interference sources in the network. However, this requires significant information overhead, which results in a worse performance at low bandwidth, i.e. when T = 700 Kbps. Note that although only user m_1 is learning, the average utility of using interactive learning schemes outperforms the myopic NE scheme. Even in a non-cooperative setting, this foresighted user actually benefits the overall system performance.

When T = 2100 Kbps, Table III shows that users are now selecting a lower power levels, since the physical transmission bandwidth is sufficient. Using the AR scheme, user m_1 again occupies channel 1 by using higher power level compared to the UB scheme. Note that using the AA scheme, user u_1 can almost reach the theoretical upper bound, since the cost of information feedback is comparatively small when T = 2100 Kbps. Again, the average utilities of the adaptive interactive learning schemes outperform the myopic NE scheme. The higher T gives a better learning environment for the user m_1 using AA scheme to approach the theoretical upper bound $U_1(\mathbf{A}_{-v}^{fors})$ than using AR scheme. Since all the users are selfish (including user m_1 who is learning), the learning user m_1 will benefit itself by suppressing the utility of m_2 as shown in Table III. This situation is not seen in Table II, since the AA scheme has low learning efficiency when the T is small.

TABLE II
SIMULATION RESULTS OF THE FIVE SCHEMES WHEN T = 700 KBPS

T = 700 Kbps		Actions $A_v = [f_v, P_v]$ (or strategies $\mathbf{S}_v$)	u_v (Kbit/joule)	$\sum_{v=1}^{5} u_v / 5$
1) Centralized Optimal (CO)	m_1	[1,3]	1022.8	1420.8
	m_2	[2,1]	0	
	m_3	[3,2]	1479.5	
	m_4	[2,1]	3096.7	
	m_5	[2,2]	1499.8	
2) Theoretical Upper Bound (UB)	m_1	[1,3]	$U_1(\mathbf{A}_{-v}^{fors})$ = 1022.8	1285.0
	m_2	[2,4]	0	
	m_3	[2,4]	765.3	
	m_4	[3,1]	3100.8	
	m_5	[3,2]	1536.1	
3) Myopic Best Response (NE)	m_1	[1,3] x 65%, [1,5] x 35%	519.0	890.15
	m_2	[2,5] x 65%, [3,5] x 35%	195.2	
	m_3	[3,2]x33%,[3,3]x33%, [3,5]x33%	530.6	
	m_4	[2,1]x65%, [3,1]x35%	2073.0	
	m_5	[2,2]x33%,[2,3]x33%, [1,3]x33%	1132.9	
4) Adaptive Reinforcement Learning at m_1 (AR)	m_1	[1,5]	555.2	1005.6 (ω_v = 0.7)
	m_2	[2,5]	113.5	
	m_3	[3,5]	345.6	
	m_4	[2,1]	2830.2	
	m_5	[2,3]	1183.7	
5) Adaptive Action Learning at m_1 (AA)	m_1	[1,3]x65%,[1,4]x27%, [1,5]x8%	529.3	1039.3 ($\lvert \mathbf{V}_v \rvert$ = 2)
	m_2	[2,5] x 85%, [3,5] x 15%	445.6	
	m_3	[3,2]x45%,[3,3]x45%, [3,5]x10%	446.8	
	m_4	[2,1]x50%, [3,1]x50%	2771.2	
	m_5	[2,2]x10%,[2,3]x10%, [1,3]x80%	1003.3	

B. Convergence of the learning schemes

In order to show the convergence of the proposed learning schemes, in Figure 7, we simulate the time plot of the two proposed learning algorithms (AR and AA) and the best response scheme without learning (NE). The network settings are the same as Table II when T = 700 Kbps. It is shown that both the two proposed learning schemes outperform the myopic best response scheme in terms of the average utility. The convergence speed of the AR scheme is about three times slower than the myopic best response (which converges to Nash equilibrium in about 5 time slots), while the AA scheme is about six times slower. The convergence speed of the AR scheme is faster than the AA scheme, since the AR scheme only need to build belief on its own utility. The AA scheme needs to build beliefs on its neighboring users' strategies, which leads to a slower convergence speed.

TABLE III
SIMULATION RESULTS OF THE FIVE SCHEMES WHEN T = 2100 KBPS

T = 2100 Kbps		Actions $A_v = [f_v, P_v]$ (or mixed strategies $\mathbf{S}_v$)	u_v (Kbit/joule)	$\sum_{v=1}^{5} u_v / 5$
1) Centralized Optimal (CO)	m_1	[1,2]	1562.2	1718.7
	m_2	[2,4]	781.2	
	m_3	[3,2]	1562.5	
	m_4	[2,1]	3125.0	
	m_5	[2,2]	1562.5	
2) Theoretical Upper Bound (UB)	m_1	[1,2]	$U_1(\mathbf{A}_{-v}^{fors})$ = 1562.2	1458.3
	m_2	[2,3]	76.8	
	m_3	[2,3]	1041.7	
	m_4	[3,1]	3125.0	
	m_5	[3,2]	1562.5	
3)Myopic Best Response (NE)	m_1	[1,2]x25%,[1,3]x25%, [2,2]x25%,[2,3]x25%	523.4	1380.7
	m_2	[1,3]x25%,[1,4]x25%, [2,3]x25%,[2,4]x25%	390.6	
	m_3	[1,2]x25%,[1,3]x25%, [2,2]x25%,[2,3]x25%	1302.1	
	m_4	[3,1]	3125.0	
	m_5	[3,2]	1562.5	
4) Adaptive Reinforcement Learning at m_1 (AR)	m_1	[1,3]	1018.2	1503.6 (ω_v = 1)
	m_2	[2,4]	757.8	
	m_3	[2,3]	1054.7	
	m_4	[3,1]	3125.0	
	m_5	[3,2]	1562.5	
5) Adaptive Action Learning at m_1 (AA)	m_1	[1,2]x50%,[2,2]x50%	1549.1	1455.7 ($\lvert \mathbf{V}_v \rvert$ = 4)
	m_2	[1,3] x50%,[2,3]x50%	0	
	m_3	[1,3] x50%,[2,3]x50%	1041.7	
	m_4	[3,1]	3125.0	
	m_5	[3,2]	1562.5	

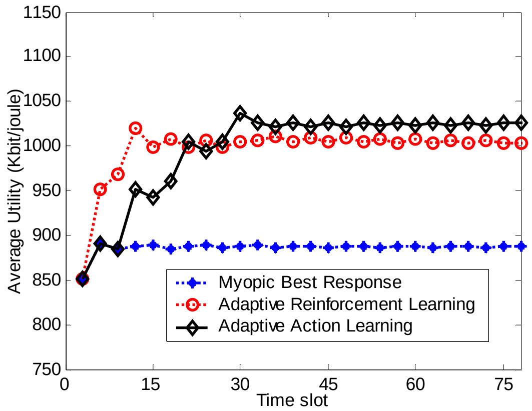

Fig. 7 Average utility vs. time slot of the proposed algorithms when T = 700 Kbps (a time slot is considered to be 10ms).

C. *Adaptive reinforcement learning using different time scales*

The reinforcement learning is very sensitive to the initial status of users' actions. Hence, in our simulations, we first train the user m_1 's initial strategy by performing myopic best response in the first 20 time slots. Then, we simulate the reinforcement learning with different values of ω_v in Figure 8 for different T. Since the input rates of the applications are fixed to 500 Kbps, the utility will saturate as the bandwidth increases. The UB scheme has another saturation when T becomes larger than 1.1 Mbps, since the larger bandwidth enables another set of actions for the users. Note that when $\omega_1 = 1$, the reinforcement learning learns the transmission strategy $\mathbf{S}_1^t$ at every time slot. The simulation results show that the performance of $\omega_1 = 0.8$ is better than $\omega_1 = 1$ when the physical bandwidth is lower than 1Mbps, since learning at a slower pace can reduce the overhead of the private information feedback. The results in Figure 8 show that the proposed adaptive reinforcement learning operates on the envelope of the solutions obtained for different ω_1, with $\omega_1 \in [0.5, 1]$. Hence, the performance of user m_1 using the adaptive reinforcement learning becomes closer to the upper bound.

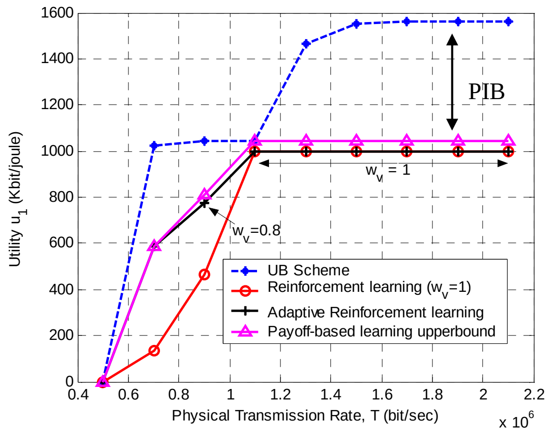

Fig. 8 Performance of user m_1 adopting adaptive reinforcement learning with private information feedback using different ω_1.

D. *Adaptive action learning from different neighboring users*

In Figure 9, we also simulate the case that the action learning models the strategy of the nearest $|\mathbf{V}_v| = 2$ users instead of $|\mathbf{M}_{-v}| = 4$ users. With smaller $|\mathbf{V}_v|$, fewer neighbors need to feed back information and hence, results in less information overhead. The simulation results show that modeling users from public information feedback can improve the performance for user m_1. However, when the physical transmission rate is lower than 1.1 Mbps, the required information overhead degrades the performance significantly and hence, it is essential to adapt the number of neighbors in the action learning to model less users in the network. The results show that using the proposed adaptive action learning, the performance of user m_1 with public information feedback becomes closer to the upper bound.

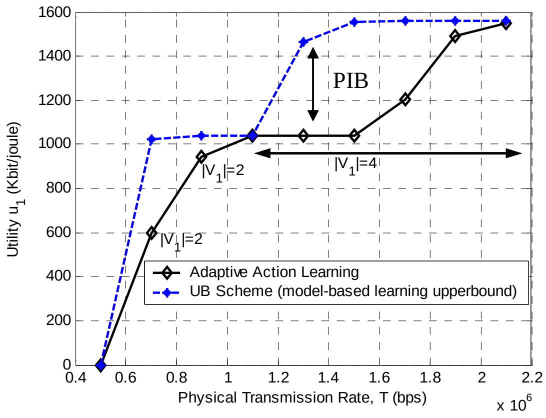

Fig. 9 Performance of user m_1 adopting adaptive action learning with public information feedback using different $|\mathbf{V}_1^t|$.

E. *Mobility effect on the interactive learning efficiency*

In the previous subsections, all the simulation results are based on the fixed topology shown in Figure 6. In this subsection, we simulate the case that all 5 receivers moves according to a well-known mobility model "random walk" [22] – receivers randomly select a direction at each time slot and move at a fixed speed ν. Starting from the topology in Figure 6, Figure 10 shows the learning efficiency over time of the AR, AA, and NE schemes for ν =0.5, 1, 2 (meter/sec) with T = 2100 Kbps. It is shown that the AA scheme has higher learning efficiency on average, since user m_1 is able to obtain the channel gain information (which is directly affected by the mobility) of the other users from the public information feedback. Moreover, as expected, as the mobility increases, the learning efficiency decreases because the receivers are moving further apart. Especially for the reinforcement learning without explicit channel gain information, the results show that the performance can be worse than myopic best response, since the learning cannot keep up with the topology changes and the user's belief about the other users becomes inaccurate when the mobility is high.

VII. Conclusions

In this paper, we provide an adaptive interactive learning framework for delay sensitive users to adapt their frequency channel selections and power levels in wireless networks in a decentralized manner. We show that a foresighted user can improve its utility significantly by learning from the information feedback. We determine performance upper bounds for the user's utility when learning from private or public information feedback, respectively. The simulation results show that the proposed adaptive interactive learning can significantly improve the performance of delay sensitive users compared to the myopic best response. It is shown that even when only one user learns from its information feedback, the overall performance can be better than the Nash equilibrium resulting from the myopic best response. Especially, if the available system bandwidth is not limited, the proposed adaptive action learning with public information feedback approaches the utility upper bound.

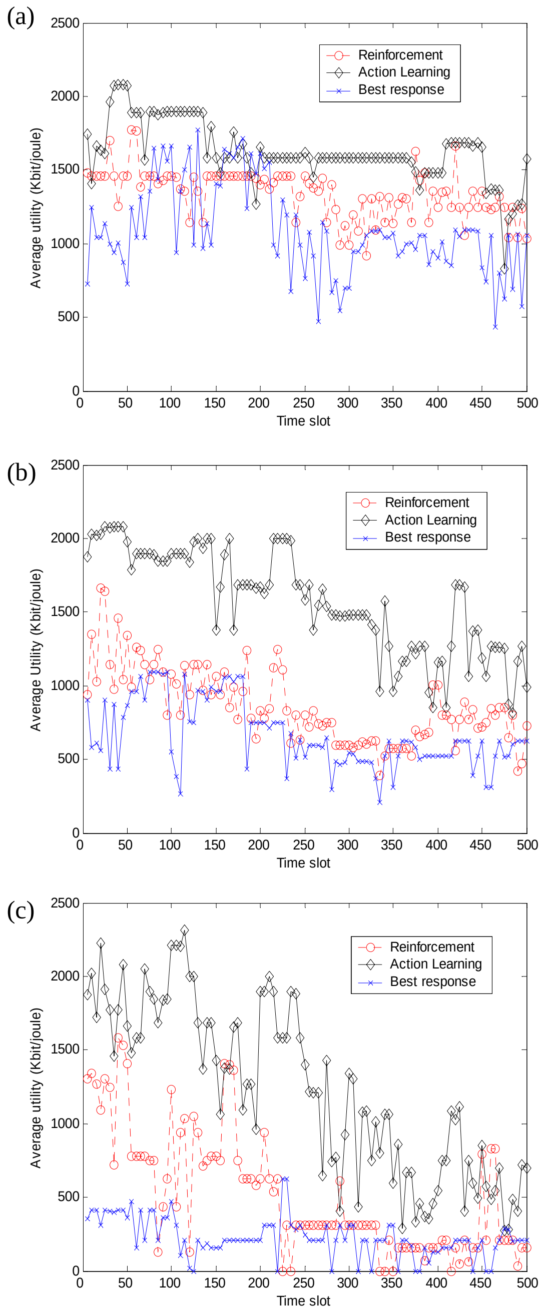

Fig. 10 Average utility over time using the adaptive interactive learning when receivers have mobility (T = 2100 Kbps) (a) $\nu = 0.5$, (b) $\nu = 1$, (c) $\nu = 2$ (m/sec).

Appendix I

Recall that T_v and p_v represent the maximum transmission rate and packet error rate of user m_v using the frequency channel f_v. T_v and p_v are estimated by the MAC/PHY layer link adaptation, which can be modeled as sigmoid functions of the SINR $\gamma_v(A_v, A_{-v})$ for user m_v:

$$p_v(f_v, \gamma_v(A_v, A_{-v})) = \frac{1}{1+\exp(\zeta(\gamma_v(A_v, A_{-v}) - \delta))}, \quad (33)$$

$$B_v(A_v, A_{-v}) = T_v(f_v)(1 - p_v(f_v, \gamma_v(A_v, A_{-v}))), \quad (34)$$

$B'_v(\sigma_v, A_v, A_{-v}) = B_v(A_v, A_{-v})\theta(\sigma_v)$, and $\theta(\sigma_v) = 1 - \rho\omega_v(|\mathbf{V}_v| + 1)$, where ζ , δ , and $\rho > 0$ are empirical constants corresponding to the modulation and coding schemes for a given packet length.

Assume that a delay sensitive application is sent by the user m_v through the network with the average input rate R_v (bits/sec). Assume that the user m_v maintains a queue with infinite buffer size in the application layer. We model the packet arrival process using Poisson process. The packet arrival rate is assumed as $\lambda_v = R_v / L_v$ (packet/sec). Considering the packet protection scheme similar to the Automatic Repeat Request protocol in IEEE 802.11 networks [14], the transmission time of a packet can be modeled as a geometric distribution [15]. For simplicity, we approximate the queuing model as M/M/1 queue with the service rate $\mu_v(\sigma_v, A_v, A_{-v}) = B'_v(\sigma_v, A_v, A_{-v}) / L_v$ (packet/sec). Denote the delay of transmitting the delay sensitive application through the network as $D_v(\sigma_v, A_v, A_{-v})$. The average delay can be obtained by

$$E[D_v(\sigma_v, A_v, A_{-v})] = \frac{1}{\mu_v(\sigma_v, A_v, A_{-v}) - \lambda_v}, \text{ for } \mu_v(\sigma_v, A_v, A_{-v}) > \lambda_v . \quad (35)$$

Using the M/M/1 queuing model, the probability that the packet of the user m_v can be received before the delay deadline d_v is

$$\mathrm{Prob}\{D_v(\sigma_v, A_v, A_{-v}) \le d_v\} = \begin{cases} 1 - \exp(-\dfrac{d_v}{E[D_v(\sigma_v, A_v, A_{-v})]}), & \text{for } \mu_v(\sigma_v, A_v, A_{-v}) > \lambda_v \\ 0 & , \text{otherwise} \end{cases} \quad (36)$$

The utility function in (2) equals to 0 unless the transmitted power is high enough to support a sufficient throughput $B'_v(\sigma_v, A_v, A_{-v})) / L_v > \lambda_v$ to keep the probability $\mathrm{Prob}\{D_v(\sigma_v, A_v, A_{-v}) \le d_v\} > 0$ (see Figure 2). Substituting equation (35) and (36) into equation (2), we have equation (11). Since $B'_v(\sigma_v)$ is a decreasing function of σ_v , the utility function is a non-increasing function of σ_v .

APPENDIX II

<u>*Proof of Proposition 4:*</u> Given the channel model $B_v(f, \gamma)$ for the frequency channel f in equation (34), user m_v , $m_v \in \Omega_f$ can apply queuing analysis with the application characteristics R_v , L_v and d_v . From equation (35) and (36), we have $\mathrm{Prob}\{D_v \le d_v\} = 1 - \frac{1}{F_v(\gamma_v)}$. The optimality condition of $\frac{\partial u_v}{\partial P_v} = 0$ becomes $-P_v \times \frac{\partial}{\partial P_v}\frac{1}{F_v(\gamma_v)} = 1 - \frac{1}{F_v(\gamma_v)}$. The left hand side can be derived as $\gamma_v \frac{\partial B_v(\gamma_v)}{\partial \gamma_v} \times \frac{d_v}{L_v}\frac{1}{F_v(\gamma_v)}$, since $P_v \frac{\partial \gamma_v}{\partial P_v} = \gamma_v$. By multiplying F_v to both sides, we have the optimality condition in Proposition 4 and the corresponding γ_v^{tar} that maximizes the utility function u_v .

REFERENCES

[1] M. J. Neely, E. Modiano, and C. E. Rohrs, “Dynamic Power Allocation and Routing for Time-Varying Wireless Networks”, *IEEE Journal on Selected Areas in Communications*, vol. 23. no1, Jan 2005.

[2] S. G. Kiani, G. E. Oien, D. Gesbert, “Maximizing multi-cell capacity using distributed power allocation and scheduling,” *IEEE Wireless Communications and Networking Conference, WCNC 2007*, pp. 1690-1694, Mar 2007.

[3] S. Singh, D. Bertsekas, “Reinforcement learning for dynamic channel allocation in cellular telephone systems,” In *Advances in Neural Information Processing Systems*, pp. 974-980, Cambridge MA, 1997.

[4] J. Zhao, H. Zheng, G.-H. Yang, “Distributed coordination in dynamic spectrum allocation networks,” in *Proc. IEEE DySPAN 2005*, Nov 2005, pp. 259-268.

[5] D. Gesbert, S. G. Kiani, A. Gjendemsjo, and G. E. Oien, “Adaptation, Coordination, and Distributed Resource Allocation in Interference-Limited Wireless Networks,” *Proceeding of IEEE*, vol. 95, no. 12, pp. 2393-2409, Dec 2007.

[6] F. Fu and M. van der Schaar, "Non-collaborative resource management for wireless multimedia applications using mechanism design," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 851-868, Jun. 2007.

[7] W. Yu, R. Lui, “Dual Methods for Nonconvex Spectrum Optimization of Multi-carrier Systems,” *IEEE Transactions on Communications*, vol. 54, no. 7, July 2006.

[8] W. Yu, G. Ginis, and J. M. Cioffi, “Distributed multi-user power control for digital subscriber lines,” *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 1105-1115, Jun. 2002.

[9] F. Meshkati, M. Chiang, H. V. Poor, and S. C. Schwartz, “A game-theoretic approach to energy-efficient power control in multi-carrier CDMA systems,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 24, pp. 1115-1129, June 2006.

[10] D. J. Goodman and N. B. Mandayam, “Power control for wireless data,” *IEEE Personal Communications*, vol. 7, pp. 48-54, Apr 2000.

[11] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, “Efficient power control via pricing in wireless data networks,” *IEEE Trans. on Commun.*, vol. 50, no. 2, pp. 291-303, Oct 2002.

[12] M. Xiao, N. B. Shroff, and E. J. P. Chong, “A utility-based power-control scheme in wireless cellular systems,” *IEEE/ACM Transactions on Networking*, vol. 11, pp. 210-221, Apr 2003.

[13] C. Long, Q. Zhang, B. Li, H. Yang, and X. Guan, “Non-cooperative power control for wireless ad hoc networks with repeated games,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 25, no. 6 pp.

1101-1112, Aug. 2007.

[14] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall, 2002.

[15] A. G. Konheim, “A Queuing Analysis of Two ARQ Protocols,” *IEEE Transactions on Communications*, vol. com-28, no. 7, July 1980.

[16] D. Bertsekas, R. Gallager, *Data Networks*, Prentice Hall, Inc. Upper Saddle River, NJ, 1987.

[17] H. Shiang and M. van der Schaar, “Informationally Decentralized Video Streaming over Multi-hop Wireless Networks,” *IEEE Trans. Multimedia*, vol. 9, issue 6, Sep 2007.

[18] S. Lal, E. S. Sousa, “Distributed resource allocation for DS-CDMA-based multimedia ad hoc wireless LANs,” *IEEE J. Sel. Areas Commun.*,vol. 17, no. 5, pp. 947-967, May 1999.

[19] H. P. Young, *Interactive learning and its Limits*, Oxford University Press, NY 2004.

[20] Y. Su and M. van der Schaar, "A New Look at Multi-user Power Control Games," in *Proc. Int. Conf. Commun.* (ICC 2008) June 2008.

[21] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, Cambridge, MA: MIT Press, 1998.

[22] T. Camp, J. Boleng, V. Davies, “A survey of mobility models for ad hoc network research,” in *Wireless Communications and Mobile Computing (WCMC)*, vol. 2, no. 5, pp. 483-502, 2002.

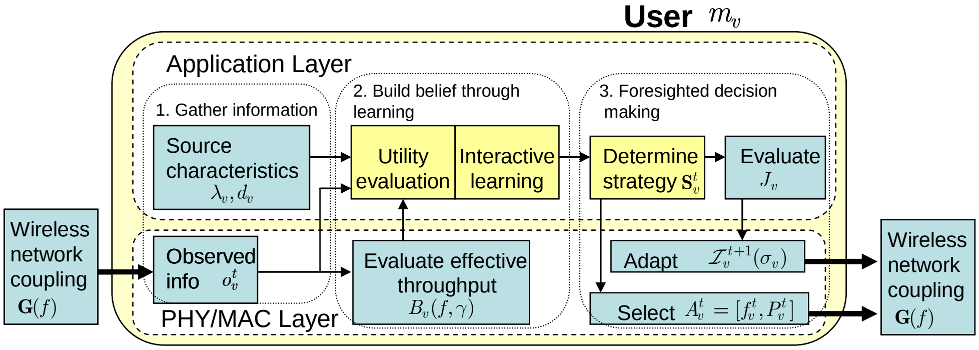
User m_v
Application Layer
1. Gather information
2. Build belief through learning
3. Foresighted decision making
Source characteristics λ_v, d_v
Utility evaluation
Interactive learning
Determine strategy $\mathbf{S}_v^t$
Evaluate J_v
Wireless network coupling $\mathbf{G}(f)$
Observed info o_v^t
Evaluate effective throughput $B_v(f,\gamma)$
Adapt $\mathcal{I}_v^{t+1}(\sigma_v)$
Select $A_v^t = [f_v^t, P_v^t]$
PHY/MAC Layer
Wireless network coupling $\mathbf{G}(f)$

Fig. 5 System block diagram for the adaptive interactive learning for dynamic resource management.